Extremes 2:1, 5–23 (1999)

Likelihood-Based Inference for Extreme Value Models

STUART G. COLES s.coles@lancaster.ac.uk
Department of Mathematics and Statistics, Lancaster University

MARK J. DIXON[1] m.j.dixon@city.ac.uk
School of Mathematics, Actuarial Science and Statistics, City University

[Received February 5, 1998; Revised July 23, 1999; Accepted July 26, 1999]

Abstract. Estimation of the extremal behavior of a process is often based on the fitting of asymptotic extreme value models to relatively short series of data. Maximum likelihood has emerged as a flexible and powerful modeling tool in such applications, but its performance with small samples has been shown to be poor relative to an alternative fitting procedure based on probability weighted moments. We argue here that the small-sample superiority of the probability weighted moments estimator is due to the assumption of a restricted parameter space, corresponding to finite population moments. To incorporate similar information in a likelihood-based analysis, we propose a penalized maximum likelihood estimator that retains the modeling flexibility and large-sample optimality of the maximum likelihood estimator, but improves on its small-sample properties. The properties of the penalized likelihood estimator are verified in a simulation study, and in application to sea-level data, which also enables the procedure to be evaluated in the context of structural models for extremes.

Key words. generalized extreme value distribution, maximum likelihood, probability weighted moments, small sample properties

1. Introduction

By almost universal consent, a starting point for modeling the extremes of a process is based on distributional models derived from asymptotic theory. Such arguments stem from the fact that whatever certainties we may have about the behavior of a process when it is in ‘‘typical’’ states, often do not apply at extreme levels. Consequently, what is hoped to be a safer approach is obtained by exploiting asymptotic arguments to derive a parametric family of models which characterize distributional tail behavior; this family can then be fitted to correspondingly extreme data. This argument dates back to Fisher and Tippett (1928), and though there have been a multitude of refinements and enhancements, the basic principle has remained the same. A comprehensive overview of the development of models for extreme value analysis is given by Smith (1990).

Of less common agreement is the method of inference by which such asymptotic models are fitted to data. As both fashions and technologies have developed, so procedures for

[1] Address for correspondence: School of Mathematics, Actuarial Science and Statistics, City University, Northampton Square, London, EC1V 0HB.

parameter estimation in extreme value models have sprouted. These include: graphical methods; method of moments; order statistic methods; probability weighted moments; maximum likelihood; and Bayesian methods. As a general framework for extreme value modeling, likelihood-based methods have many advantages, both theoretical and practical. In most situations extreme value models are ''regular'' in the usual sense of likelihood theory (Smith (1985)), so standard results demonstrate their asymptotic optimality. Moreover, approximate inferences, such as confidence intervals, are immediate. Likelihood functions can also be constructed for complex modeling situations, enabling, for example, non-stationarity, covariate effects and regression modeling. This combination of asymptotic optimality, ready-solved inference properties and modeling flexibility represents a comprehensive package for alternative methods to compete against.

One argument that has been used against maximum likelihood in extreme value models is its small-sample properties. In particular, Hosking et al. (1985) argued that in small samples, probability-weighted moments estimation is superior to maximum likelihood in terms of bias and mean-square error. Despite the many advantages of maximum likelihood noted above, poor performance in small samples remains a serious criticism, since it is not uncommon in practice to need to make inferences about extremes with very few data—the rarity of extreme events means that even long observational periods may lead to very few data that can be incorporated into an extreme value model. The aim of this paper is to examine more closely the comparison between maximum likelihood and probability weighted moments for estimating parameters of extreme value models with small data sets. We focus, in particular, on the assumption of finite population means in the probability moments estimating equations. In effect, this is prior information on the parameters, so in a fair comparison equivalent information should also be included in a likelihood analysis.

We restrict attention to the generalized extreme value distribution (Jenkinson, 1955), which comprises the classical extreme value models, sometimes referred to as the ''three types'' (Gumbel, Frèchet and (negative) Weibull respectively), though a parallel argument could be based on other asymptotic tail models. Our aim is to understand more clearly the apparent failings of maximum likelihood in this situation. In Section 2 we define the generalized extreme value model, giving its asymptotic motivation, and describe the procedures of maximum likelihood and probability weighted moments for parameter estimation within this family of models. In Section 3 we undertake a simulation study to compare in detail the two inference procedures. In light of these results, in Section 4 we consider ways of modifying the maximum likelihood estimator (MLE). In particular, we propose a penalized maximum likelihood estimator (PMLE) for point estimation, and demonstrate that whilst retaining all of the advantages of the MLE, its small-sample properties are comparable with those of the probability weighted moments estimator. The PMLE also inherits all of the asymptotic properties of the MLE, but we demonstrate that for small samples where estimates of precision are required, it may be better to carry out a fully Bayesian analysis of which the PMLE is an approximate point summary. Finally, in Section 5, we give a comparison of moment and likelihood-based estimators in the modeling of annual maximum sea-levels around the UK coast, demonstrating that the advantages of the PMLE carry over to more complex modeling situations.

2. Model and inference procedures

The generalized extreme value distribution is derived from the following classical characterization of extremal behavior. Let $X_1, X_2, \ldots$ be a sequence of independent variables with common distribution function F, and let $M_n = \max\{X_1, X_2, \ldots X_n\}$. After linear re-normalization, to avoid the obvious degeneracy of the distribution of M_n, as $n \to \infty$, the entire class of non-degenerate limiting distributions of M_n is provided by the generalized extreme value family

$$G(z; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi} \right\} \tag{2.1}$$

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$. The case of $\xi = 0$ is interpreted as the limit of (2.1) as $\xi \to 0$. Since the model (2.1) includes location and scale parameters, μ and $\sigma > 0$ respectively, it follows that the distribution function of M_n itself, for large enough n, can be approximated by a member of the generalized extreme value family. The parameter ξ in model (2.1) is a shape parameter whose value is dominant in determining tail behavior; $\xi < 0$ corresponds to distribution with an upper bound, while increasingly large positive values correspond to an increasingly heavy upper tail.

This characterization is often referred to as the ''classical'' extreme value theorem. More recent theory justifies the use of approximation (2.1) for the distribution of M_n even when the series $X_1, X_2, \ldots$ exhibits short-range temporal dependence (Leadbetter et al., 1983, for example).

A natural modeling approach, having observed data $x_1, x_2 \ldots$ from F, is to form block maxima, with blocks often taken conveniently to correspond to a year of observations, and then to fit the family (2.1) to these block maxima. The problem we focus on is the estimation of $(\mu\,\sigma\,\xi)$, assuming that (2.1) is the correct model for block maximum data $z_1, z_2, \ldots, z_k$. The performance of different estimators is often judged in terms of the accuracy of estimation of extreme quantiles, since these quantities are often the principal requirement of an extreme value analysis. Inverting (2.1), the $1-p$ quantile, q_p, satisfies

$$q_p = \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right]. \tag{2.2}$$

Substituting estimates of $(\mu\ \sigma\ \xi)$ into (2.2) gives an estimate of q_p for any chosen p.

2.1. Maximum likelihood (ML)

Since the support of the distribution G is dependent on the parameters, the asymptotic regularity of the model is not immediately obvious. However, this issue was clarified by Smith (1985) who proved that maximum likelihood estimates exist whenever $\xi > -1$ and are regular whenever $\xi > -0.5$. These apparent restrictions are of little practical importance, since $\xi > -1$ corresponds to a distribution with a very short upper tail, a situation which is rarely encountered in applications for which extreme value modeling is required. Thus, at least in the usual situation of $\xi > -0.5$, maximum likelihood estimates can be obtained and approximate inferences derived from standard asymptotic likelihood theory (Cox and Hinkley, 1974, for example).

Other studies of maximum likelihood performance in extreme value settings are given by Cohen (1984), Smith (1986) and Smith and Weissman (1985).

The likelihood function for independent block maxima $z_1, \dots z_n$ from model (2.1) is

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{k} \frac{dG(z_i; \mu, \sigma, \xi)}{dz_i}. \tag{2.3}$$

Likelihood functions for more complex situations, including missing data, non-stationarity, temporal dependence and covariate effects are also straightforward to construct (Davison and Smith; 1990 and Smith; 1989, for example). Prescott and Walden (1980) give an explicit algorithm to maximize the likelihood (2.3), while Splus code for model-fitting that permits the inclusion of generic covariate models is available online from http://www.maths.lancs.ac.uk/~coless/.

2.2. *Probability weighted moments (PWM)*

An alternative estimation scheme to maximum likelihood is to equate theoretical and sample moments. It is well-known (Cohen and Whitten, 1982, for example) that this naive method of moments performs badly for the estimation of extreme value parameters when the shape parameter is large. A referee has pointed out that this could be due in part to the requirement that $\xi < 1/3$; we will argue below that the equivalent constraint for probability weighted moments is the substantive reason for their improved performance relative to maximum likelihood in small samples. The method of probability weighted moments is analogous to the method of moments—matching theoretical and empirical functionals—but based on quantities that give increased weight to the tail information. The use of probability weighted moments for extreme value models was first proposed by Landwehr et al. (1979), who studied the Gumbel model, $\xi = 0$ in (2.1). Extending this analysis, Hosking et al. (1985) studied the performance of probability weighted moments for the complete generalized extreme value family.

Specializing a more general definition, the probability weighted moments of a variable X with distribution function F are defined by

$$\beta_r = E[X\{F(X)\}^r] \qquad r = 0, 1, 2, \dots .$$

An obvious estimator of β_r, based on an observed sample $x_1, \ldots, x_n$, is its empirical counterpart:

$$\hat{\beta}_r = \frac{1}{n}\sum_{j=1}^{n} x_j\{\tilde{F}(x_j)\}^r, \tag{2.4}$$

where $\tilde{F}$ is an empirical estimate of the distribution function F.

Conveniently, for the generalized extreme value distribution, (2.1), analytical expressions can be obtained for the probability weighted moments:

$$\beta_r = (r+1)^{-1}[\mu - \sigma\{1 - (r+1)^{\xi}\Gamma(1-\xi)\}/\xi], \tag{2.5}$$

provided $\xi < 1$. Hence, the method of probability weighted moments estimate of $(\mu\ \sigma\ \xi)$ is obtained by equating (2.5) and (2.4) for $r = 0, 1, 2$. Though this system of equations is non-linear, Hosking et al. (1985) show that an approximate solution is given by

$$\hat{\xi} = -7.859c - 2.9554c^2, \quad \text{where } c = \frac{2\hat{\beta}_1 - \hat{\beta}_0}{3\hat{\beta}_2 - \hat{\beta}_0} - \frac{\log 2}{\log 3}$$

$$\hat{\sigma} = \frac{(\hat{\beta}_0 - 2\hat{\beta}_1)\hat{\xi}}{\Gamma(1-\xi)(1-2^{\hat{\xi}})}$$

$$\hat{\mu} = \hat{\beta}_0 - \hat{\sigma}\{\Gamma(1-\xi) - 1\}/\hat{\xi}.$$

They compared both the asymptotic and small sample properties of the probability weighted moments estimator with the maximum likelihood estimator of the generalized extreme value model parameters. In summary, they find that the relative asymptotic efficiency of the probability weighted moment estimator is around 70% when $\xi \approx 0$, but is substantially lower as $|\xi|$ increases. In contrast, for small samples, they observe through simulation that the probability weighted moments estimator has considerably better properties than the maximum likelihood estimator in terms of both bias and mean square error for values of ξ between -0.4 and 0.4. It is this small sample behavior that we wish to examine more closely.

3. Comparison of methods

A simulation study confirms the observations of Hosking et al. (1985). We simulated $N = 20000$ samples of size n, from the generalized extreme value distribution with shape parameter ξ, for a range of values of n and ξ. Without loss of generality, the location and scale parameters were fixed at $\mu = 0$ and $\sigma = 1$. For each sample the maximum likelihood estimate and probability weighted moments estimate of $(\mu\,\sigma\,\xi)$ were obtained and substituted into (2.2) to give the corresponding estimates of the distributional quantile, q_p, for various values of p. Averaging over the samples leads to estimates of bias and mean square error.

The maximum likelihood estimates were obtained by applying a standard numerical optimization algorithm to the likelihood using simple moment estimators as initial parameter values. The procedure is remarkably robust, though with the smallest sample sizes and default starting values, the algorithm very occasionally failed to converge. In most cases, this difficulty was overcome by changing the initial values. Rather more troublesome is the occurrence, even more occasionally, and only for very small sample sizes, of samples for which the algorithm failed to converge for any initial values. The percentage of samples in which this happens depends on the sample size, and the value of ξ, but is typically less than 2% even in the worst cases. Since both types of problem arose infrequently, leading to samples that demonstrated no particularly unusual characteristics, we rejected these samples in our study, anticipating that this would not unduly bias the results.

For the smallest sample size considered of $n = 15$, and for $\xi = -0.2$ and $\xi = 0.2$, the results are summarized in Table 1. In terms of both bias and mean square error, maximum likelihood is seen to be a poor competitor to the probability weighted moments estimator, especially for the estimate of a particularly extreme quantile, q_p, when ξ is positive. Further insight into the failings of the maximum likelihood estimator is gained by looking at the sampling distributions of the different estimators. Based on the simulated values in the case $\xi = 0.2$, Figure 1 shows a kernel density estimate of the sampling distribution of $\log(q_p)$ for $p = 0.001$. Even on this logarithmic scale, the distribution of the maximum likelihood estimator exhibits substantial positive skew, which explains the poor results identified in Table 1. Figure 1 also suggests that robust measures of estimator performance, which are not affected by the skewness in the tail of the distribution, will be more favorable to the maximum likelihood estimator. This is confirmed in Table 1 which also gives the median bias of both the maximum likelihood and probability weighted moments estimators. By this measure, maximum likelihood is equally competitive. Similar remarks can be made about robust measures of variability.

Table 1. Bias and root mean square error (RMSE) of maximum likelihood (ML) estimates and probability weighted moments (PWM) estimates of q_p, the distributional quantile of a generalized extreme value distribution with shape parameter ξ, based on a sample size of 15.

	$\xi = -0.2$		$\xi = 0.2$	
	$q_{0.01}$	$q_{0.0001}$	$q_{0.01}$	$q_{0.0001}$
ML bias	0.02	24.84	0.78	$>10^3$
PWM bias	0.08	0.47	-0.01	0.55
ML RMSE	1.91	$>10^4$	$>10^2$	$>10^4$
PWM RMSE	0.10	1.16	0.31	6.78
ML median bias	-0.10	-0.15	-0.08	-0.08
PWM median bias	0.06	0.20	-0.14	-0.20

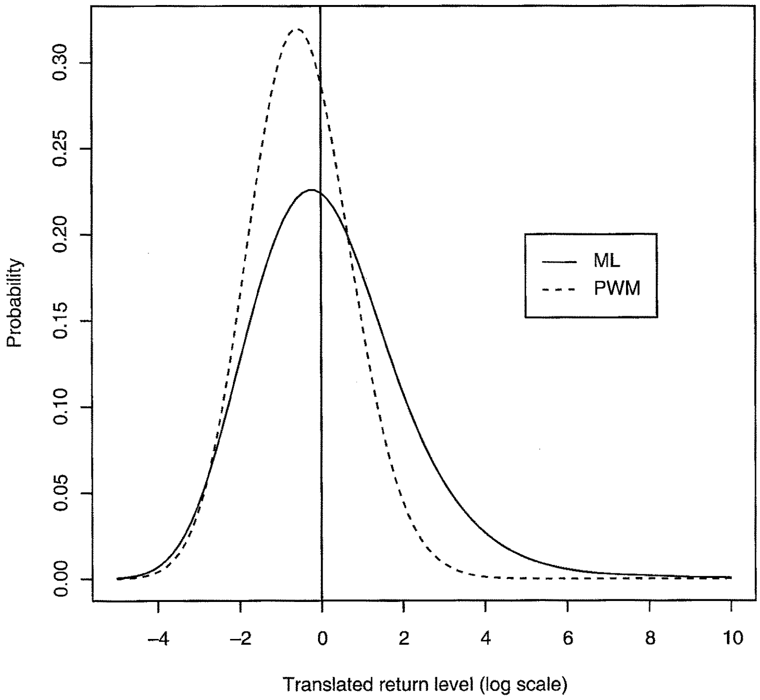

Figure 1. Kernel density estimates of the sampling distribution of the logarithm of q_{0001} in the case $n = 15$, and $\xi = 0.4$. The estimates have been location shifted by the true value, so that the vertical line, at 0 corresponds to the true value of the (logged) return level. The solid and broken lines represent the sampling distribution for ML and PWM respectively.

The poor relative performance of the maximum likelihood estimator of extreme quantiles can be explained by examining the sampling behavior of the estimators of the original parameters $(\mu\,\sigma\,\xi)$ of model (2.1). Examination of density estimates of the respective sampling distributions (not shown) reveals that the main difference between the maximum likelihood and probability weighted moments estimators arises in the distribution of the shape parameter estimate, which, in the case of the maximum likelihood estimator, has noticeable positive skew. Moreover, the non-linearity of mapping (2.2) as a function of ξ means that a small positive bias for ξ translates to a substantial bias for q_p.

The relatively poor estimation of ξ by maximum likelihood can be largely attributed to the different assumptions made by the estimating equations. In particular, the probability weighted moments estimator assumes *a priori* that $\xi < 1$, equivalent to specifying that the distribution has finite mean. Thus, the parameter space of $(-\infty, \infty)$ is mapped to $(-\infty, 1)$ in the estimated parameter space, leading to a reduction in the sampling variation of the estimate of ξ. The price of this is a negative bias in the estimate of ξ, which increases with the value of ξ. However, the non-linearity in equation (2.2) as a function of ξ means that, judged in terms of mean square error of extreme quantiles, underestimation of ξ is penalized much less heavily than over-estimation. Thus, the probability weighted moments estimator admits a certain amount of bias-variance trade-off in the estimate of ξ, which after transformation to extreme quantiles, q_p, avoids the heavy upper tail of the sampling distribution of the maximum likelihood estimator of q_p.

This discussion highlights the role of criteria for estimator assessment, and the issue is particularly sensitive in extreme value modeling. Estimates of extreme quantiles are often used as tolerance settings in, for example, building construction. In such situations underestimation is likely to be grossly more costly than overestimation, in which case mean square error is an inappropriate estimation criterion, and more appropriate criteria may favor the maximum likelihood estimator, which avoids the negative bias in the estimation of ξ. We shall not pursue this argument further since the selection of an appropriate cost function in any particular application will be specific to the circumstances of that application. We observe, however, that any scheme which tends to underestimate ξ, however arbitrary, will inherit the same apparent improvement over maximum likelihood in terms of mean square error of extreme quantile estimates.

Returning to the issue of the constrained parameter space for the probability weighted moments estimator, the restriction $\xi < 1$ can be viewed as prior information, which, in a fair comparison of methods, should be available also to the likelihood-based analysis. Simply restricting the parameter space for ξ to $(-\infty, 1)$ leads to maximum likelihood estimates on the boundary $\xi = 1$ for samples which would have had maximum likelihood estimate $\hat{\xi} > 1$ in the unrestricted case. A more satisfactory adjustment would apply a bijective mapping of the parameter space for ξ, from $(-\infty, \infty)$, into $(-\infty, 1)$.

One possibility is to adopt a prior distribution for ξ with support restricted to $(-\infty, 1)$, and to apply a fully Bayesian analysis, adopting the mean or some other average of the corresponding posterior distribution as the estimate of ξ. Bayesian techniques for extremes have been considered by Smith and Naylor (1987) and Coles and Tawn (1996), each coming to the conclusion that such an analysis can give a more informative likelihood-based analysis of extreme value data than maximum likelihood. In particular, Coles and Tawn (1996) demonstrate through a case study the inferential improvements which can be obtained by the inclusion of expert prior information in an extreme value analysis. Smith and Naylor (1987) deal with the more common situation in which genuine prior information is unavailable, but a Bayesian analysis is carried out on the basis of a reference prior. In spirit, this is our approach also, though for the most part we reduce the computational burden by the use of penalized maximum likelihood estimators. Within this framework we look for a suitable penalty function—equivalent to determining an appropriate reference prior—for incorporating the required structural information about ξ.

4. Penalized likelihood (PL)

4.1. The penalized maximum likelihood estimator

Penalized likelihood is a straightforward method of incorporating into an inference information that is supplementary to that provided by the data. A standard application is to non-parametric smoothing, in which the appropriate likelihood is balanced by a function that penalizes roughness. Silverman (1985) gives a simple overview and cites many more detailed references. The present context is rather different; we use a penalty function to provide the likelihood with the information that the value of ξ is smaller than 1, and that values close to 1 are less likely than smaller values. This application is closer in spirit to a Bayesian formulation in which structural information on ξ is formulated as a prior distribution. After some experimentation we restricted attention to penalty functions of the form

$$P(\xi) = \begin{cases} 1 & \text{if } \xi \le 0 \\ \exp\left\{-\lambda\left(\frac{1}{1-\xi} - 1\right)^{\alpha}\right\} & \text{if } 0 < \xi < 1 \\ 0 & \text{if } \xi \ge 1 \end{cases} \tag{4.1}$$

for a range of non-negative values of α and λ. The corresponding penalized likelihood function is

$$L_{pen}(\mu, \sigma, \xi) = L(\mu, \sigma, \xi) \times P(\xi), \tag{4.2}$$

where L is the likelihood function defined by (2.3). Large values of α in the penalty function correspond to a more severe relative penalty for values of ξ which are large, but less than 1, while λ determines the overall weighting attached to the penalty. The penalty function (on a logarithmic scale) is shown in Figure 2 for various values of α and λ. After further experimentation, it was found that the combination $\alpha = \lambda = 1$ leads to a reasonable performance across a range of values for ξ and sample sizes; hence all our results are reported with respect to this choice. The value of $(\mu\ \sigma\ \xi)$ which maximizes (4.2) is the maximum penalized likelihood estimator (PMLE).

To illustrate the behavior of penalized likelihood, Figure 3 shows the sampling distribution of the PMLE of ξ in the case $n = 25$ for a range of values for ξ. For negative values of ξ, the PMLE is almost indistinguishable from the MLE. However, for positive values of ξ, the PMLE has a behavior much closer to that of the probability weighted moments estimator, thus inheriting the characteristics of smaller variance at the expense of negative bias. In terms of both bias and variance the estimator appears to be at least as good as the probability weighted moments estimator.

The relative properties of the PMLE are illustrated more systematically in Figure 4. Bias and mean square error of estimates of $q_{0.001}$ are plotted as a function of $\log(n)$, where n is the sample size, for a range of values of ξ. First, relative to the maximum likelihood estimates, it is clear that the penalized estimator inherits the usual large sample properties, without being affected by particularly poor performance with smaller samples. Second, with respect to the probability weighted moments estimator, the PMLE is almost uniformly better in terms of both bias and mean square error. The only notable exception is in the case $\xi = 0.4$ with very small samples, for which the PMLE admits larger bias, but with the benefit of considerably smaller mean square error. Figure 4 also illustrates the effect of the skewness in the distribution of $\hat{\xi}$ on the quantile q_p. For example, in the case $\xi = 0.2$, there is a positive bias in the probability weighted moments estimator of q_p,

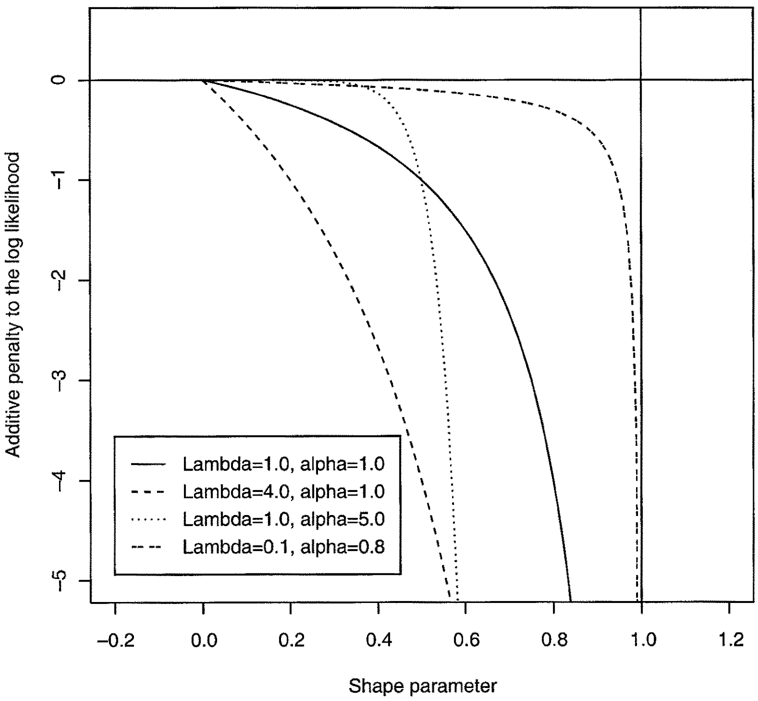

Figure 2. Logarithm of the penalty function for various values of α and λ plotted against ξ.

although Figure 3 indicates a negative bias in the shape parameter. Overall, Figure 4 suggests that for all sample sizes the PMLE has properties which at least match, and in some cases improve upon, those of maximum likelihood and probability weighted moments. Similar behavior was also observed for other quantiles. Figure 4 also suggests that for negative values of ξ the effect of the penalty function is negligible for sample sizes as small as 25, but that for large positive values its effect on sampling properties remains up to samples of size 500 or so.

4.2. *Precision of estimation*

When the true value of $\xi < 1$, the penalty function modifies the likelihood by an amount that is asymptotically negligible. Thus, in this case, all standard asymptotic results for the MLE of the GEV distribution apply also to the PMLE. In particular, this means that when $-0.5 < \xi < 1$, the PMLE is regular, in the usual likelihood sense (Smith, 1985). Hence, approximate confidence intervals can be calculated using either the asymptotic normality of the PMLE, or from the profile penalized likelihood function. The accuracy of these intervals may, however, be quite different from those of the corresponding maximum likelihood estimator, and for small samples the accuracy may be quite poor. Thus, in

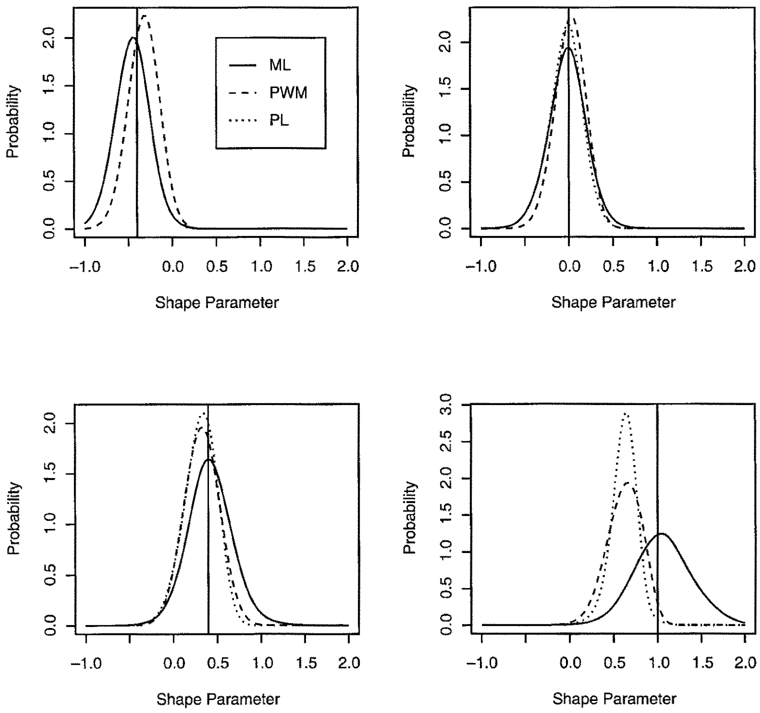

Figure 3. Kernel density estimates of the sampling distribution of ξ. The vertical lines indicate the true values of ξ which are (a) -0.4, (b) 0.0, (c) 0.4, and (d) 1.0.

situations where accurate estimates of precision are required, it may be more convenient to undertake a fully Bayesian analysis for which the PMLE is, at least approximately, an appropriate point summary. This is most simply achieved by MCMC (see Coles and Tawn, 1996, for example). Formally adopting a prior that is proportional to (4.1) for ξ, with vague independent priors on the other parameters, would lead to a Bayesian inference in which the PMLE is the posterior mode. Improper priors can be problematic in MCMC algorithms however, so it is probably better to approximate this analysis with non-informative, though proper, priors for μ and σ, and to apply a similar modification to the lower tail of ξ.

We have adopted such a procedure in a simulation study to compare coverage probabilities of nominal 95% confidence and credible intervals of $q_{0.01}$, based on samples of size 25. Results from Hall et al. (1999) suggest that for similar models, intervals constructed from the profile likelihood or via bootstrap are likely to have better coverage probabilities than simpler approximations based on the asymptotic distribution of the MLE. Our simulation results, summarized in Table 2, suggest that Bayesian credible intervals are also likely to be more accurate. The asymptotic approximations are seen to be poorer for the PMLE than the MLE to an extent that depends on the value of ξ, and that even judged by this frequentist criterion, Bayesian credible intervals are to be preferred.

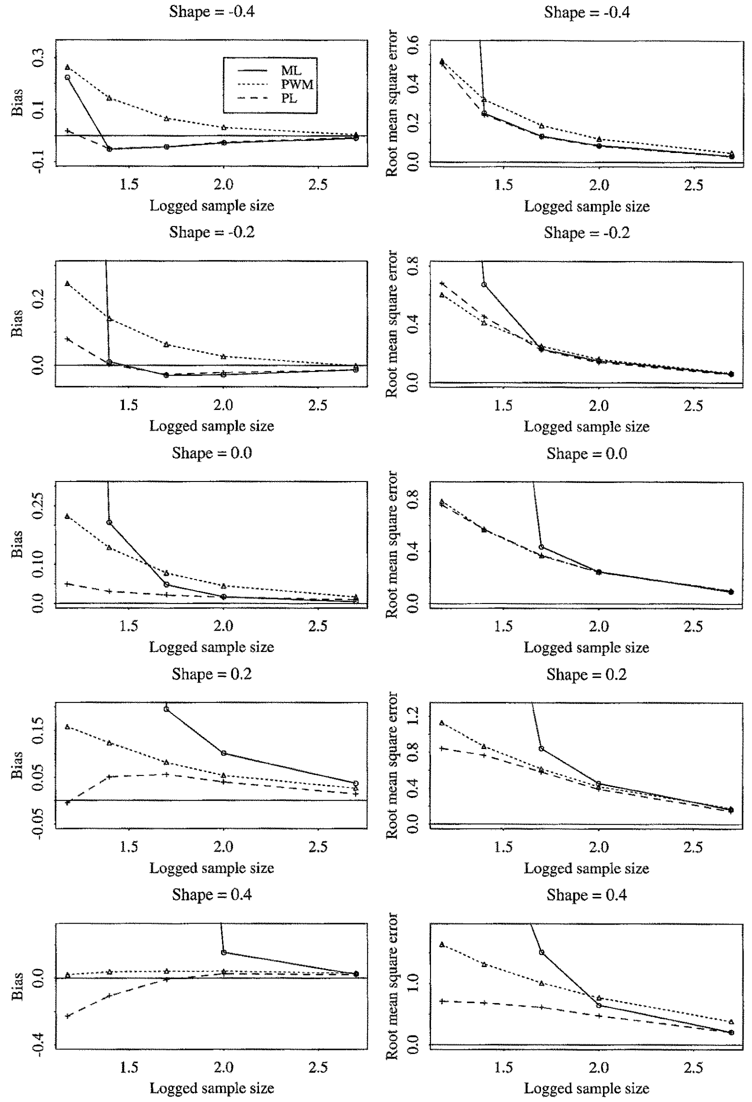

Figure 4. Performance of the three methods, maximum likelihood (ML), probability weighted moments (PWM) and penalized likelihood (PL), in terms of bias and root mean squared error of q_{0001} plotted against sample size (logged to base 10) between 15 and 500. The figures are for shape parameters between -0.4 and 0.4.

5. Application to UK sea-level data

In this section we compare the performance of the various estimators in modeling annual maximum sea-level data from 58 mainland UK coastal sites. The number of annual maximum data available for each site ranges from 10, at Margate, to 136, at Sheerness, with a mean of 43 per site; details of these data are given in Graff (1981) and Coles and Tawn (1990), who assume that the annual maximum data are reasonably modeled as independent observations from the GEV distribution. Under this same assumption we address three points in Sections 5.1 to 5.3 respectively:

- The occasional large over-estimation of the shape parameter for very small samples, and the corresponding advantages of the PWM and PL estimators in terms of extreme quantile estimation;
- The respective bias-variance trade off in each of the three estimation methods;
- The handling of covariates.

5.1. *Heavy tail estimates*

It is not uncommon to apply extreme value analyzes to as few as 15 or 20 annual maxima, sample sizes for which the limitations of maximum likelihood identified in the earlier sections can arise. To illustrate the problem, we examine those sea-level data sets with long data records, and analyze each 15-year window of annual maximum data within the record using the GEV model. This enables assessment relative to estimates based on the entire series, which can be regarded as a ''gold standard'' for comparison. The first example has been rather artificially constructed, but illustrates the comparisons well. For each of the 58 sites which have more than 14 years of data, we construct a number of data sets of length 15 years, starting from the first year for which data is available and incrementing the start year throughout the data span. For example Sheerness has $136 - 14 = 122$ such data sets, 1819–1938, 1820–1939, . . ., 1967–1983. Over all 58 sites this procedure gives rise to a total of 1771 data sets of 15-year windows, to which the GEV was fitted using the three methods, ML, PWM and PL. real data sets of 15 years that would be available for analysis had the observations been made only over that 15 year period. Table 3 highlights a number of the analyzes for which the ML estimates appear anomalous.

Table 2. Coverage probabilities obtained through simulation of nominal 95% confidence interval and Bayesian credible interval for the quantile $q_{0.01}$ based on samples of size 25. The confidence intervals for both MLE and PMLE are based on calculated as $\hat{q}_{0.01} \pm 1.96SE$, where the standard error, SE, is obtained from the information matrix in the standard way. The Bayesian intervals are constructed using vague, but proper, priors.

ξ	MLE	PMLE	Bayes
– 0.4	91%	89%	89%
– 0.2	81%	79%	91%
0	75%	67%	91%
0.2	81%	79%	89%
0.4	86%	80%	94%

Table 3. Estimates of the shape parameter and the 100 year return level for 5 UK sites, using a subset of the annual maximum data, of 15 years. Span denotes the total length of data available at the site, block denotes the span of data used, $\hat{\xi}$ is the shape parameter estimate, and $\hat{z}_p$ is the 100 year return level (or 99% annual maximum quantile) estimate. The final three columns show the return level estimates obtained by using all of the available data at the site.

			$\hat{\xi}$			$\hat{z}_p$ for 15 yr block			$\hat{z}_p$ for full data		
Site	Block	Span	MLE	PWM	PL	MLE	PWM	PL	MLE	PWM	PL
Dover	1941–1955	62	0.74	0.16	0.37	7.84	5.42	5.04	4.73	4.89	4.70
Tilbury	1934–1950	46	1.03	− 0.03	0.09	16.54	5.27	4.95	4.97	5.12	4.93
Southend	1944–1960	57	0.73	0.03	0.25	6.80	4.68	4.42	4.46	4.59	4.44
Grt Yarmouth	1939–1954	77	1.19	0.32	0.46	23.37	4.43	3.98	3.33	3.36	3.29
Immingham	1964–1979	69	1.03	0.02	0.32	12.55	5.47	5.25	4.81	4.89	4.81

The results for each site are similar. For example, at Great Yarmouth, the period 1939 to 1954 leads to a very high MLE of ξ, and a correspondingly implausible 100-year return level estimate (or 99% annual maximum quantile) of over 23 meters. In contrast, the PWM and PL estimates for the span 1939 to 1954 are around 4 m, which is much closer to the estimate of around 3.3 m based on the full 77-year data span.

5.2. *Bias-variance trade-off*

Application of the three estimation methods to each of the 58 data sites, this time using the entire data record at each site, also illustrates the differing degrees of bias-variance trade-off in the three estimation schemes identified in the simulation study. Figure 5 shows the estimates of the 100 year return level from PWM plotted against the corresponding PL and ML for each site. In general the PWM yields slightly larger estimates than the MLE or PMLE. The noticeable exception is for North Shields which has respective parameter estimates of 0.70, 0.25, and 0.49, under ML,PWM, and PL, with corresponding 100-year return level estimates of 5.50 m 4.28 m and 4.33 m.

This pattern of behavior is in accord with our simulation studies. The average sample size is around 40, while the mean of the shape parameter estimates for the 58 sites is − 0.09. Figure 4 shows that for these values, the PWM tends to overestimate quantiles slightly, while for smaller samples the ML has a large positive bias as a result of occasional, but substantial, overestimation, as appears to be the case for North Shields. In contrast, the PWM and PL methods seemingly avoid the large bias at North Shields, but in the case of PWM it is at the expense of a small positive bias for most sites. Of course, these conclusions are tentative since the true parameter values are unknown. However, spatial estimates based on tide-surge decomposition using substantially more data than the annual maxima are available in Dixon and Tawn (1997) for each of these sites—we can again regard these estimates as a ''gold standard'' for comparison. On this basis the ''mean-square bias'' for the PMLE is 0.0570 compared with 0.0633 for estimates based on PWM; thus, the PWM estimates appear to have a greater propensity to bias than the PMLEs.

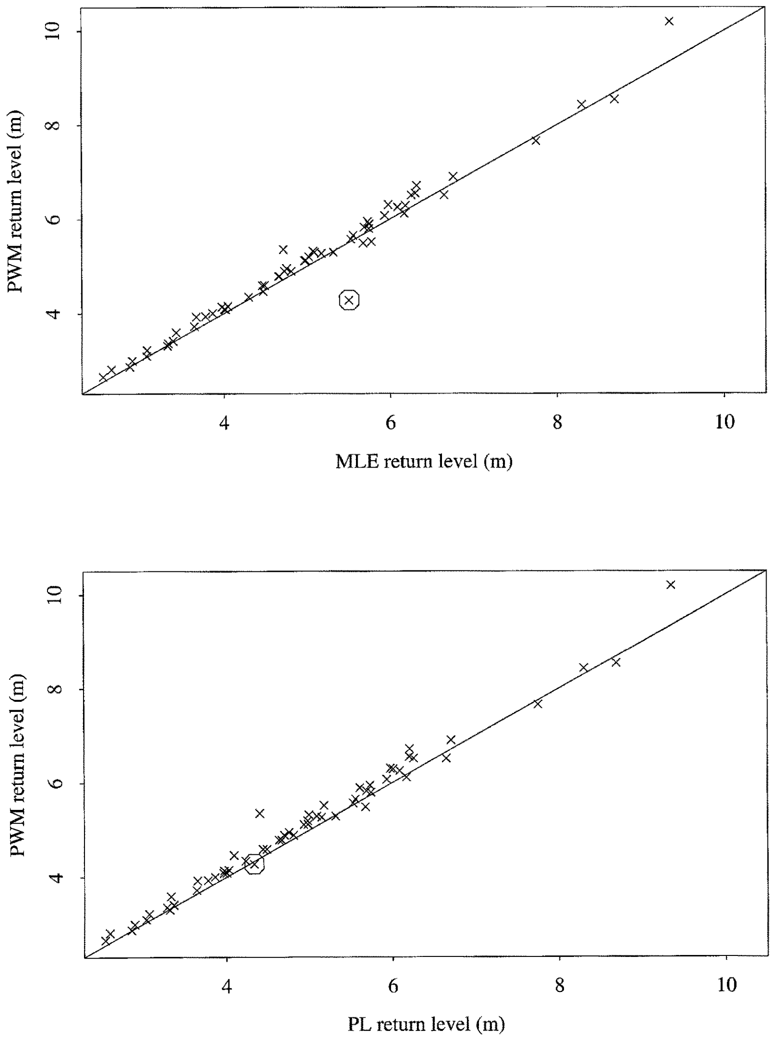

Figure 5. Return level estimates for 58 UK annual maximum data sites obtained under the PWN plotted against estimates obtained using the ML and PL estimators. The estimates for North Shields have been circled. The lines $y = x$ are also given on the figures.

5.3. Covariates

In actual fact, the sea-level data contain long-term trends which have been modeled by Coles and Tawn (1990) and Dixon and Tawn (1992) by extending the GEV model to regress the location parameter linearly with time:

$$Z_t \sim GEV(a + bt, \sigma, \xi).$$

Fitting this model is straightforward for ML and PL; we simply maximize the (penalized) likelihood with respect to the four parameters. Corresponding PWM equations can also be specified for this four parameter model; Scarfe (1992) has done this for a differently parametrized model. However, solution of the equations is not straightforward, and the approximate linearization which simplified the three-parameter model is no longer available. More general covariate relationships would involve greater complexity still.

A more naive method to fit the model consists of preliminary removal of the trend using ordinary least-squares, followed by a probability weighted moments fit of the residuals. Thus, $\hat{b}$ is a least-squares estimate of b and a, σ and ξ are obtained as PWM estimates of the GEV fitted to $Z_t^* = Z_t - \hat{b} \times t, t = 1, \ldots n$. We have compared this method, which amounts to simple linear regression for b, with the four-parameter MLE and PMLE estimates of trend, for each window of length 20 years in the sea-level data, giving rise to 1520 analyzes. Figure 6 shows the ML trend estimates plotted against the PWM (i.e. the least squares estimate) and the PL estimates. The high degree of scatter in points in Figure 6a indicates that for this small sample size, the least-squares procedure can give very different trend estimates compared with ML parameter estimation based on the GEV. In contrast, the strong linearity in Figure 6b suggests that PL will provide trend estimates of comparable accuracy to those of ML.

A referee has pointed out that the PWM estimates could be improved by iteration. This seems most natural to achieve by assuming that $\tilde{Z}_t = Z_t - \hat{b} \times t$ are independently distributed as GEV($(\hat{a}, \hat{\sigma}, \hat{\xi})$, where $\hat{a}, \hat{\sigma}, \hat{\xi}$ are the PWM estimates, and deriving a one-parameter likelihood for b. This leads to a new estimate $\hat{b}$ of b. Then, the whole procedure may be iterated to convergence. We have tried this and found that the final estimate of b is indeed very similar to that obtained using the four-parameter likelihood models. Figure 7 illustrates the various estimates obtained for the Sheerness data, which is circled in Figure 6a. From these data alone, it is not clear which trend estimate is to be preferred. However, based on a spatial analyzes of these data, Coles and Tawn (1990) obtain a similar trend estimate to that given by MLE. The point seems to be that an iterated PWM algorithm performs similarly to the likelihood analyzes, but that this methodology itself requires a likelihood formulation at each iteration of the algorithm. In contrast, the naive least squares estimate can lead to quite misleading trend estimates.

6. Conclusions

Our main point is not to promote one estimator in favor of another in terms of its detailed sampling properties. Our preference for likelihood-based methods is based less on sampling properties than on their flexibility for modeling the types of data structure normally encountered in environmental processes: non-stationarity, temporal dependence, covariate effects, for example. However, what we have shown in this paper is that even within standard ''textbook'' criteria for assessing the performance of estimators, methods based on the likelihood are competitive, even with small sample sizes. In particular, the penalized estimator we have proposed has both computational simplicity and the facility to handle complex data structures in the same way as the maximum

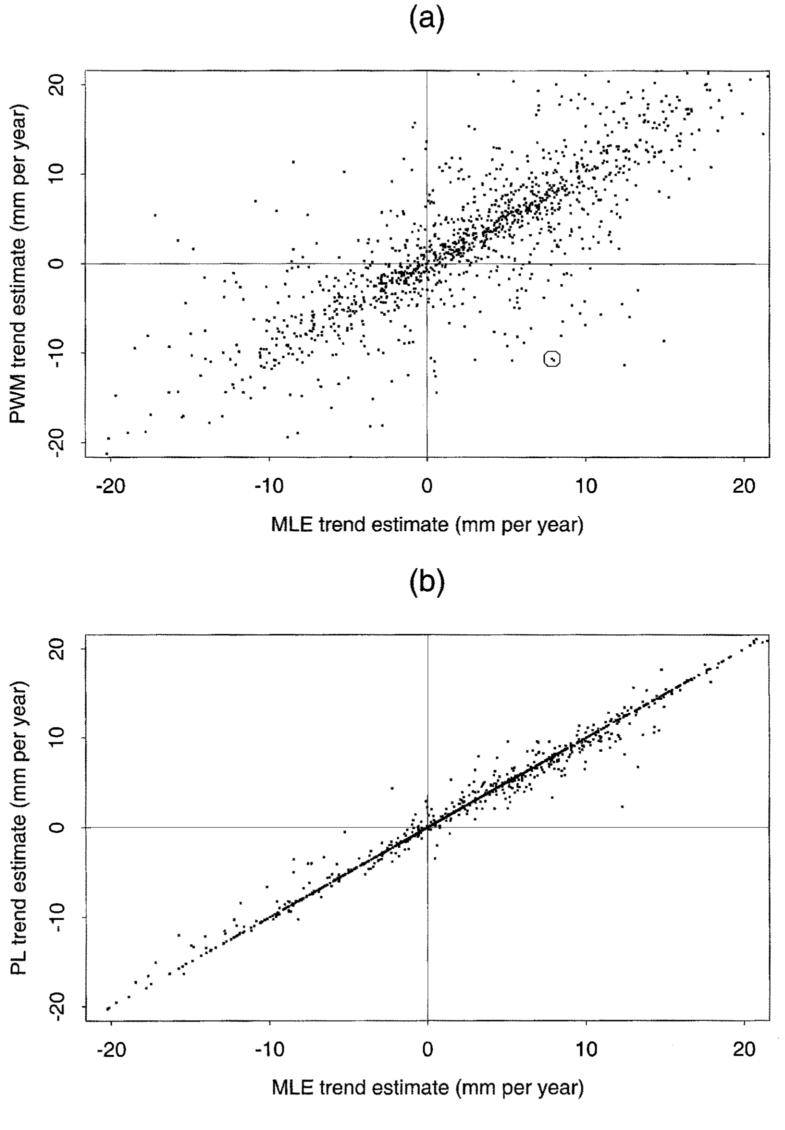

Figure 6. Trend estimates for the UK sea-level data, for all blocks of 20 consecutive years. (a) Trend estimates obtained using standard linear regression against the ML GEV trend estimates. The circled point denotes the Sheerness data for 1900–1920. (b) PL trend estimates against the ML GEV trend estimates.

likelihood estimator, but has small sample properties which out-perform those of the probability weighted moments estimator.

Application to 58 UK data sites reinforces our conclusions: penalized likelihood appears to improve upon both maximum likelihood and probability weighted moments,

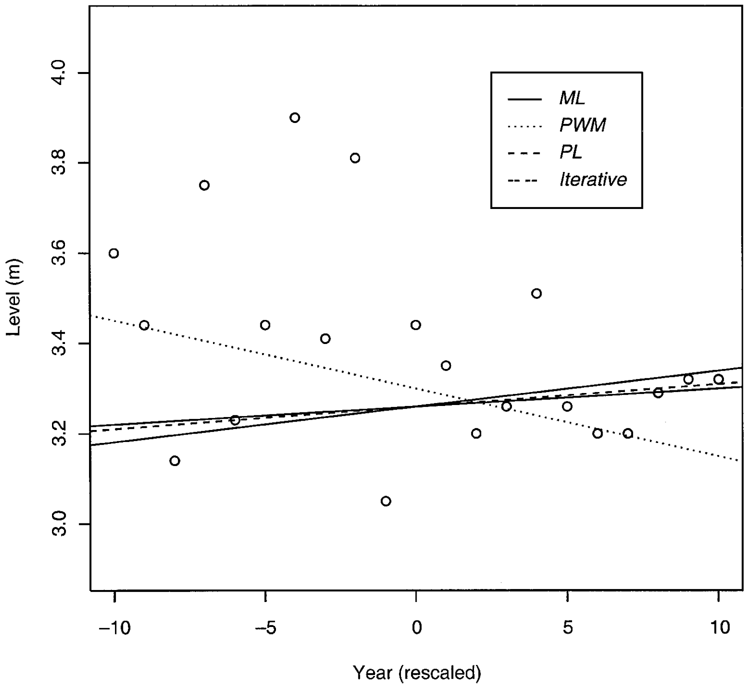

Figure 7. Trend estimates for Sheerness data, years 1900–1920, using the ML and the PL methods, standard linear regression, and a hybrid interative scheme.

and is as easy to apply as maximum likelihood. This is especially important in covariate settings, such as our simple case of a linear trend in sea-levels.

Acknowledgments

We thank Jonathan Tawn for many helpful discussions, Richard Smith for comments on an earlier version, and referees and editors for much constructive criticism. MJD was supported by a MAFF grant.

References

[1] Cohen, J.P., ''The Asymptotic Behavior of Maximum Likelihood Estimates for Univariate Extremes,'' In Tiago de Oliveira (ed.), *Statistical Extremes and Applications*, Reidel, Dordrecht, pp. 435–442, 1984.

[2] Cohen, A.C. and Whitten, B.J., ''Modified Maximum Likelihood and Modified Moment Estimators for the Three-Parameter Weibull Distribution,'' *Comm. Stat. Theor. Meth.* A11, 2631–2656, (1982).

[3] Coles, S.G. and Tawn, J.A., ''Statistics of Coastal Flood Prevention,'' *Phil. Trans. R. Soc. Lond. A* 332, 457–476, (1990).

[4] Coles, S.G. and Tawn, J.A., ''A Bayesian Analysis of Extreme Rainfall Data,'' *Appl. Statist.* 45, 463–478,

(1996).

[5] Cox, D.R. and Hinkley, D.V., *Theoretical Statistics*, Chapman and Hall, London, 1974.

[6] Davison, A.C. and Smith, R.L., ''Models for Exceedances Over High thresholds (with Discussion),'' *J. Roy. Statist. Soc. B* 52, 393–442, (1990).

[7] Dixon, M.J. and Tawn, J.A., ''Trends in U.K. Extreme Sea Levels: a Spatial Approach,'' *Geophys. J. Int.* 111, 607–616, (1992).

[8] Dixon, M.J. and Tawn, J.A., *Estimates Of Extreme Sea Conditions: Spatial Analyzes for the UK*, Proudman Oceanographic Laboratory Internal document: in press, 210 pages, 1997.

[9] Fisher, R.A. and Tippett, L.H.C., ''On the Estimation of the Frequency Distributions of the Largest or Smallest Member of a Sample,'' *Proc. Camb. Phil. Soc.* 24, 180–190, (1928).

[10] Graff, J., ''An Investigation of the Frequency Distributions of Annual Sea-Level Maxima at Ports Around Great Britain,'' *Estuarine Coastal Shelf Sci.* 12, 389–449, (1981).

[11] Hosking, J.R.M., ''Algorithm AS215: Maximum Likelihood Estimation of the Parameters of the Generalized Extreme-Value Distribution,'' *Appl. Statist.* 34, 301–310, (1985).

[12] Hall, P., Peng, L., and Tajvidi, N., *On prediction intervals based on predictive likelihood or bootstrap methods*, Submitted. 1999.

[13] Hosking, J.R.M., Wallis, J.R., and Wood, E.F., ''Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments,'' *Technometrics* 27, 251–261, (1985).

[14] Hosking, J.R.M. and Wallis, J.R., ''Parameter and Quantile Estimation for the Generalized Pareto Distribution,'' *Technometrics* 29, 339–349, (1987).

[15] Jenkinson, A.F., ''The Frequency Distribution of the Annual Maximum (or minimum) of Meteorological Events,'' *Q. J. R. Meteorol. Soc.* 81, 158–171, (1955).

[16] Landwehr, J.M., Matalas, N.C., and Wallis, J.R., ''Probability Weighted Moments Compared with Some Traditional Techniques for Estimating Gumbel Parameters and Quantiles,'' *Wat. Res. Res.* 15, 1055–1064, (1979).

[17] Leadbetter, M.R., Lindgren, G., and Rootzen, H., *Extremes and Related Properties of Random Sequences and Series*, Springer Verlag, New York, 1983.

[18] Prescott, P. and Walden, A.T., ''Maximum Likelihood Estimation of the Parameters of the Generalized Extreme Value Distribution,'' *Biometrika* 67, 723–724, (1980).

[19] Scarf, P.A., ''Estimation for a Four Parameter Generalized Extreme Value Distribution,'' *Communications in Statistics—Theory and Methods* 21, 2185–2201, (1992).

[20] Silverman, B.W., ''Penalized Maximum Likelihood Estimation,'' In *Encyclopedia of Statistical Sciences*, 6, Ed. S. Kotz and N. L. Johnson, 664–667, 1985.

[21] Prescott, P. and Walden, A.T., ''Maximum Likelihood Estimation of the Parameters of the Three-Parameter Generalized Extreme-Value Distribution from Censored Samples,'' *J. Statist. Comput. Simul.* 16, 241–250, (1983).

[22] Smith, R.L., ''Maximum Likelihood Estimation in a Class of Non-Regular Cases,'' *Biometrika* 72, 67–92, (1985).

[23] Smith, R.L., ''Maximum Likelihood Estimation for the NEAR(2) Model,'' *J. R. Statist. Soc. B* 48, 251–257, (1986).

[24] Smith, R.L., ''Extreme Value Analysis of Environmental Time Series: an Application to Trend Estimation in Ground Level Ozone,'' *Statist. Sci.* 4, 367–393, (1989).

[25] Smith, R.L., ''Extreme Value Theory,'' In *Handbook of Applicable Mathematics*, 7, 437–471, ed. W. Ledermann, John Wiley, Chichester, 1990.

[26] Smith, R.L. and Naylor, J.C., ''A Comparison of Maximum Likelihood and Bayesian Estimators for the Three-Parameter Weibull Distribution,'' *Appl. Statist.* 36, 358–369, (1987).

[27] Smith, R.L. and Weissman, I., ''Maximum Likelihood Estimation of the Lower Tail of a Probability Distribution,'' *J.R. Statist. Soc. B* 47, 285–298, (1985).