Finite sample inference for extreme value distributions

July 17, 1998

Mark J. Dixon, Anthony W. Ledford, and Paul K. Marriott[1]

SUMMARY

We consider the problem of small sample inference for the generalised extreme value distribution. In particular, we show the existence of approximate and exact ancillary statistics for this distribution and that small sample likelihood based inference is greatly improved by conditioning on these statistics. Ignoring the ancillary statistics in inference can have severe consequences in some standard applications of extreme value theory. We illustrate this via simulation and by analysis of two data sets, one based on sea-levels and the other on insurance claims.

Keywords: ancillary statistic, asymptotic ancillarity, conditional inference, estimator performance, generalised extreme value distribution, maximum likelihood estimation.

1 Introduction

The motivation for this study arose from an examination of small sample inference for extreme value distributions. For small samples, a long-standing but unresolved question is how best to estimate the parameters of the generalised extreme value (GEV) distribution which has distribution function

$$G(z;\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-1}\right\} \tag{1.1}$$

where $[y]_+ = \max\{y, 0\}$, and $\mu, \sigma(>0)$, and ξ are location, scale and shape parameters respectively. The case $\xi = 0$ is interpreted as the limit of (**??**) as $\xi \to 0$, and the cases Fréchet, Gumbel and (negative) Weibull correspond to $\xi > 0$, $\xi = 0$, and $\xi < 0$ respectively. The value of the parameter ξ in model (**??**) is dominant in determining tail behaviour; $\xi < 0$ corresponds to a distribution with an upper bound, while increasingly large positive values correspond to an increasingly heavy upper tail. The GEV arises as the limiting distribution of the re-scaled maximum of a sequence of random variables, and is commonly used in environmental and financial applications.

There have been a variety of proposed methods for estimation in this family and a corresponding collection of studies to investigate the relative performance of estimators, (see Hosking *et al*, 1985 and Coles and Dixon, 1997). The methods considered in these studies include maximum likelihood, method of moments, probability weighted moments and quantiles methods. For other similar non-regular distributions, such as the (negative) Weibull, a special case of the GEV, a variety of methods have been considered, including corrected maximum likelihood, (Cheng and Iles, 1987) and grouped likelihood methods. For this type of non-regular problem, the optimality properties of maximum likelihood fail to hold (see Cox and Hinkley, 1974), and small sample inference based on the maximum likelihood estimator (MLE) can be poor. Thus a debate arises about which, estimation method, if any, is optimal.

[1]MJD, Department of Statistics, University of Newcastle-upon-Tyne, Newcastle; AWL and PKM Department of Mathematic and Statistics, University of Surrey, Guildford, Surrey, GU2 5XH.

In this paper, we examine the use of conditional inference for the GEV for finite sample sizes. In particular, we show that there exists ancillary statistics and find that previous conclusions about the relative performance of various estimators are misleading because they have been obtained using inappropriate unconditional inference.

A common feature of this type of non-regular problem is that, as the maximum likelihood estimate is not sufficient for $\theta = (\mu, \sigma, \xi)$, it does not capture all the information for inference about θ. In these cases, conditional methods can be viewed as a way of recovering the information that is contained within the data, but not in the MLE. In cases where the MLE is not sufficient for θ, suppose that the information "lost" in the MLE is contained in another statistic, A. Further if A has a distribution which is independent of the parameter θ, then it is said to be ancillary for θ, and unconditional inference based solely on the MLE can be misleading and inappropriate (Cox and Hinkley, 1974).

The existence and/or form of an ancillary for a given problem is not usually obvious, and much research into conditional inference has concentrated on developing methods for obtaining ancillary statistics in a given class of problems. There has been very little consideration of conditional inference for end point problems. We address this here for the case of the GEV.

In Section 2, we show that there exist both exact and approximate ancillary statistics for the GEV, and demonstrate that standard (unconditional) inference may give misleading results. We also demonstrate the role of ancillary statistics as measures of goodness-of-fit. In Section 3 we show by simulation that the proposed ancillary statistics are informative about inference on the parameters of interest, and compare the performance of alternative estimators in the conditional framework. We show that conditional inference may lead to improved estimation, and so affects the choice of estimation procedure in practical situations. We demonstrate the practical importance of conditioning in Section 4 using UK sea-level and an insurance claims data examples.

2 Conditional inference for the GEV

In this section we review briefly standard unconditional inference for the GEV and describe how conditional inference may be implemented.

2.1 Distribution of maximum likelihood estimate

Smith (1985) investigated the asymptotic properties of the maximum likelihood estimator (MLE) for models with density

$$f(x; \theta, \xi, \phi) = (x - \theta)^{-1-1} \quad g(x - \theta; \phi) \text{ for } x > \theta$$

where g is such that $g(y; \phi) \to c(\phi) > 0$ as $y \to 0$, and θ and ϕ are unknown parameters. The GEV distribution is a special case of this. For $\xi > -1/2$ the maximum likelihood estimate has standard asymptotic first order properties, in particular it is asymptotically normal, unbiased and efficient. For $-1 < \xi < -1/2$ it is super efficient, and for $\xi < -1$, maximum likelihood estimates do not exist. By defining a modified maximum likelihood estimator Cheng and Iles (1987) extended standard asymptotic regularity and efficiency to the region $\xi \in (-1, \infty)$.

Simulation experiments indicate that for small sample sizes there can be problems in using the MLE, even when the true ξ lies in the region $\xi \in (-1/2, \infty)$. This is because the product of densities can be very sensitive to certain data configurations. We return to this issue later.

2.2 A simplified model

In the following we motivate our more general findings by considering inference for the following simplified GEV model with fixed location and scale parameters, $\mu = 0, \sigma = 1$ and unknown $\xi > -1$. We observe $z_1, \ldots, z_n$ independent identically distributed observations from $GEV(0, 1, \xi)$ and wish to estimate ξ.

Define $m = \min\{z_i, i \in 1, \ldots, n\}$ and $M = \max\{z_i, i \in 1, \ldots, n\}$. If $m < 0$ and $M > 0$ then the log-likelihood function will have support only on the interval $\xi \in (-1/M, -1/m)$ because the term $(1 + \xi z_i)$ in the likelihood

$$l(\{z_1, \ldots, z_n\}; \xi) = \prod_{i=1}^{n} \left\{ [1 + \xi z_i]_+^{-(1+)} \quad G(z_i; 0, 1, \xi) \right\}$$

must be non-negative. Thus the support for the likelihood function is data dependent and is determined by the largest and smallest observation. This data dependence of the support of the likelihood is in fact true in general no matter what the sign of m and M. As an example, consider the sample of size 20 given in Appendix A, generated from $GEV(0, 1, 0.4)$. The log-likelihood function is plotted in the top panel of Figure **??**. For this data $m = -1.259$ and $M = 4.260$. The support of the likelihood thus lies on the interval $(-0.23, 0.79)$, indicated by the dotted lines in Figure **??**. We infer from this that the true ξ *must* also lie in this region.

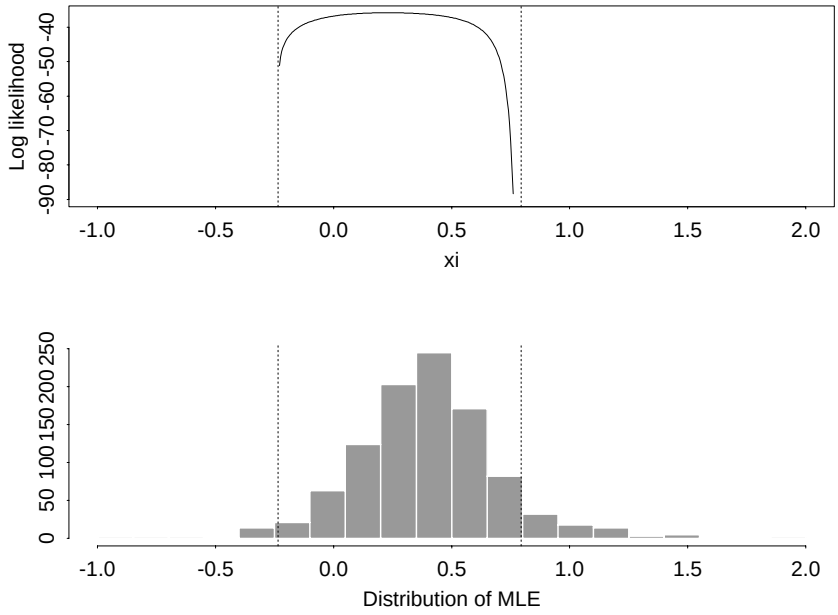

Figure 1: Log-likelihood function for GEV

In contrast, the bottom panel in Figure **??** shows the unconditional distribution of the maximum likelihood estimate for $n = 20$ and $\xi = 0.4$ as an histogram, constructed by simulation. It is clear that inference based on this unconditional distribution will be misleading. In particular we *know* from the data set that the true ξ lies in the interval $(-0.23, 0.79)$. This follows since parameter values outside this region are

logically inconsistent with the data as they lie outside the possible support of the likelihood. Unconditional inference completely fails to incorporate this important point.

Bayesian methods provide an additional inferential perspective here, as *any* prior will give a posterior that is zero outside the interval $(-0.23, 0.79)$.

2.3 Conditional Inference

There is a large amount of literature which connects ideas of conditioning on ancillary statistics and the shape of the likelihood function. This goes back to Fisher's (1925) example of the line/circle model where he argues that inference should be conditional on the information contained in the observed Hessian. In general the argument is that when the maximum likelihood estimate is not sufficient it is important to recover as much information as possible, and that often this information is found in the observed Hessian. Amari (1985) showed that asymptotically the statistic which contains the next highest amount of information after the maximum likelihood estimate can be characterised geometrically. Further, he showed that given a choice of ancillary statistics, as in the multinomial example of Basu (1964), then the statistic which contains the largest amount of information should be chosen. Marriott and Vos (1996) directly connect this information criterion with the behaviour of the shape of the likelihood function. Similar ideas are explored by Efron and Hinkley (1978) where they argue that the observed information matrix is preferable to the expected since it better approximates the variance of the conditional distribution of the maximum likelihood estimate. Further they show that a function of the observed information matrix is an asymptotic ancillary, and so in general it is possible to recover the information that the Hessian contains by conditioning. The relationship between the conditional distribution of the maximum likelihood estimate and the likelihood function is also a key part of Barndorff-Nielsen's very powerful p^*-formula which has been shown to have wide applicability in many classes of families, see Barndorff-Neilsen (1988) and Barndorff-Neilsen and Cox (1994). Reparameterising the parameter space according to the shape of the log-likelihood function, using so called *directed likelihood*, has been shown to be a very effective tool in classical and Bayesian analysis, see Sweeting (1996) and the references therein. In particular, quantities based on directed likelihood are automatically conditional on any second order ancillary. Good references for conditional inference and the role of ancillary statistics is Reid (1995?) and Cox and Hinkley (1974).

In the example of Section 2.2 there are in fact a number of exact ancillary statistics which give direct information about the shape and support of the likelihood function. For example, suppose we wished to conduct inference conditionally on the fact that the observed log-likelihood function had no upper end point, i.e. that its support is of the form $(-a, \infty)$. It is easy to see that this happens if and only if m, the minimum observed data point, is positive. The probability of this event happening is given by

$$\Pr(m > 0; \xi) = \Pr(z > 0; \xi)^n = \{1 - G(0; 0, 1, \xi)\}^n = \{1 - \exp(-1)\}^n$$

which is independent of ξ. Hence the sign of m is an exact ancillary statistic. The conditionality principle, Cox and Hinkley (1974), indicates that we should therefore conduct inference conditionally on this fact. Alternatively it is clear that the proportion of negative observations within the sample is also an exact ancillary statistic. While these ancillary statistics have only academic interest since there is no direct extension to the unrestricted case of μ and σ unknown, they do motivate the proposed ancillary in the next section.

2.4 A class of approximate ancillaries

We return now to the general case of inference in the GEV distribution when all three parameters μ, σ and ξ are unknown. The above exact ancillaries do not extend to this family in an immediate way. Therefore we look for a class of approximate or asymptotic ancillary statistics in order to capture the information in the shape of the likelihood function. As noted in Section 2.3, asymptotic ancillaries are frequently constructed from the higher derivatives of the likelihood function around the maximum likelihood estimate. However as Figure **??** and simulation experiments show, the GEV log-likelihood function is frequently not well approximated by polynomial families due to its compact support. In particular it is often very far from being approximately parabolic, and so the information given by the Hessian is not the appropriate information for inference.

Informal analysis of the simulated samples reveals that the position or configuration of the points relative to $\hat{\mu}$, the MLE of μ, contains information about the shape of the likelihood function. This suggests that if the data are transformed onto a common scale, then the position of the observed points will provide the basis for an informative ancillary statistic. Specifically our interest will be the region of support of the likelihood function.

Denoting the order statistics by $Z^{(r)} : r = 1, \ldots, n$, we examine the r-dimensional statistic

$$\{F(z^{(r)}; \hat{\mu}, \hat{\sigma}, \hat{\xi}), r = 1, \ldots, n\},$$

i.e. the order statistics transformed through the fitted GEV distribution. Transforming the data points z by their true distribution function $F(z; \mu, \sigma, \xi)$ will give the order statistics of a $U(0,1)$ distribution, which is clearly independent of μ, σ and ξ. Hence we might expect $\{F(z^{(r)}; \hat{\mu}, \hat{\sigma}, \hat{\xi}), r = 1, \ldots, n\}$ to be approximately independent of μ, σ and ξ and hence approximately ancillary.

Theorem 1 Define

$$W_r = F(z^{(r)}; \hat{\mu}, \hat{\sigma}, \hat{\xi}),$$

$r = 1, \ldots, n$, where $z^{(r)}$ is the r^{th} smallest order statistic. Then the statistic (W_1, W_n) is a second order ancillary statistic for the parameters (μ, σ, ξ).

Proof: See Appendix B.

.

Lemma 2 shows that information not contained in the maximum likelihood estimate can be recovered, at least partially, by using the information contained in the approximate ancillary statistic (W_1, W_n).

Lemma 2 The statistic (W_1, W_n) together with the maximum likelihood estimate determines the support of the likelihood function.

Proof: The log-likelihood is given by

$$\sum_{i=1}^{n} \{-\log(\sigma) - (1 + 1/\xi) \log[1 + \xi(\frac{z_i - \mu}{\sigma})] - [1 + \xi(\frac{z_i - \mu}{\sigma})]^{-1} \quad \}$$

which has support over those (μ, σ, ξ) for which for each $i = 1, \ldots, n$

$$\xi(\frac{z_i - \mu}{\sigma}) > -1.$$

The solution space to these linear constraints is completely determined by $m = \min(z_i)$ and $M = \max(z_i)$, $i = 1, \ldots, n$. Clearly, m and M can be determined by the statistic $(\hat{\mu}, \hat{\sigma}, \hat{\xi}, W_1, W_n)$ since

$$m = F^{-1}(W^{(1)}; \hat{\mu}, \hat{\sigma}, \hat{\xi}), M = F^{-1}(W^{(n)}; \hat{\mu}, \hat{\sigma}, \hat{\xi}).$$

QED

These two results show that by recording (W_1, W_n) together with the maximum likelihood estimate the support of the likelihood function may be determined. By conditioning on these approximately ancillary statistics it may be possible to recover this information in inference. In Section 3 we investigate conditioning for sample sizes typical of practical applications.

2.5 Goodness of fit

An alternative use of $\{W_1, \ldots, W_n\}$ is to construct measures of goodness of fit. The Moran goodness of fit statistic, (Cox and Hinkley, 1974), compares the fitted and empirical probability distribution functions, and in particular it uses a summary statistic to investigate how close the $(W_1, \ldots, W_n)$ is to the order statistics of a uniform distribution. Our method which focuses on (W_1, W_n), has a similar interpretation and assesses how well the estimated distribution fits in the tails. Hence W_n being abnormally small, for example, would indicate that the fitted model has a much heavier tail than would be expected. Conditional inference is able to exploit this information.

The main reason that standard asymptotic normality results break down for the GEV, according to Cheng and Iles (1987), is that the likelihood, constructed as the product of densities, can be misleading as a measure of the probability of observing an event. In particular since the density function can have a very large maximum for some parameter values, observations near this maximum can dominate the fitting process. There is therefore information in the statistic

$$(W_1, \ldots, W_n)$$

regarding the uniformity of fit on all the observations, especially those in the tail. It is this information which we exploit by conditioning.

2.6 The ancillary a.

The previous section notes that W_1 measures the goodness of fit in the lower tail, and W_n in the upper. In order to have reliable inference we require a good fit in both tails. We therefore concentrate, by using Theorem 3, on a single statistic which captures both properties.

Theorem 3 The statistic

$$a = \max\{W_1, 1 - W_n\},$$

is second order ancillary.

Proof: See Appendix B.

.

We note that W_1 being abnormally large, or W_n small, indicates a poor fit in one of the tails. Hence if a is abnormally large we can infer a poor fit in one or both tails.

3 Simulation study

In this section we use simulation to evaluate the properties of the proposed asymptotic ancillary for sample sizes which we regard as typical and important for practical extreme value analysis.

3.1 Estimation

For some samples, especially when n is small (less than 15) and $|\xi| > 0.5$, the standard maximum likelihood estimate does not exist because the likelihood sometimes tends to infinity as the boundary of support of the likelihood approaches the largest (or smallest) observation. In this case a singularity occurs which may lead to poor estimation for other parameters. The poor performance of maximum likelihood in non-regular situations such as this is well known. See Smith (1985) and Cheng and Illes (1987).

Cheng and Illes (1987) suggest that this problem may be solved by allowing for the inherent discreteness of real data that occurs due to finite accuracy recording. They maximise the product

$$\prod_{i=1}^{n}[G(z_i + h; \theta) - G(z_i - h; \theta)]$$

rather than the usual product of densities. Here h is a small number determined by the accuracy of the recorded data. Harter and Moore (1966), Cheng and Iles (1987) and Smith (1985 and 1990) suggest that when the usual MLE is not obtainable, the maximum (or minimum as appropriate) observation may be used as an estimate of the endpoint of the distribution, α say. The other parameters are then found by maximum likelihood estimation based on the remaining observations.

To ensure robust numerical procedures we adopt a similar approach to that of the above authors. Maximisations are achieved using a quasi-Newton optimisation routine (A fortran algorithm, NAG(1997) E04JAF).

3.2 Ancillarity

Before describing the simulation results, we note that since μ and σ simply correspond to location and scale transformations in the data, our main interest is in the effect of conditioning on estimation of the shape parameter ξ. We hence have $\mu = 0$ and $\sigma = 1$ in all simulations. For each simulated sample of size n, we obtain an estimate $\hat{\theta}$ of θ as described above and the corresponding ancillary $W = (W_1, W_n)$ given by

$$W_r = F(z^{(r)}; \hat{\mu}, \hat{\sigma}, \hat{\xi}); r = 1, n.$$

From this we calculate

$$a = \max\{W_1, 1 - W_n\}.$$

3.2.1 Asymptotic ancillarity of a

We now examine the approximate independence of the distribution of a on ξ for various values of n and ξ. Figure 2 illustrates the distribution of a for parameter values $\xi = -0.4, 0.0, 0.4, 0.8$ and for $n = 10$. We illustrate with a kernel density plot and boxplots. There is some small dependence on ξ at this sample size. However Figure 3 shows the same information for sample size 20. Even at this small sample size we see that the asymptotic approximation is a good one.

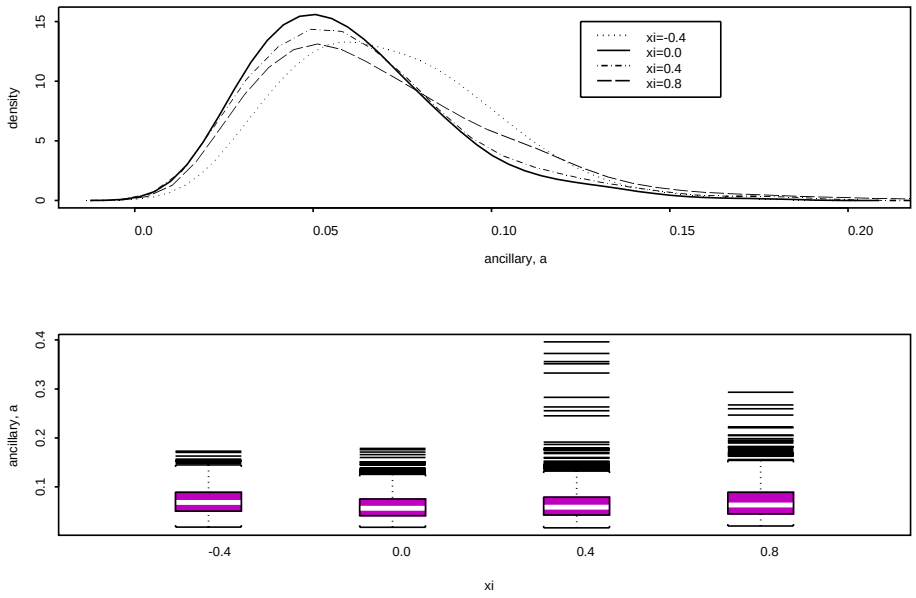

Figure 2: The dependence of a on ξ: Sample size 10

3.3 The effects of conditioning on a.

We now see, again in a simulation study, exactly what is the effect of conditioning on the ancillary statistic a. Recall that the previous theory predicts that we would expect a good fit if a is not too large. Further we would predict that maximum likelihood estimation will be reliable when we have this good fit. Conversely when there is a poor fit in the tails the reliability of the estimate should be open to question. This view is consistent with the traditional view of an ancillary in that it measures the quality of the estimate, in particular the size of the standard error.

Figure 4 shows, for sample size $n = 15$ and $\xi = 0.4$, the relationship between a and $\hat{\xi}$. The top panel is a scatterplot of $\hat{\xi}$ and a. The vertical line corresponds to the $\xi = 0.4$, its true value, and the horizontal line corresponds to $a = 0.06$ which we are using as the conditioning threshold.

Below $a = 0.06$ we see that the estimates of ξ lie in the range $(-0.5, 1.3)$ however above this value the variation of $\hat{\xi}$ is much larger and includes values well above and well below the true value. This point is further illustrated in the lower panel where we show density plots of the unconditional distribution of $\hat{\xi}$ and its distribution conditional on $a < 0.06$ (the dot-dashed line) and on $a > 0.06$ (the dotted line). The vertical line is again $\xi = 0.4$ the true value.

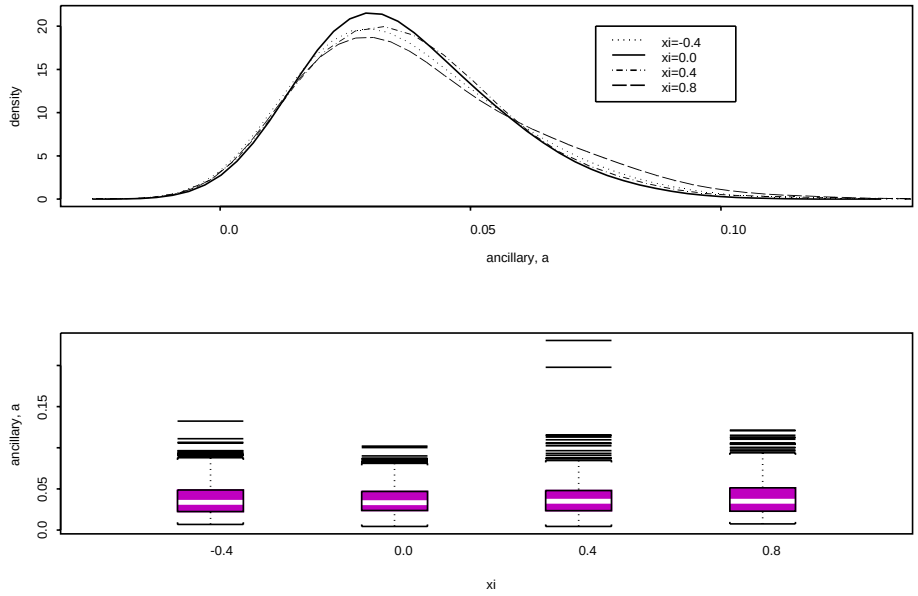

Figure 3: The dependence of a on ξ: Sample size 20

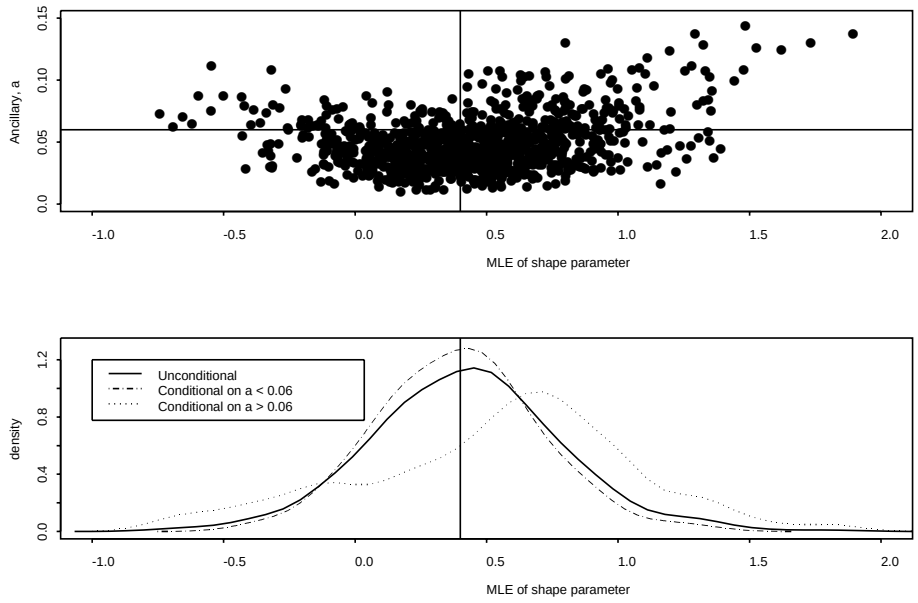

Figure 4: The effect of conditioning: Sample size 15

We immediately see that the previous predictions are confirmed. If a is small enough and we condition on the ancillary we have much tighter confidence intervals and we exclude the very heavy tailed, or very short tailed estimates. Indeed for the large a values we have a bimodal density function for the conditional

distribution.

3.4 Discussion of study

As has already been noted, there is a long standing question in extreme value applications about which method, if any, is optimal for estimating the GEV parameters from data. Although Smith (1985) proved that maximum likelihood is asymptotically optimal in the usual sense when $\xi > -0.5$, Hosking *et al* (1985) showed by simulation that, for small samples and when the shape parameter is close to zero, maximum likelihood performs very poorly in terms of mean bias and mean squared error. They also suggested that of the alternatives, the method of probability weighted moments was the best competitor. Coles and Dixon (1998) extended this study, and found that underlying the superior performance of the PWM estimator is the assumption that $\xi < 1$. While this may be plausible in much practical work it does place the comparison of the PWM and maximum likelihood estimation on a different footing. In particular the distribution of the maximum likelihood estimate has support which includes values $\xi > 1$.

The simulations suggest that comparisons based on unconditional inference are misleading. Note that, for example in Figure 4 it is seen that values of $\xi > 1$ rarely occur conditionally on $a < 0.6$. As noted by Coles and Dixon (1997), which method is optimal depends on specific features of the analysis such as the loss function and the prior beliefs about the value of ξ. It is important to realise that the very large estimated values which can occur with large a can be very important in applications. We discuss this issue in more detail in Section 4.3. Practical issues are considered in the next section.

4 Applications

Before considering the application of our methods to specific data sets, we outline the motivation behind using, and the existing implementation of, the GEV in applications. The underlying result of extreme value theory is the extremal types theorem. In summary, the extremal types theorem tells us that the distribution of the maximum of a sequence of random variables, subject to suitable, broad conditions, will be (approximately) a member of the GEV family. The approximation will generally improve as the number of variables from which the maximum is taken is increased. (See Leadbetter et al or Ries at al?? for details). This theory can be directly applied by fitting the GEV to a data set consisting of observations that are derived by taking the maximum of a sequence of values over a suitably long period.

Having obtained a suitable data set, the next step is to estimate the GEV parameters. A brief review of current applications suggests that the main competitors are maximum likelihood and probability weighted moments. Thus a pragmatic approach is to initially apply each of these methods and compare the results. If the estimates do not agree, it is not clear how one should proceed. The current advice would seem to be that unless the MLE is preferred for other reasons, the PWM estimates are preferable due to the heavy tails of the MLE estimator noted in Section ?.

This (unconditional inference) approach, and the application of our conditioning ideas, are now illustrated using 3 data sets, two UK maximum sea-level data sets and one insurance claims data set.

4.1 Sea-level data

The first two data sets consists of annual maximum sea-level data from North Shields and Rye, two UK coastal sites. Details of these data are given in Graff (1981) and Coles and Tawn (1990). For illustration, we ignore the possibility of jointly modelling the data from neighbouring sites, and at each site, the aim of the analysis is to provide estimates of high quantiles of the fitted distribution in order to assess the required design height of a sea-wall. In the oceanographic literature these are usually referred to as return levels. In particular, we will concentrate on estimating the 100 year return level which corresponds to the 0.01 upper quantile of the distribution of the annual maximum sea-level. Figure **??** shows these data plotted against time: there are 35 and 17 data points at North Shields and Rye respectively. **Antnee**

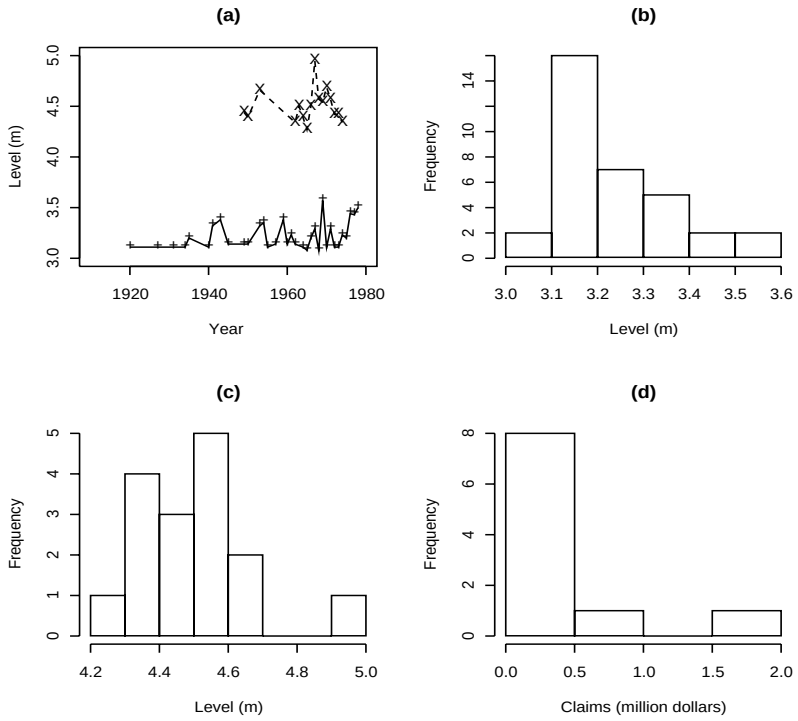

Figure 5: Plot and histograms of sea-level and insurance data sets. (a). ——, and - - - - are annual maximum sea-levels from North shields and Rye respectively. (b) and (c): histograms of annual maximum sea-levels from North shields and Rye respectively. (d) Histogram of the insurance data.

can you check in the book, and see if it is dollars or what?

4.2 Insurance data

This data set contains 3-yearly maxima insurance claims globally from 1963 to 1992, standardised to account for inflation. The data are obtained from Embrechts, Klupperberg and Mikosch (1997) who consider various extreme value analyses of them. The motivation for analysis of these data is to aid the setting of premiums for reinsurance companies. As with the sea-level application, it is the high quantiles of the fitted annual maximum distribution that are of interest. Figure **??**d shows an histogram of these data. Note that the data are heavily skewed to the right.

4.3 Data analyses

Table ?? summarises the results from fitting the GEV to the three data sets using both ML and PWM estimation.

Data set	MLE ($\hat{z}_{0.01}$)	PWM ($\tilde{z}_{0.01}$)	Ancillary	Unconditional	Conditional
North Shields	4.90m	3.87m	0.07	3.87m	3.87m
Rye	5.08m	5.32m	0.020	5.32m	5.08m
Insurance data	£5.61million	£2.58million	0.04	3.87	5.61

Table 1: GEV fits at the three data sites. The second and third column $\hat{z}_p$, and $\tilde{z}_p$ represent the 0.01 upper quantile of the GEV fitted by MLE and PWM respectively. The Unconditioanl columns shows the concluions that would be drawn based on the MLE unless the PWM has a substantially lower ξ value, in which case the PWM is used. The conditional column uses the MLE unless the ancillary is above 0.06.

For North Sheilds it is clear from the ancillary value that the MLE is likely to be heavily biased, and that we should use the PWM value of 3.87m as opposed to the value 4.9m. In fact this is supported by the spatial analyses, using additional information from neighbouring sites, of Dixon Tawn and Vassie (1998). Although the final conclusion is the same as in the unconditional procedure, we have now identified the MLE as being poor.

For Rye, the MLE and PWM estimates of ξ are 0.11 and -0.10 respectively. In this case, may well go for PWM, due to heavy positive bias of MLE. In fact using simulation, we can produce the sampling distribution of each estimator both unconditionally, and conditionally on the ancillary observed. The results are summarised in Table ??.

Statistic	MLE uncon	PWM uncon	MLE Cond on $a = 0.02$	PWM Cond on $a = 0.02$
RMB	0.16	0.04	-0.03	0.07

Table 2:

These results suggest the use of the MLE in this case. Thus the conditioning has provided a different estimate, in this case, 30cm lower than previously.

Of these three data examples, it is the insurance data that best illustrates the importance of our conditioning. The ancillary value of 0.04 tells us that we expect no large biases in the MLE. In fact, in cases where the true ξ is close to or greater than 1, Dixon and Coles (1998) have shown that the PWM substantially underestimates ξ and hence high quantiles of the fitted GEV. Thus it is likely that previous conclusions, under the unconditional framework, underestimate the claim quantile by £2.3m. Our conditioning has corrected this problem.

4.4 Applications discussion

There are then two issues outstanding for applications. Firstly what value of a indicates reliability for a given sample size? Secondly what should the procedure be when the ancillary implies that the maximum likelihood estimate may be unreliable?

The first of these issues is complicated by the question of what the effect of a large bias in the estimation would have on the application in question. We have concentrated in this paper on the error measured on the scale of ξ. In fact for many applications this scale can seriously underestimate the effects of a poor estimate. For example it may well be that return values are of primary interest. In which case the large estimates of ξ observed in our simulation study with large a correspond to very heavy tails and extremely large return values. In this case the results of conditioning are even more important.

Inspection of Figure 4 indicates that there is perhaps a qualitative change in behaviour at the value $a = 0.06$. Extensive simulations and inspections of similar plots indicates that similar changes occur for the following values of a at the sample sizes.

Critical Value	Sample Size	Proportion above threshold
0.09	10	0.2
0.06	15	0.2
0.06	20	0.1
0.06	50	≤ 0.01

We also include in the above table the proportion of samples which have fallen above the threshold. As would be expected as the sample size increases differences between conditional and unconditional inference disappear, as almost all samples will fall below the threshold. The threshold in our simulations (0.06) is fairly robust to changes in sample size, hence might be a sensible bench mark to use in applications.

The second question is of course harder to answer. If we observed a value of the ancillary which indicates that there may be high bias in the estimate we can of course resort to other estimation methods, as discussed above. The main message we think that this analysis has given is that the maximum likelihood estimate in this case should not be taken at its face value, and more detailed modelling an analysis must be required.

Acknowledgements

The authors would like to thank T. Sweeting, M. Crowder, A. Kimber and J. Tawn for helpful advice while preparing this paper.

Appendix A

The data set used in Section 2.2 is:

{-1.25883676, -0.61013271, -0.53189307, -0.47713044, -0.28058898, -0.22787223, -0.16460202, -0.13072007, 0.01507085, 0.07008835, 0.07998088, 0.75858182, 1.59335496, 1.62750011, 1.95151275, 2.32614352, 2.95821996, 3.27915414, 3.86269522, 4.26005164. }

Appendix B

Proof of second order ancillarity

Let the CDF for x be $F\ (x)$, let the log-likelihood function be $\ell(\theta; (x_1, \ldots, x_n))$. Assume we have observed data $(x_1, \ldots, x_n)$, let $\hat{\theta}$ be the maximum likelihood estimate for the full data, and let $\hat{\theta}_{[i]}$ be the maximum likelihood estimate for the dataset excluding the sample x_i.

We use the following standard result

Lemma If the Fisher information is finite and non-zero we have that

$$\hat{\theta} - \hat{\theta}_{[i]} = O_p(\frac{1}{n})$$

Proof. Elementary.

.

We want to investigate the statistic $\hat{F}(x_i)$ so consider the probability

$$\begin{aligned} \Pr(\hat{F}(x_i) < u) &= \Pr(\hat{F}_{[i]}\ (x_i) + O_p(\frac{1}{n}) < u) \\ &= \Pr(\hat{F}_{[i]}\ (x_i) < u) + O(\frac{1}{n}) \end{aligned} \tag{4.1}$$

assuming the smoothness of the $F(x;\theta)$ as both functions of x and θ and using the lemma above.

Further, by conditioning we have

$$\Pr(\hat{F}_{[i]}\ (x_i) < u) = \int \Pr(\hat{F}_{[i]}\ (x_i) < u|\hat{\theta}_{[i]}) p\ (\hat{\theta}_{[i]}) d\hat{\theta}_{[i]} \tag{4.2}$$

where $p\ (\hat{\theta}_{[i]})$ is the density of $\hat{\theta}_{[i]}$ under θ. Note that $\hat{\theta}_{[i]}$ is a function of $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ and hence is independent of x_i.

Combining equations (**??**) and (**??**) we have

$$\begin{aligned} \Pr(\hat{F}(x) < u) + O_p(\frac{1}{n}) &= \int \Pr(x_i < \hat{F}^{-1}_{[i]}(u)|\hat{\theta}_{[i]}) p\ (\hat{\theta}_{[i]}) d\hat{\theta}_{[i]} \\ &= \int F\ (\hat{F}^{-1}_{[i]}(u)) p\ (\hat{\theta}_{[i]}) d\hat{\theta}_{[i]} \end{aligned} \tag{4.3}$$

Since $\Pr(x_1 < y|y, \theta) = F\ (y)$ if x_1 is independent of y.

Consider then the integrand

$$F\ (\hat{F}^{-1}_{[i]}(u)) := F\ (\hat{G}_{[i]}\ (u)),$$

say.

Expand $\hat{G}_{[i]}\ (u)$ around θ and using the fact that $\hat{\theta}_{[i]} - \theta = O_p(1/\sqrt{n-1})$. This gives

$$F\ [\hat{G}_{[i]}\ (u)] = F\ [G\ (u) + (\hat{\theta}_{[i]} - \theta)\frac{\partial G}{\partial \theta}(u) + O(\frac{1}{n})]$$

Taylor expanding $F\ (x)$ now with respect to x around $G\ (u)$ gives

$$\begin{aligned} F\ [G_{\hat{[i]}}\ (u)] &= F\ [G\ (u)] + (\hat{\theta}_{[i]} - \theta)\frac{\partial G}{\partial \theta}(u)\frac{\partial F}{\partial x}[G\ (u)] + O_p(\frac{1}{n}) \\ &= u + (\hat{\theta}_{[i]} - \theta)\frac{\partial G}{\partial \theta}(u)\frac{\partial F}{\partial x}[G\ (u)] + O_p(\frac{1}{n}) \end{aligned}$$

Now, substituting into the integral in equation (**??**) and again assuming the asymptotic unbiasedness of $\hat{\theta}_{[i]}$ gives

$$\begin{aligned} \Pr(F\hat{}(x) < u) &= \int \left(u + (\hat{\theta}_{[i]} - \theta)\frac{\partial G}{\partial \theta}(u)\frac{\partial F}{\partial x}(G\ (u)) + O_p(\frac{1}{n}) \right) p\ (\hat{\theta}_{[i]}) d\hat{\theta}_{[i]} + O_p(\frac{1}{n}) \\ &= u + O_p(\frac{1}{n}) \end{aligned}$$

Hence this statistic is at least second order ancillary. .

To finish the proof of Theorem 1 we note that to the correct order the statistics

$$F\hat{}(x_i), F\hat{}(x_j)$$

for $i \neq j$, are independent. Hence their joint distribution will be ancillary to the correct asymptotic order.

It then follows by standard arguments that the distribution of the order statistics of any n^{tple} will be ancillary to the required order. This completes the proof of Theorem 1.

Theorem 3 follows immediately by standard properties of order statistics.

References

Amari S-I, (1985), *Differential-Geometric Methods in Statistics.* Springer, Lecture Notes in Statistics, **28**, Berlin.

Bardorff-Nielsen, O.E., (1988) *Parametric Statistical Families and Likelihood.* Springer, New York.

Bandorff-Nielsen and Cox (1994),*Inference and Asymptotics* Monographs on Statistics and Applied Probability, **52**, Chapman and Hall, London

Basu, D., (1964), Recovery of ancillary information, *Sankya A*, **26**, 3-16.

Cheng, C. H. and Iles, T. C. (1987). Corrected maximum likelihood in non-regular problems. *JRSS Series B*, **49**, 95-101

Coles, S. G. and Dixon, M. J. (1997). Likelihood based inference for extreme value distributions. *In preparation.*

Coles, S. G. and Tawn, J. A. (1990). Statistics of coastal flood prevention. *Phil. Trans. R. Soc. Lond., A*, **332**, 457–476.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London. *J. Roy. Statist. Soc. B*, **52**, 393–442.

Emrecths book (Antnee fill in)

Fisher, R.A., (1925), Theory of statistical estimation, *Proceedings of the Cambridge Philosophical Society*, **22**, 700-725.

Graff, J. (1981). An investigation of the frequency distributions of annual sea-level maxima at ports around Great Britain. *Estuarine Coastal Shelf Sci.*, **12**, 389–449.

Harter and Moore (1966)

Hosking, J.R.M., Wallis, J.R. and Wood, E.F. (1985). Estimation of the generalized extreme-value distribution by the method of probability- weighted moments. *Technometrics* **27**, 251–261.

Marriott, P.K. and Vos, P. (1997), On the global geometry of parameteric models and information recovery, submitted to *Bernoulli*
NAG (1997)

Reid, N. (1995?)

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika* **72**, 67-92.

Smith, R. L. (1990). Extreme value theory. In *Handbook of Applicable Mathematics*, **7**, 437–471, ed. W. Ledermann, John Wiley, Chichester.

Sweeting, T., (1996), *Bayesian Stats* Approximate Bayesian computation based on signed roots of log-density ratios (with discussion). *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds.), Oxford: University Press, 427-444.