

# Mixed Model-based Additive Models for Sample Extremes

BY S. A. PADOAN

*Dipartimento di Statistica,  
Università di Padova, via Cesare Battisti 241/243, 35100 Padova, ITALY*

AND M.P. WAND

*School of Mathematics and Applied Statistics,  
University of Wollongong, Northfields Avenue, Wollongong 2522, AUSTRALIA*

28th March, 2008

ABSTRACT

We consider additive models fitting and inference when the response variable is a sample extreme. Non-linear covariate effects are handled using the mixed model representation of penalised splines. A fitting algorithm based on likelihood approximations is derived. The efficacy of the resulting methodology is demonstrated via application to simulated and real data.

*Some keywords:* Generalised extreme value distribution; Generalised linear mixed models; Modified profile likelihood; Penalised likelihood; Penalised splines; Semiparametric regression.

## 1 Introduction

Statistical modelling of extreme values has flourished since the mid-1980s. One of many motivating examples is extreme climate events (e.g. maximum annual temperatures) that are possibly linked with global warming. The book of Coles (2001) provides a comprehensive introduction to this topic. The generalised extreme value distribution has emerged as the most common family for modelling such data.

Recent work by, for example, Davison and Ramesh (2000), Chavez-Demoulin and Davison (2005) and Yee and Stephenson (2007) has demonstrated the usefulness of non-parametric regression, or smoothing, in extreme value contexts. The first of these papers used a local likelihood approach, while the second used smoothing splines. Chavez-Demoulin and Davison (2005) and Yee and Stephenson (2007) also treated the additive model extension, where the effect of several covariates can be considered simultaneously and flexibly.

In this note we explore an alternative approach to additive model fitting and inference for sample extreme responses. It is based on the mixed model/splines paradigm that has achieved a great deal of success in other contexts during the last decade. Ruppert, Wand and Carroll (2003) provides a summary of this general approach. It is well used and developed when response variable is normally distributed or has a distribution within the exponential family, but has not been explored for extremes. A compelling feature of

this approach is that the smoothing parameters correspond to variance components, so maximum likelihood or Bayesian techniques can be applied for model fitting, assessment and inference. There is no need for secondary procedures of such as cross-validation to choose smoothing parameters. Another advantage is that complications such as spatial or temporal correlation, missing data and measurement error are more easily incorporated.

There are a number of routes that can be taken to perform fitting and inference for mixed model-based additive models. Here we explore the simplest and fastest: approximate maximum likelihood. Consequently, this work represents an extension of the so-called Penalised Quasi-Likelihood (PQL) ideas of e.g. Breslow and Clayton (1993) to extreme responses.

Section 2 lays out the model and fitting and inference strategy. Some simulation results, that demonstrates good practical performance, are presented in Section 3. Application to maximum temperatures data is described in Section 4. An appendix provides mathematical details.

## 2 Mixed model-based additive models for sample extremes

### 2.1 Non-stationary extremes sequences

Let  $X_1, \dots, X_n$  be an independent and identically distributed (i.i.d.) set of random variables and let  $M_n = \max(X_1, \dots, X_n)$  denote the sample maximum. Then the limiting distribution as  $n \rightarrow \infty$  of  $(M_n - a_n)/b_n$  (if such a sequences of constants  $\{b_n > 0\}$  and  $\{a_n\}$  exist) must be a member of the *generalised extreme value (GEV)* family of distributions (e.g. von Mises, 1954; Jenkinson, 1955). A random variable  $Y$  has a GEV distribution, denoted by  $Y \sim \text{GEV}(\mu, \psi, \xi)$  if its cumulative distribution function is given by:

$$F(y; \mu, \psi, \xi) = \exp \left[ - \left\{ 1 + \xi \left( \frac{y - \mu}{\psi} \right) \right\}_+^{-1/\xi} \right], \quad -\infty < \mu, \xi < \infty, \quad \psi > 0,$$

where  $y_+ = \max(0, y)$  and  $\mu, \psi$  and  $\xi$  are respectively location, scale and shape parameters. The GEV distribution may be divided into the following three sub-families: Fréchet distribution (Fisher-Tippett type III) for  $\xi > 0$ , Weibull distribution (Fisher-Tippett type II) for  $\xi < 0$  and Gumbel-type distribution (Fisher-Tippett type I) when  $\xi \rightarrow 0$ ; see Fisher and Tippett (1928).

Now suppose we observe  $n$  sample maxima  $y_1, \dots, y_n$  as well as corresponding covariate vectors  $\mathbf{x}_1, \dots, \mathbf{x}_n$ . The  $y_i$  are obtained from approximately equi-sized samples of a variable of interest. A common situation is  $y_i$  corresponding to the annual maximum of a daily measurement, such as rainfall in a particular town, for year  $i$  ( $1 \leq i \leq n$ ). General GEV regression models (e.g. Coles, 2001) take the form

$$y_i | \mathbf{x}_i \sim \text{GEV}(\mu(\mathbf{x}_i), \psi(\mathbf{x}_i), \xi(\mathbf{x}_i)), \quad (1)$$

where, for example,  $\mu(\mathbf{x}_i) = g((\mathbf{X}\boldsymbol{\beta})_i)$ ,  $g$  is a link function,  $\boldsymbol{\beta}$  is a vector of regression coefficients and  $\mathbf{X}$  is a design matrix associated with the  $\mathbf{x}_i$ s. Similar structures may be imposed upon  $\psi(\mathbf{x}_i)$  and  $\xi(\mathbf{x}_i)$ . The regression coefficients can be estimated via maximum likelihood. Davison and Ramesh (2000) and Hall and Tajvidi (2000) argue that parametric models for (1) can be too restrictive, and have advocated non-parametric approaches. Other contributions to smoothing for sample extremes include Pauli and Coles (2001) and Chavez-Demoulin and Davison (2005), the latter incorporating generalised additive model structure.

### 2.2 Generalised extreme value mixed model-based splines

Ruppert *et al.* (2003) and Wand (2003) have discussed how penalised splines can be carried out in a mixed model framework for Gaussian and exponential family models. Here we

focus on the GEV case.

Assuming that the location parameter in the GEV distribution is smooth on an interval  $[a, b]$  in the  $x_i$  domain then the simplest time-nonhomogeneous nonparametric regression model is given by

$$y_i \sim \text{GEV}(\mu(x_i), \psi, \xi). \quad (2)$$

Mixed model-based penalised spline models for  $\mu$  take the general form

$$\eta(x) \equiv g\{\mu(x)\} = \beta_0 + \beta_1 x + \sum_{k=1}^K u_k z_k(x); \quad u_1, \dots, u_K \text{ i.i.d. } N(0, \sigma^2)$$

where  $g$  is a link function and  $z_1, \dots, z_K$  is an appropriate set of spline basis functions. The simplest version is  $z_k(x) = (x - \kappa_k)_+$ , where  $\kappa_1, \dots, \kappa_K$  is a dense set of knots within the range of the  $x_i$ 's. More sophisticated basis functions are recommended for consideration. See, for example, Welham, Cullis, Kenward and Thompson (2006) and Wand and Ormerod (2008). The latter reference describes the  $z_k$  corresponding to the  $\mathbb{R}$  function `smooth.spline()`. The choice of  $K$  has a secondary effect and, for many signals, about 20 knots are sufficient.

Let  $\mathbf{y} = (y_1, \dots, y_n)$  and define the design matrices

$$\mathbf{X} = [1 \ x_i]_{1 \leq i \leq n}, \quad \mathbf{Z} = [z_k(x_i)]_{\substack{1 \leq i \leq n \\ 1 \leq k \leq K}}$$

associated with fixed effects  $\boldsymbol{\beta} = [\beta_0 \ \beta_1]^T$  and random effects  $\mathbf{u} = [u_1, \dots, u_K]^T$ . Given  $\mathbf{u}$ , the  $y_i$  are conditionally independent with distribution,

$$y_i | \mathbf{u} \sim \text{GEV}(g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i, \psi, \xi),$$

Note that  $\boldsymbol{\mu} \equiv g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})$  is related to the conditional mean of  $\mathbf{y}$  given  $\mathbf{u}$  via

$$E(\mathbf{y} | \mathbf{u}) = \begin{cases} \boldsymbol{\mu} + \mathbf{1}\psi\{\Gamma(1 - \xi) - 1\}/\xi, & \text{for } \xi \neq 0 \\ \boldsymbol{\mu} + \mathbf{1}\psi\gamma, & \text{for } \xi = 0 \end{cases}$$

where  $\mathbf{1}^T = (1, \dots, 1)$  is a vector of  $n$  one values,  $\Gamma$  is the Gamma function and  $\gamma = 0.57721566 \dots$  is Euler's constant.

Let  $\mathbf{C} = [\mathbf{X} | \mathbf{Z}]$  be the matrix obtained combining the columns of design matrices  $\mathbf{X}$  and  $\mathbf{Z}$ , and with vector  $\boldsymbol{\nu} = [\boldsymbol{\beta}^T \ \mathbf{u}^T]^T$  the  $K + 2$  coefficients of fixed and random effects. With this notation the conditional probability density function of  $y_i | \mathbf{u}$  and the probability density function of  $\mathbf{u}$  random effects have the expressions

$$f(\mathbf{u}; \sigma^2) = (2\pi\sigma^2)^{-K/2} \exp\left(-\frac{\|\mathbf{u}\|^2}{2\sigma^2}\right) \quad \text{and}$$

$$f(\mathbf{y} | \mathbf{u}; \boldsymbol{\beta}, \psi, \xi) = \prod_{i=1}^n \frac{1}{\psi} \left\{ 1 + \xi \left( \frac{(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i}{\psi} \right) \right\}^{-\frac{1}{\xi} - 1} \exp \left[ - \left\{ 1 + \xi \left( \frac{(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i}{\psi} \right) \right\}^{-\frac{1}{\xi}} \right].$$

The norm for fitting (2) is estimation of the parameters via maximisation of the likelihood:

$$\mathcal{L}(\boldsymbol{\beta}, \psi, \xi, \sigma^2) = f(\mathbf{y}; \boldsymbol{\beta}, \psi, \xi, \sigma^2) = \int_{\mathbb{R}^K} f(\mathbf{y} | \mathbf{u}; \boldsymbol{\beta}, \psi, \xi) f(\mathbf{u}; \sigma^2) d\mathbf{u} \quad (3)$$

and prediction of the random effects via the best predictor  $\hat{\mathbf{u}} = E(\mathbf{u} | \mathbf{y})$ . However, both involve intractable integrals over  $\mathbb{R}^K$ . We propose to overcome this hinderance by appealing to the ideas of penalised log-likelihood (Green, 1987; Breslow and Clayton, 1993). This involves application of Laplace's method to approximate (3). The set of coefficients

$\boldsymbol{\nu}$  is then treated as a parameter vector, but with the random effects vector  $\mathbf{u}$  being penalised according to the restriction  $\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ . The resulting penalised log-likelihood is

$$\begin{aligned} \ell_{\text{PL}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2) &= \log\{f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \psi, \xi)\} + \log\{f(\mathbf{u}; \sigma^2)\} \\ &= -\frac{1+\xi}{\xi} \mathbf{1}^T \log \left\{ \mathbf{1} + \xi \left( \frac{\mathbf{y} - \mathbf{C}\boldsymbol{\nu}}{\psi} \right) \right\} - \mathbf{1}^T \left\{ \mathbf{1} + \xi \left( \frac{\mathbf{y} - \mathbf{C}\boldsymbol{\nu}}{\psi} \right) \right\}^{-\frac{1}{\xi}} \\ &\quad - n \log(\psi) - \frac{K}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \|\mathbf{u}\|^2. \end{aligned}$$

Even though  $\ell_{\text{PL}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2)$  is not a bona fide log-likelihood, we will treat it as such in the ensuing discussion. Since it involves multivariate parameters of different types, its direct maximisation is numerically challenging. Also, it can lead to unreliable estimators when  $\boldsymbol{\nu}$  is high-dimensional. We instead propose use of an iterative scheme in which estimates of components of the parameter vector are updated while keeping the other components fixed. Naïve updating would involve successive maximisation of profile likelihoods. However, since profile likelihoods are not genuine likelihoods, inferences based on their maximisation can be misleading (see e.g. Severini, 1998). This has led to several refinements of profile likelihood to address this problem. The most widely accepted is modified profile log-likelihood (Barndorff-Nielsen & Cox, 1994). We therefore propose to estimate  $(\boldsymbol{\nu}, \psi, \xi, \sigma^2)$  through successive maximisation of modified profile log-likelihoods. A similar scheme is used by Breslow & Clayton (1993) in the case of exponential family mixed models.

The modified profile log-likelihood of  $(\psi, \xi, \sigma^2)$  is

$$\ell_{\text{PL}}(\hat{\boldsymbol{\nu}}_{\psi, \xi, \sigma^2}, \psi, \xi, \sigma^2) - \frac{1}{2} \log |\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\hat{\boldsymbol{\nu}}_{\psi, \xi, \sigma^2}, \psi, \xi, \sigma^2)| + |\mathbf{D}_{\hat{\boldsymbol{\nu}}_{\psi, \xi, \sigma^2}} \hat{\boldsymbol{\nu}}|$$

where  $\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\hat{\boldsymbol{\nu}}_{\psi, \xi, \sigma^2})$  is the observed information matrix for  $\boldsymbol{\nu}$  for fixed  $(\psi, \xi, \sigma^2)$  at the corresponding maximum penalised likelihood estimate  $\hat{\boldsymbol{\nu}}_{\psi, \xi, \sigma^2}$  and the D notation for vector differentiation is defined in Section A.1. Because of approximate orthogonality between trend and scale/shape in the GEV distribution (Tawn, 1988), the last term is asymptotically negligible. Henceforth, we work with the approximation

$$\ell_{\text{M}}(\psi, \xi, \sigma^2) = \ell_{\text{PL}}(\hat{\boldsymbol{\nu}}_{\psi, \xi, \sigma^2}, \psi, \xi, \sigma^2) - \frac{1}{2} \log |\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\hat{\boldsymbol{\nu}}_{\psi, \xi, \sigma^2}, \psi, \xi, \sigma^2)|.$$

An explicit expression for  $\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\hat{\boldsymbol{\nu}}_{\psi, \xi, \sigma^2})$  is given in Appendix A.2.

The modified profile log-likelihood of  $\sigma^2$  is approximately (using, again, an approximate orthogonality argument)

$$\begin{aligned} \ell_{\text{M}}(\sigma^2) &= \ell_{\text{PL}}(\hat{\boldsymbol{\nu}}_{\psi, \xi, \sigma^2}, (\hat{\psi}, \hat{\xi})_{\sigma^2}, \sigma^2) - \frac{1}{2} \log |\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\hat{\boldsymbol{\nu}}_{(\hat{\psi}, \hat{\xi})_{\sigma^2}, \sigma^2}, (\hat{\psi}, \hat{\xi})_{\sigma^2}, \sigma^2)| \\ &\quad - \frac{1}{2} \log |\mathcal{I}_{(\psi, \xi)(\psi, \xi)}((\hat{\psi}, \hat{\xi})_{\sigma^2}, \sigma^2)| \end{aligned}$$

where  $\mathcal{I}_{(\psi, \xi)(\psi, \xi)}((\hat{\psi}, \hat{\xi})_{\sigma^2}, \sigma^2)$  is the observed information matrix for  $(\psi, \xi)$  for fixed  $\sigma^2$  at the corresponding maximum modified profile likelihood estimate  $(\hat{\psi}, \hat{\xi})_{\sigma^2}$ . An explicit expression for  $\mathcal{I}_{(\psi, \xi)(\psi, \xi)}((\hat{\psi}, \hat{\xi})_{\sigma^2}, \sigma^2)$  is given in Appendix A.3.

We thus propose to estimate the parameters  $\boldsymbol{\beta}, \psi, \xi, \sigma^2$  and random effects  $\mathbf{u}$  via the iterative scheme:

1. Set starting values:  $\hat{\boldsymbol{\nu}}, \hat{\psi}, \hat{\xi}, \hat{\sigma}^2$ .
2. Update  $\hat{\boldsymbol{\nu}}$  by maximising the penalised log-likelihood  $\ell_{\text{PL}}(\boldsymbol{\nu}, \hat{\psi}, \hat{\xi}, \hat{\sigma}^2)$ .
3. Update  $(\hat{\psi}, \hat{\xi})$  by maximising the modified profile log-likelihood  $\ell_{\text{M}}(\psi, \xi, \hat{\sigma}^2)$ .

4. Update  $\hat{\sigma}^2$  by maximising the modified profile log-likelihood  $\ell_M(\sigma^2)$ .
5. Repeat steps 2–4 until convergence.

The maximum penalised likelihood estimate of  $\nu$  can be obtained by the Newton-Raphson method with substitution of the observed information by the expected information matrix (Prescott and Walden, 1980). The required derivatives are given in Appendix A.2.

The estimates of  $(\psi, \xi)$  and  $\sigma^2$  can be obtained using a quasi-Newton numerical maximisation routine (e.g. Broyden, 1967). However, as argued by Smith (1985), asymptotic likelihood results for the GEV distribution are subject to restrictions.

### 2.3 Additive models extension

We now consider the extension where several variates may impact on the sample extremes  $y_1, \dots, y_n$ . If  $\mathbf{x}_i$  is  $d$ -variate then

$$g\{\mu(\mathbf{x}_i)\} = \eta(\mathbf{x}_i) = f_1(x_{i1}) + \dots + f_d(x_{id})$$

defines a general additive model for the  $\mu(\mathbf{x}_i)$ . Here the  $f_j$  are general smooth functions. The mixed model-based penalised splines of Section 2.2 can accommodate this extension by setting

$$\mathbf{X} = [1 \ \mathbf{x}_{i1} \cdots \ \mathbf{x}_{id}]_{1 \leq i \leq n},$$

and  $\mathbf{Z}$  defined similarly, with suitable spline basis functions. Also,

$$\text{Cov}(\mathbf{u}) = \mathbf{G}_{\sigma^2} \equiv \text{blockdiag}(\sigma_1^2 \mathbf{I}_{K_1}, \dots, \sigma_d^2 \mathbf{I}_{K_d}),$$

with  $K_j$  the number of spline basis functions used for  $f_j$ . The fitting procedure described in the previous section is basically the same, but with longer  $\beta$  and  $\mathbf{u}$  vectors and  $\sigma^2$  replaced by the vector  $\sigma^2 = (\sigma_1^2, \dots, \sigma_d^2)$ .

## 3 Simulation study

We investigated the performance of mixed model-based penalised splines for extremes with a simulation study. Let

$$\mu(x) = 2x + \cos(4\pi x), \quad 0 \leq x \leq 1.$$

Data was generated in two steps. Firstly a sample  $x_1, \dots, x_n$  was drawn from a uniform distribution on  $(0, 1)$ . Secondly, given the  $x_i$ 's,  $n$  realisations were drawn according to  $y_i \sim \text{GEV}(\mu(x_i), \psi, \xi)$ . The shape parameter  $\xi$  was set to  $-0.4, 0$  and  $0.4$  corresponding to the three different types of GEV distributions. Also, different values of the scale parameters were considered. We performed 500 data replications for each configuration. In each case estimation was performed using the likelihood-based algorithm of Section 2.

Results for estimation of the scalar parameters are summarised in Table 1. The estimates are seen to be reasonably accurate for all three-type distributions even though the estimation method involves approximated likelihood functions. With the Fréchet distribution we observe a bias on the shape parameter for sample size equal to 100. This bias effect is a consequence of the approximated likelihood functions. A bias can arise from an inadequacy of Laplace's approximation integral when used with the heavy tail distribution (Fréchet). However, for larger sample size the variability of estimates distribution decrease gradually in accordance with the standard asymptotic likelihood theory.

Distribution	$n$	$\hat{\psi}$	$\hat{\xi}$	$\hat{\sigma}$
Fréchet	500	0.611 (0.029)	0.398 (0.040)	9.6 (1.0)
-	100	0.645 (0.073)	0.365 (0.099)	9.1 (2.1)
	<i>true</i>	0.6	0.4	
Gumbel	500	0.401 (0.014)	-0.004 (0.033)	9.1 (0.6)
-	100	0.403 (0.036)	-0.021 (0.078)	9.7 (1.1)
	<i>true</i>	0.4	0	
Weibull	500	0.500 (0.018)	-0.397 (0.027)	9.5 (0.8)
-	100	0.509 (0.055)	-0.406 (0.085)	9.8 (1.9)
	<i>true</i>	0.5	-0.4	

Table 1: *Smoothing and nuisance parameters estimates of three-types: the second column indicates which smoothing function is used while the third column indicates the sample size. From columns 4–6 GEV scale, shape and variance components estimates are given. Standard errors are in brackets.*

Figure 1 conveys the performance of the function estimation component. The estimates showed correspond to the 10th, 50th and 90th percentiles of the replication-wise deviance measures; given by

$$D(\boldsymbol{\mu}; \hat{\boldsymbol{\mu}}) = 2\{\ell_{\boldsymbol{\mu}}(\boldsymbol{\mu}) - \ell_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}})\},$$

where  $\ell_{\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}})$  is log-likelihood computed for  $y_i = \hat{\mu}_i$ . For the larger sample sizes the fitted curves approximately match the true curves for all three percentiles and distributions. With sample size 100 the results are still acceptable. In addition, we notice for the Fréchet case how a lack of accuracy for the fitted curve corresponding to the 90th percentile is expected by the nature of the heavy-tailed distribution.

## 4 Application to English temperature data

In this section we consider the maximum Central England Temperature. The data-set consists of daily maximum temperatures representative of a roughly triangular area of England enclosed by Lancashire, London and Bristol recorded from 1878 to 2006. The maxima correspond to a time period of one year so are based on equi-sized samples. The annual mean of the North Atlantic Oscillation and Southern Oscillation Index are also considered. The North Atlantic Oscillation index measures the difference of mean atmospheric sea-level pressures near the Azores and near Iceland. The Southern Oscillation Index measures the difference of mean atmospheric sea-level pressures near Tahiti and Darwin, Australia. All three daily and monthly series are available on the web respectively from: <http://hadobs.metoffice.com/hadcet/index.html> and <http://www.cru.uea.ac.uk/cru/data/pci.htm>. A large amount of literature has established that the North Atlantic Oscillation has climatic effects on European and North American winters and Southern Oscillation on Australia’s climate. Because of this, we were curious to see if these two environmental processes could also have an effect on the annual maxima temperature in central England. The aim of this analysis is not to provide an exhaustive investigation of the temperature behaviour, but rather to illustrate how our tools can be used to assess dependence of the extreme on covariates. We first fitted a mixed model-based GEV additive model to maximum annual temperature with smooth functions of time, North Atlantic Oscillation and Southern Oscillation Index as predictors. The identity link function was used. The slope coefficients and variance components for both North Atlantic Oscillation and Southern Oscillation Index were far from significant – there is no evidence in the data that these two variables impact extreme

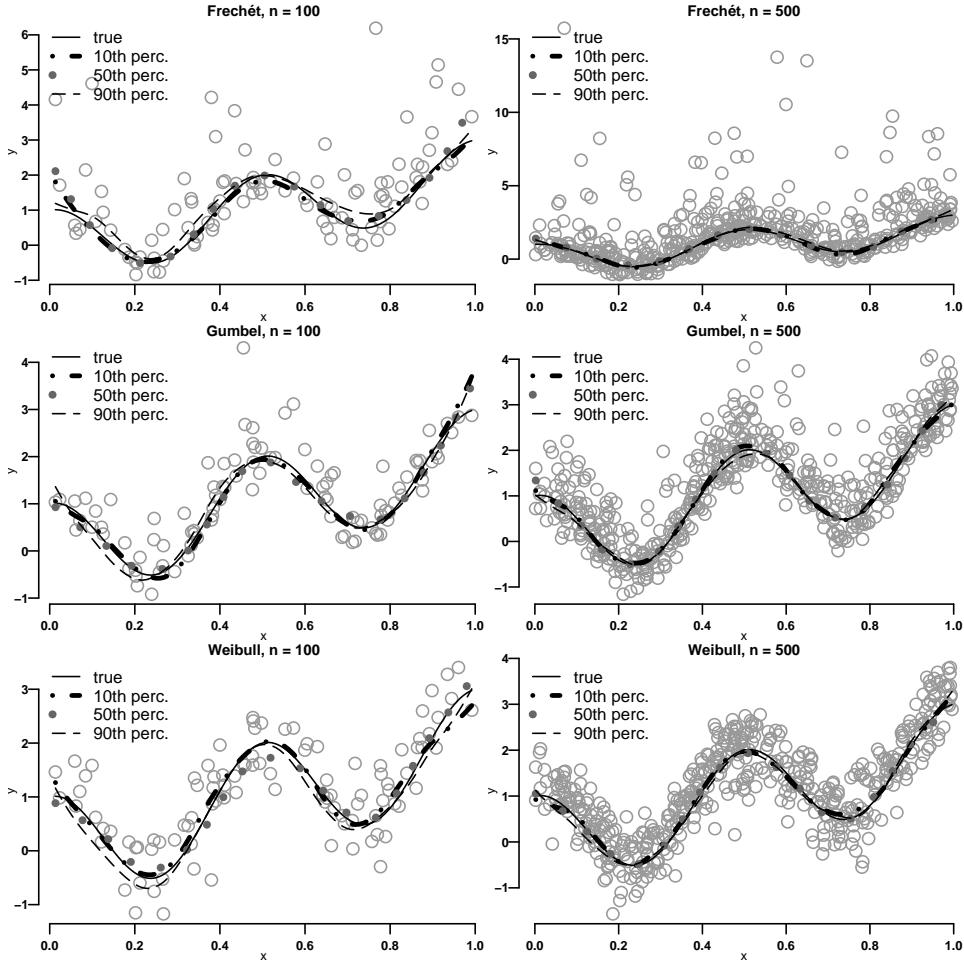


Figure 1: Fitting of the location parameter: the smoothing function  $\mu$  and its fitting for the three-type distributions and different sample sizes are plotted on the panels.

temperature in central England. We then re-fit using time as the only predictor. The estimated mean trend is plotted in Figure 2. The shaded regions are variability bands; see the Appendix for details. From 1978 to 2006 the range of the maxima temperature trend is about 2 degrees Celsius. Initially the trend has increased from nearly 20 degrees to 21 degrees in about 1991. In around the last 15 years the trend has increased again of the same amount arriving at nearly 22 degrees. Thus it seems from this brief analysis that in recent years the maxima temperature trend has been accelerating. The non-mean parameter estimates were  $\hat{\sigma}^2 = 0.055$ ,  $\hat{\psi} = 1.32$  and  $\hat{\xi} = -0.11$ .

We then decided to assess the adequacy of the linear model for the effect of time against the non-linear spline alternative. This involves the hypothesis test,  $H_0 : \sigma^2 = 0$  against  $H_1 : \sigma^2 > 0$ . Note that this is not a trivial problem because, under the null hypothesis, the variance components are on the boundary of their parameter space. For example, the standard test performed within the likelihood ratio paradigm does not have the usual chi-squared distribution but rather a mixture of chi-squareds (e.g. Self and Liang, 1987). Moreover, under the mixed model paradigm, the random effects induce a dependence factor (e.g. Miller, 1977) and the mixture asymptotic distribution does not hold for penalized spline models (Crainiceanu and Ruppert, 2004). In order to test the null hypothesis we used the likelihood ratio test, but for the reasons just discussed we determine the null distribution of the likelihood ratio test statistic by a simulation-based alternative. The critical value of the test has been obtained by Monte Carlo simulation. More precisely, for the data vector  $\mathbf{y}$  and the variance component  $\sigma^2$ , the ratio test statistic

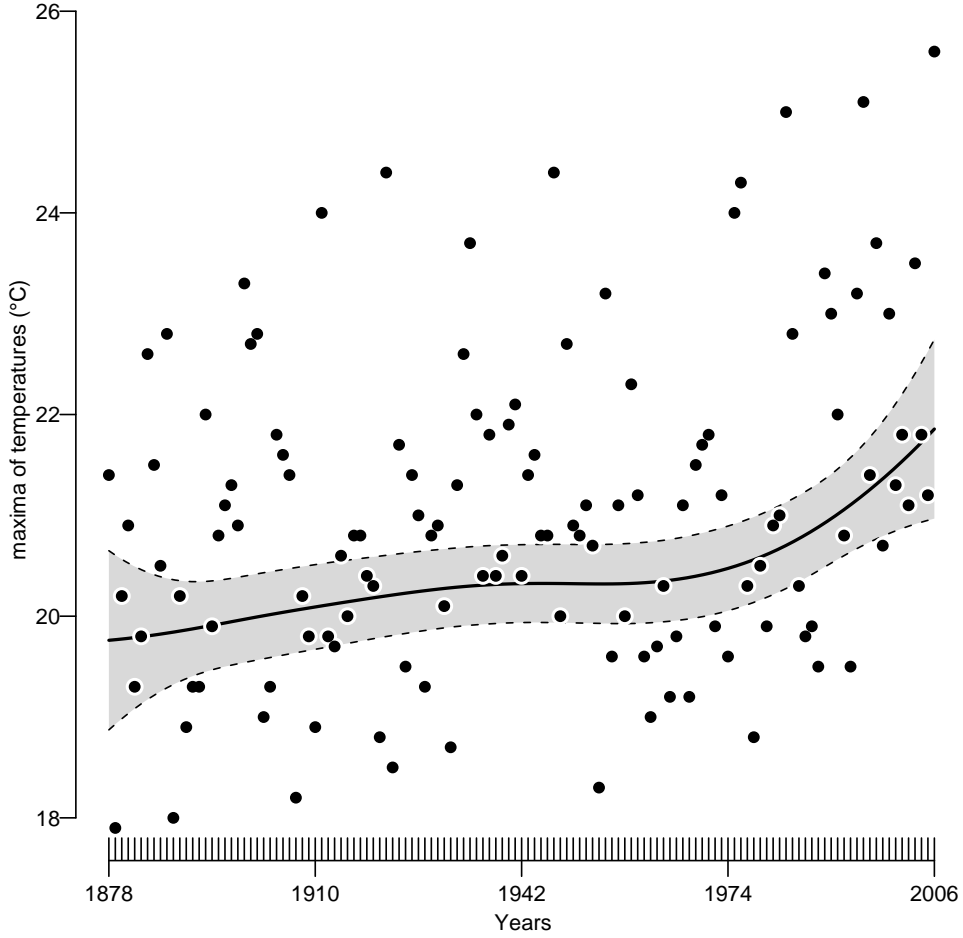


Figure 2: *Central England Temperatures example. The three panels are, respectively, maxima annual temperatures versus time, North Atlantic Oscillation and Southern Oscillation Index. Continuous lines express the fitting trend and the shaded regions are variability bands ( $\pm 2$  estimated standard deviations).*

is

$$\text{LLR}(\mathbf{y}) = 2\{\ell(\hat{\sigma}^2; \mathbf{y}) - \ell(\hat{\sigma}_0^2; \mathbf{y})\}, \quad (4)$$

where  $\hat{\sigma}_0^2$  maximizes the penalized log likelihood under the null hypothesis that the variance components could be removed from the model, and  $\hat{\sigma}^2$  under the alternative. We compute the statistic (4) with the observed data, and denote it by  $\text{LLR}(\mathbf{y}^{\text{obs}})$ . Fixing the model parameter equal to  $\hat{\sigma}_0^{2,\text{obs}}$ , the maximum likelihood estimates obtained under null hypothesis with the observed data, we simulate  $M = 10000$  synthetic data from the spline mixed-model for extremes (under the null hypothesis it consists of a GEV model with linear trend). Then for each simulated data we estimate the smoothing, the GEV and the dispersion parameters according to the models under the null and the alternative hypotheses and so we compute the test statistic  $\text{LLR}(\mathbf{y}^{\text{sim}})$  by using (4). In this way we obtain a sequence of values that simulate the distribution of the likelihood ratio test under the null hypothesis. Finally the  $p$ -value of the test is the proportion of simulated values  $\text{LLR}(\mathbf{y}^{\text{sim}})$  that exceed the statistic computed with the real data. In other words

$$\text{p-value} = \frac{\sum_{m=1}^M I\{\text{LLR}(\mathbf{y}^{\text{sim}}) > \text{LLR}(\mathbf{y}^{\text{obs}})\}}{M},$$

where  $I\{B\}$  is the indicator function of the set  $B$ . Using this simulation-based method we obtained a  $p$ -value of 0.00020. There is very strong evidence in favour of the non-linear spline relationship depicted in Figure 2.

## Appendix

In this appendix we present explicit expressions for the derivatives required for the likelihood-based fitting scheme given in Section 2.2.

### A.1 Vector notation

Let  $f$  be a real-valued function in the  $d \times 1$  vector  $\mathbf{x} = (x_1, \dots, x_d)$ . Then the derivative vector  $D_{\mathbf{x}}f(\mathbf{x})$ , is the  $1 \times d$  with  $i$ th entry  $\partial f(\mathbf{x})/\partial x_i$ . The corresponding Hessian matrix is given by  $H_{\mathbf{x}}f(\mathbf{x}) = D_{\mathbf{x}}\{D_{\mathbf{x}}f(\mathbf{x})\}^T$ .

If  $\mathbf{a} = (a_1, \dots, a_d)$  and  $\mathbf{b} = (b_1, \dots, b_d)$  are two  $d \times 1$  vectors then element-wise multiplication is denoted by  $\mathbf{a} \odot \mathbf{b} = (a_1b_1, \dots, a_db_d)$ . The expression  $\mathbf{a}/\mathbf{b}$  denotes element-wise division  $(a_1/b_1, \dots, a_d/b_d)$ . Scalar functions applied to vectors are also evaluated element-wise. For example,  $\mathbf{a}^{-1/\xi} = (a_1^{-1/\xi}, \dots, a_d^{-1/\xi})$ .

### A.2 Expression for $\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2)$

For the additive model extension, the penalised log-likelihood may be written as

$$\ell_{\text{PL}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2) = h(\mathbf{y}, \boldsymbol{\nu}, \psi, \xi) - \frac{1}{2}\mathbf{u}^T \mathbf{G}_{\sigma^2}^{-1} \mathbf{u}$$

where

$$h(\mathbf{y}, \boldsymbol{\nu}, \psi, \xi) \equiv -n \log(\psi) - \frac{1+\xi}{\xi} \mathbf{1}^T \log\{\mathbf{1} + \xi(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})/\psi\} - \mathbf{1}^T \{\mathbf{1} + \xi(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})/\psi\}^{-\frac{1}{\xi}}.$$

Vector differential calculus methods (e.g. Wand, 2002) lead to

$$D_{\boldsymbol{\nu}}\ell_{\text{PL}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2) = h_{\boldsymbol{\nu}}(\mathbf{y}, \boldsymbol{\nu}, \psi, \xi)^T \mathbf{C} - \begin{bmatrix} \mathbf{0} \\ \mathbf{u}^T \mathbf{G}_{\sigma^2}^{-1} \end{bmatrix}$$

and

$$H_{\boldsymbol{\nu}}\ell_{\text{PL}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2) = \mathbf{C}^T \text{diag}\{h_{\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y}, \boldsymbol{\nu}, \psi, \xi)\} \mathbf{C} - \text{blockdiag}(\mathbf{0}, \mathbf{G}_{\sigma^2}^{-1})$$

where

$$h_{\boldsymbol{\nu}}(\mathbf{y}, \boldsymbol{\nu}, \psi, \xi) \equiv \frac{(1 + \xi)\mathbf{1} - \{\mathbf{1} + \xi(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})/\psi\}^{-1/\xi}}{\psi\{\mathbf{1} + \xi(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})/\psi\}}$$

and

$$h_{\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y}, \boldsymbol{\nu}, \psi, \xi) \equiv \frac{(1 + \xi)[\xi\mathbf{1} - \{\mathbf{1} + \xi(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})/\psi\}^{-1/\xi}]}{\psi^2\{\mathbf{1} + \xi(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})/\psi\}^2}.$$

The required observed information matrix expression is then

$$\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2) = \mathbf{C}^T \text{diag}\{-h_{\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y}, \boldsymbol{\nu}, \psi, \xi)\} \mathbf{C} + \text{blockdiag}(\mathbf{0}, \mathbf{G}_{\sigma^2}^{-1}).$$

The (penalised likelihood-based) expected information matrix is

$$E\{\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2)\} = \{(1 + \xi)/\psi\}^2 \Gamma(1 + 2\xi) \mathbf{C}^T \mathbf{C} + \text{blockdiag}(\mathbf{0}, \mathbf{G}_{\sigma^2}^{-1})$$

which is consistent with results in Prescott and Walden (1980) and Tawn (1988).

### A.3 Expression for $\mathcal{I}_{(\psi,\xi)(\psi,\xi)}(\psi, \xi, \sigma^2)$

First note that

$$\mathcal{I}_{(\psi,\xi)(\psi,\xi)}(\psi, \xi, \sigma^2) = - \begin{bmatrix} H_{\psi\psi} & H_{\psi\xi} \\ H_{\psi\xi} & H_{\xi\xi} \end{bmatrix}$$

where

$$\begin{aligned} H_{\psi\psi} &\equiv \frac{\partial^2}{\partial \psi^2} \left\{ \ell_{\text{PL}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2) - \frac{1}{2} \log |\mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2)| \right\} \\ &= \frac{\partial^2}{\partial \psi^2} \left\{ h(\mathbf{y}, \boldsymbol{\nu}, \psi, \xi) - \frac{1}{2} \log \left| \mathbf{C}^T \text{diag}\{-h_{\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y}, \boldsymbol{\nu}, \psi, \xi)\} \mathbf{C} + \text{blockdiag}(\mathbf{0}, \mathbf{G}_{\sigma^2}^{-1}) \right| \right\} \end{aligned}$$

and  $H_{\psi\xi}$  and  $H_{\xi\xi}$  are defined analogously as the other second-order partial derivatives. Then vector calculus methods lead to

$$\begin{aligned} H_{\psi\psi} &= h_{\psi\psi}(\mathbf{y}, \psi, \xi) + \frac{1}{2} \text{tr} \left[ \mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2)^{-1} \mathbf{C}^T \text{diag}\{h_{\psi\psi\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y}, \psi, \xi)\} \mathbf{C} \right], \\ H_{\psi\xi} &= h_{\psi\xi}(\mathbf{y}, \psi, \xi) + \frac{1}{2} \text{tr} \left[ \mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2)^{-1} \mathbf{C}^T \text{diag}\{h_{\psi\xi\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y}, \psi, \xi)\} \mathbf{C} \right] \quad \text{and} \\ H_{\xi\xi} &= h_{\xi\xi}(\mathbf{y}, \psi, \xi) + \frac{1}{2} \text{tr} \left[ \mathcal{I}_{\boldsymbol{\nu}\boldsymbol{\nu}}(\boldsymbol{\nu}, \psi, \xi, \sigma^2)^{-1} \mathbf{C}^T \text{diag}\{h_{\xi\xi\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y}, \psi, \xi)\} \mathbf{C} \right]. \end{aligned}$$

Here  $\mathbf{r} \equiv (\mathbf{y} - \mathbf{C}\boldsymbol{\nu})/\psi$ ,

$$h_{\psi\psi}(\mathbf{y}, \psi, \xi) \equiv \frac{n}{\psi^2} + \mathbf{1}^T \left[ \frac{\{(1+\xi)\mathbf{r}^2\} \odot \left\{ \xi \mathbf{1} - (1+\xi\mathbf{r})^{-\frac{1}{\xi}} \right\} + 2\mathbf{r} \odot (1+\xi\mathbf{r}) \odot \left\{ (1+\xi\mathbf{r})^{-\frac{1}{\xi}} - (1+\xi)\mathbf{1} \right\}}{\psi^2(1+\xi\mathbf{r})^2} \right],$$

$$\begin{aligned} h_{\psi\xi}(\mathbf{y}, \psi, \xi) &\equiv \mathbf{1}^T \left[ \frac{\mathbf{r} \odot (1+\xi\mathbf{r}) \odot \left[ \mathbf{1} - (1+\xi\mathbf{r})^{-\frac{1}{\xi}} \odot \left\{ \frac{\log(1+\xi\mathbf{r})}{\xi^2} - \frac{\mathbf{r}}{\xi(1+\xi\mathbf{r})} \right\} \right]}{\psi(1+\xi\mathbf{r})^2} \right] \\ &\quad + \mathbf{1}^T \left[ \frac{\mathbf{r}^2 \odot \left\{ (1+\xi\mathbf{r})^{-\frac{1}{\xi}} - (1+\xi)\mathbf{1} \right\}}{\psi(1+\xi\mathbf{r})^2} \right], \end{aligned}$$

$$\begin{aligned} h_{\xi\xi}(\mathbf{y}, \psi, \xi) &\equiv -\mathbf{1}^T \left[ \frac{\log(1+\xi\mathbf{r}) \odot \left\{ (1+\xi\mathbf{r}) \odot \log(1+\xi\mathbf{r}) - 2\xi(\mathbf{r} + \mathbf{1}) \right\} + 2\xi^2\mathbf{r}}{\xi^4(1+\xi\mathbf{r})^{1+\frac{1}{\xi}}} \right] \\ &\quad + \mathbf{1}^T \left[ \frac{\xi\mathbf{r} \odot \left\{ \xi\mathbf{r}(\xi+3) + 2 \right\} - 2(1+\xi\mathbf{r}) \odot \log(1+\xi\mathbf{r})}{\xi^3(1+\xi\mathbf{r})^2} \right], \end{aligned}$$

$$\begin{aligned} h_{\psi\psi\boldsymbol{\nu}\boldsymbol{\nu}}(\mathbf{y}, \psi, \xi) &\equiv \frac{6(1+\xi)(\xi\mathbf{r})^2 \odot \left\{ \xi \mathbf{1} - (1+\xi\mathbf{r})^{-\frac{1}{\xi}} \right\} - (1+6\xi+5\xi^2)(\mathbf{r}^2) \odot (1+\xi\mathbf{r})^{-\frac{1}{\xi}}}{\psi^4(1+\xi\mathbf{r})^4} \\ &\quad + \frac{6(1+\xi) \left\{ \xi \mathbf{1} - (1+\xi\mathbf{r})^{-\frac{1}{\xi}} \right\}}{\psi^4(1+\xi\mathbf{r})^2} \\ &\quad + \frac{(1+\xi)\mathbf{r} \odot \left[ 6(1+\xi\mathbf{r})^{-\frac{1}{\xi}} - 12\xi \left\{ \xi \mathbf{1} - (1+\xi\mathbf{r})^{-\frac{1}{\xi}} \right\} \right]}{\psi^4(1+\xi\mathbf{r})^3}, \end{aligned}$$

$$\begin{aligned}
h_{\psi\xi\nu\nu}(\mathbf{y}, \psi, \xi) &\equiv \frac{3\xi(2 + \xi)\mathbf{r} - (7 + 8\xi)\mathbf{r} \odot (\mathbf{1} + \xi\mathbf{r})^{-\frac{1}{\xi}}}{\psi^3(\mathbf{1} + \xi\mathbf{r})^3} \\
&+ \frac{2(1 + \xi) \left[ \mathbf{1} - (\mathbf{1} + \xi\mathbf{r})^{-1/\xi} \odot \left\{ \frac{\log(\mathbf{1} + \xi\mathbf{r})}{\xi^2} - \frac{\mathbf{r}}{\xi(\mathbf{1} + \xi\mathbf{r})} \right\} \right] \odot (\xi\mathbf{r} - \mathbf{1})}{\psi^3(\mathbf{1} + \xi\mathbf{r})^3} \\
&- \frac{(\mathbf{1} + \xi)\mathbf{r} \odot (\mathbf{1} + \xi\mathbf{r})^{-1/\xi} \odot \left\{ \frac{\log(\mathbf{1} + \xi\mathbf{r})}{\xi^2} - \frac{\mathbf{r}}{\xi(\mathbf{1} + \xi\mathbf{r})} \right\}}{\psi^3(\mathbf{1} + \xi\mathbf{r})^3} \\
&+ \frac{3(1 + \xi)\mathbf{r} \odot (2\xi\mathbf{1} + \mathbf{r}\psi) \odot (\mathbf{1} + \xi\mathbf{r})^{-1/\xi} - 6\xi^2(1 + \xi)\mathbf{r}}{\psi^4(\mathbf{1} + \xi\mathbf{r})^4} \\
&- \frac{2 \left\{ \xi\mathbf{1} - (\mathbf{1} + \xi\mathbf{r})^{-\frac{1}{\xi}} \right\}}{\psi^3(\mathbf{1} + \xi\mathbf{r})^2}
\end{aligned}$$

and

$$\begin{aligned}
h_{\xi\xi\nu\nu}(\mathbf{y}, \psi, \xi) &\equiv \frac{-(1 + \xi)(\mathbf{1} + \xi\mathbf{r})^{-1/\xi} \odot \left\{ \frac{\log(\mathbf{1} + \xi\mathbf{r})}{\xi^2} - \frac{\mathbf{r}}{\xi(\mathbf{1} + \xi\mathbf{r})} \right\}^2}{\psi^2(\mathbf{1} + \xi\mathbf{r})^2} \\
&- \frac{(\mathbf{1} + \xi)(\mathbf{1} + \xi\mathbf{r}) \odot \left\{ \frac{2\mathbf{r} + \mathbf{r}^2\xi}{\xi^2(\mathbf{1} + \xi\mathbf{r})} - \frac{2\log(\mathbf{1} + \xi\mathbf{r})}{\xi^3} \right\}}{\psi^2(\mathbf{1} + \xi\mathbf{r})^2} \\
&+ \frac{\left[ \mathbf{1} - (\mathbf{1} + \xi\mathbf{r})^{-1/\xi} \odot \left\{ \frac{\log(\mathbf{1} + \xi\mathbf{r})}{\xi^2} - \frac{\mathbf{r}}{\xi(\mathbf{1} + \xi\mathbf{r})} \right\} \right] \odot (2\mathbf{1} - 4\mathbf{r} - 2\xi\mathbf{r})}{\psi^2(\mathbf{1} + \xi\mathbf{r})^3} \\
&+ \frac{2\mathbf{r} \odot \left\{ \xi\mathbf{1} - (\mathbf{1} + \xi\mathbf{r})^{-1/\xi} \right\} \odot (3\xi\mathbf{1} - 2\xi\mathbf{r} + \mathbf{1})}{\psi^2(\mathbf{1} + \xi\mathbf{r})^4}.
\end{aligned}$$

#### 4.1 Variability bands

A naïve expression for variability bands at  $x_i$  is given by:

$$(\mathbf{C}\hat{\boldsymbol{\nu}})_i \pm 2\sqrt{\{\mathbf{C}\widehat{\text{Cov}}(\hat{\boldsymbol{\nu}}|\mathbf{u})\mathbf{C}^T\}_{ii}}.$$

where  $\widehat{\text{Cov}}(\hat{\boldsymbol{\nu}}|\mathbf{u})$  is the estimated covariance matrix of  $\hat{\boldsymbol{\nu}}$  given  $\mathbf{u}$ . It is based on  $\text{Cov}(\hat{\boldsymbol{\nu}}|\mathbf{u}) = \mathcal{I}_{\nu\nu}(\boldsymbol{\nu}, \psi, \xi, \boldsymbol{\sigma}^2)^{-1}$  with  $\boldsymbol{\nu}, \psi, \xi$  and  $\boldsymbol{\sigma}^2$  replaced by their estimates.

## Acknowledgment

This work was commenced while both authors were visiting School of Mathematics and Statistics, University of New South Wales, Sydney, Australia. Their hospitality is gratefully acknowledged. Thanks are also due to Scott Sisson, Stuart Coles and Francesco Pauli for their helpful suggestions. This research was partially supported by Australian Research Council Discovery Project DP0877055.

## References

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. London: Chapman & Hall.

- Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.
- Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Applied Statistics*, **54**, 207–222.
- Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.
- Crainiceanu, C. and Ruppert, D. (2004) Likelihood ratio tests in linear mixed models with one variance component, *Journal of the Royal Statistical Society, Series B*, **66**, 165–185.
- Broyden, C. G. (1967). Quasi-Newton methods and their application to function minimization. *Mathematics of Computation*, **21**, 368–81.
- Davison, A. C. and Ramesh, N. I. (2000). Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society, Series B*, **62**, 191–208.
- Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.
- Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, **55**, 245–259.
- Hall, P. and Tajvidi, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, **15**, 153–167.
- Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological events . *Quarterly Journal of the Royal Meteorological Society*, **81**, 158–272.
- Miller, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics*, **5**, 746–762.
- Pauli, F. and Coles, S. G. (2001). Penalized likelihood inference in extreme value analysis. *Journal of Applied Statistics*, **28**, 547–560.
- Prescott, P. and Walden, A. T. (1980). Maximum likelihood of the parameters of the generalized extreme-value distribution. *Biometrika*, **67**, 723–724.
- Ruppert, D., Wand, M. P. and Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.
- Self, S.G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.
- Severini, T.A. (1998). An approximation to the modified profile likelihood function. *Biometrika*, **85**, 403–411.
- Smith, R.L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72**, 67–90.
- Tawn, J.A. (1988). An extreme value theory model for dependent observations. *Journal of*

*Hydrology*, **101**, 227–250.

von Mises, R. (1954). La distribution de la plus grande de  $n$  valeurs. In *Selected Papers, Volume II*, pages 271–294. American Mathematical Society, Providence, Rhode Island, USA.

Wand, M. P. (2002). Vector differential calculus in statistics. *The American Statistician*, **56**, 55–62.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*. **18**, 223–249.

Wand, M. P. and Ormerod, J. T. (2008). On semiparametric regression with O' Sullivan penalised splines. *Australian and New Zealand Journal of Statistics*, in press.

Welham, S. J., Cullis, B. R., Kenward, M. G. and Thompson, R. (2006). A comparison of mixed model splines. *Australian and New Zealand Journal of Statistics*, **49**, 1–23.

Yee, T.W. and Stephenson, A.G. (2007). Vector generalized linear and additive extreme value models. *Extremes*, **10**, 1–19.