Appl. Statist. (2005)
54, *Part* 3, *pp.* 507–554

Generalized additive models for location, scale and shape

R. A. Rigby and D. M. Stasinopoulos

London Metropolitan University, UK

[*Read before* The Royal Statistical Society *on Tuesday, November 23rd, 2004, the President*, Professor A. P. Grieve, *in the Chair*]

Summary. A general class of statistical models for a univariate response variable is presented which we call the generalized additive model for location, scale and shape (GAMLSS). The model assumes independent observations of the response variable y given the parameters, the explanatory variables and the values of the random effects. The distribution for the response variable in the GAMLSS can be selected from a very general family of distributions including highly skew or kurtotic continuous and discrete distributions. The systematic part of the model is expanded to allow modelling not only of the mean (or location) but also of the other parameters of the distribution of y, as parametric and/or additive nonparametric (smooth) functions of explanatory variables and/or random-effects terms. Maximum (penalized) likelihood estimation is used to fit the (non)parametric models. A Newton–Raphson or Fisher scoring algorithm is used to maximize the (penalized) likelihood. The additive terms in the model are fitted by using a backfitting algorithm. Censored data are easily incorporated into the framework. Five data sets from different fields of application are analysed to emphasize the generality of the GAMLSS class of models.

Keywords: Beta–binomial distribution; Box–Cox transformation; Centile estimation; Cubic smoothing splines; Generalized linear mixed model; *LMS* method; Negative binomial distribution; Non-normality; Nonparametric models; Overdispersion; Penalized likelihood; Random effects; Skewness and kurtosis

1. Introduction

The quantity of data collected and requiring statistical analysis has been increasing rapidly over recent years, allowing the fitting of more complex and potentially more realistic models. In this paper we develop a very general regression-type model in which both the systematic and the random parts of the model are highly flexible and where the fitting algorithm is sufficiently fast to allow the rapid exploration of very large and complex data sets.

Within the framework of univariate regression modelling techniques the generalized linear model (GLM) and generalized additive model (GAM) hold a prominent place (Nelder and Wedderburn (1972) and Hastie and Tibshirani (1990) respectively). Both models assume an exponential family distribution for the response variable y in which the mean μ of y is modelled as a function of explanatory variables and the variance of y, given by $V(y) = \phi\, v(\mu)$, depends on a constant dispersion parameter ϕ and on the mean μ, through the variance function $v(\mu)$. Furthermore, for an exponential family distribution both the skewness and the kurtosis of y are, in general, functions of μ and ϕ. Hence, in the GLM and GAM models, the variance, skewness

Address for correspondence: R. A. Rigby, Statistics, OR and Mathematics (STORM) Research Centre, London Metropolitan University, 166–220 Holloway Road, London, N7 8DB, UK.
E-mail: r.rigby@londonmet.ac.uk

 0035–9254/05/54507

and kurtosis are not modelled explicitly in terms of the explanatory variables but implicitly through their dependence on μ.

Another important class of models, the linear mixed (random-effects) models, which provide a very broad framework for modelling dependent data particularly associated with spatial, hierarchical and longitudinal sampling schemes, assume normality for the conditional distribution of y given the random effects and therefore cannot model skewness and kurtosis explicitly.

The generalized linear mixed model (GLMM) combines the GLM and linear mixed model, by introducing a (usually normal) random-effects term in the linear predictor for the mean of a GLM. Bayesian procedures to fit GLMMs by using the EM algorithm and Markov chain Monte Carlo methods were described by McCulloch (1997) and Zeger and Karim (1991). Lin and Zhang (1999) gave an example of a generalized additive mixed model (GAMM). Fahrmeir and Lang (2001) discussed GAMM modelling using Bayesian inference. Fahrmeir and Tutz (2001) discussed alternative estimation procedures for the GLMM and GAMM. The GLMM and GAMM, although more flexible than the GLM and GAM, also assume an exponential family conditional distribution for y and rarely allow the modelling of parameters other than the mean (or location) of the distribution of the response variable as functions of the explanatory variables. Their fitting often depends on Markov chain Monte Carlo or integrated (marginal distribution) likelihoods (e.g. Gaussian quadrature), making them highly computationally intensive and time consuming, at least at present, for large data sets where the model selection requires the investigation of many alternative models. Various approximate procedures for fitting a GLMM have been proposed (Breslow and Clayton, 1993; Breslow and Lin, 1995; Lee and Nelder, 1996, 2001a, b). An alternative approach is to use nonparametric maximum likelihood based on finite mixtures; Aitkin (1999).

In this paper we develop a general class of univariate regression models which we call the generalized additive model for location, scale and shape (GAMLSS), where the exponential family assumption is relaxed and replaced by a very general distribution family. Within this new framework, the systematic part of the model is expanded to allow not only the mean (or location) but all the parameters of the conditional distribution of y to be modelled as parametric and/or additive nonparametric (smooth) functions of explanatory variables and/or random-effects terms. The model fitting of a GAMLSS is achieved by either of two different algorithmic procedures. The first algorithm (RS) is based on the algorithm that was used for the fitting of the mean and dispersion additive models of Rigby and Stasinopoulos (1996a), whereas the second (CG) is based on the Cole and Green (1992) algorithm.

Section 2 formally introduces the GAMLSS. Parametric terms in the linear predictors are considered in Section 3.1, and several specific forms of additive terms which can be incorporated in the predictors are considered in Section 3.2. These include nonparametric smooth function terms, using cubic splines or smoothness priors, random-walk terms and many random-effects terms (including terms for simple overdispersion, longitudinal random effects, random-coefficient models, multilevel hierarchical models and crossed and spatial random effects). A major advantage of the GAMLSS framework is that any combinations of the above terms can be incorporated easily in the model. This is discussed in Section 3.3.

Section 4 describes specific families of distributions for the dependent variable which have been implemented in the GAMLSS. Incorporating censored data and centile estimation are also discussed there. The RS and CG algorithms (based on the Newton–Raphson or Fisher scoring algorithm) for maximizing the (penalized) likelihood of the data under a GAMLSS are discussed in Section 5. The details and justification of the algorithms are given in Appendices B and C respectively. The inferential framework for the GAMLSS is considered in Appendix A, where alternative inferential approaches are considered. Model selection, inference and residual

diagnostics are considered in Section 6. Section 7 gives five practical examples. Section 8 concludes the paper.

2. The generalized additive model for location, scale and shape

2.1. Definition

The p parameters $\boldsymbol{\theta}^{\mathrm{T}} = (\theta_1, \theta_2, \ldots, \theta_p)$ of a population probability (density) function $f(y|\boldsymbol{\theta})$ are modelled here by using additive models. Specifically the model assumes that, for $i = 1, 2, \ldots, n$, observations y_i are independent conditional on $\boldsymbol{\theta}^i$, with probability (density) function $f(y_i|\boldsymbol{\theta}^i)$, where $\boldsymbol{\theta}^{i\mathrm{T}} = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})$ is a vector of p parameters related to explanatory variables and random effects. (If covariate values are stochastic or observations y_i depend on their past values then $f(y_i|\boldsymbol{\theta}^i)$ is understood to be conditional on these values.)

Let $\mathbf{y}^{\mathrm{T}} = (y_1, y_2, \ldots, y_n)$ be the vector of the response variable observations. Also, for $k = 1, 2, \ldots, p$, let $g_k(\cdot)$ be a known monotonic link function relating $\boldsymbol{\theta}_k$ to explanatory variables and random effects through an additive model given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk} \boldsymbol{\gamma}_{jk} \tag{1}$$

where $\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_k$ are vectors of length n, e.g. $\boldsymbol{\theta}_k^{\mathrm{T}} = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{nk})$, $\boldsymbol{\beta}_k^{\mathrm{T}} = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J'_k k})$ is a parameter vector of length J'_k, $\mathbf{X}_k$ is a known design matrix of order $n \times J'_k$, $\mathbf{Z}_{jk}$ is a fixed known $n \times q_{jk}$ design matrix and $\boldsymbol{\gamma}_{jk}$ is a q_{jk}-dimensional random variable. We call model (1) the GAMLSS.

The vectors $\boldsymbol{\gamma}_{jk}$ for $j = 1, 2, \ldots, J_k$ could be combined into a single vector $\boldsymbol{\gamma}_k$ with a single design matrix $\mathbf{Z}_k$; however, formulation (1) is preferred here as it is suited to the backfitting algorithm (see Appendix B) and allows combinations of different types of additive random-effects terms to be incorporated easily in the model (see Section 3.3).

If, for $k = 1, 2, \ldots, p$, $J_k = 0$ then model (1) reduces to a fully parametric model given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k. \tag{2}$$

If $\mathbf{Z}_{jk} = \mathbf{I}_n$, where $\mathbf{I}_n$ is an $n \times n$ identity matrix, and $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ for all combinations of j and k in model (1), this gives

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{3}$$

where $\mathbf{x}_{jk}$ for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$ are vectors of length n. The function h_{jk} is an unknown function of the explanatory variable X_{jk} and $\mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ is the vector which evaluates the function h_{jk} at $\mathbf{x}_{jk}$. The explanatory vectors $\mathbf{x}_{jk}$ are assumed to be known. We call the model in equation (3) the semiparametric GAMLSS. Model (3) is an important special case of model (1). If $\mathbf{Z}_{jk} = \mathbf{I}_n$ and $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ for specific combinations of j and k in model (1), then the resulting model contains parametric, nonparametric and random-effects terms.

The first two population parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in model (1) are usually characterized as location and scale parameters, denoted here by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, whereas the remaining parameter(s), if any, are characterized as shape parameters, although the model may be applied more generally to the parameters of any population distribution.

For many families of population distributions a maximum of two shape parameters $\boldsymbol{\nu}$ $(= \boldsymbol{\theta}_3)$ and $\boldsymbol{\tau}$ $(= \boldsymbol{\theta}_4)$ suffice, giving the model

$$\left.\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1}, \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2}, \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3}, \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4}. \end{aligned}\right\} \tag{4}$$

The GAMLSS model (1) is more general than the GLM, GAM, GLMM or GAMM in that the distribution of the dependent variable is not limited to the exponential family and all parameters (not just the mean) are modelled in terms of both fixed and random effects.

2.2. Model estimation

Crucial to the way that additive components are fitted within the GAMLSS framework is the backfitting algorithm and the fact that quadratic penalties in the likelihood result from assuming a normally distributed random effect in the linear predictor. The resulting estimation uses shrinking (smoothing) matrices within a backfitting algorithm, as shown below.

Assume in model (1) that the $\boldsymbol{\gamma}_{jk}$ have independent (prior) normal distributions with $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^{-})$, where $\mathbf{G}_{jk}^{-}$ is the (generalized) inverse of a $q_{jk} \times q_{jk}$ symmetric matrix $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$, which may depend on a vector of hyperparameters $\boldsymbol{\lambda}_{jk}$, and where if $\mathbf{G}_{jk}$ is singular then $\boldsymbol{\gamma}_{jk}$ is understood to have an improper prior density function proportional to $\exp(-\frac{1}{2}\boldsymbol{\gamma}_{jk}^{\mathrm{T}}\mathbf{G}_{jk}\boldsymbol{\gamma}_{jk})$. Subsequently in the paper we refer to $\mathbf{G}_{jk}$ rather than to $\mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$ for simplicity of notation, although the dependence of $\mathbf{G}_{jk}$ on hyperparameters $\boldsymbol{\lambda}_{jk}$ remains throughout.

The assumption of independence between different random-effects vectors $\boldsymbol{\gamma}_{jk}$ is essential within the GAMLSS framework. However, if, for a particular k, two or more random-effect vectors are not independent, they can be combined into a single random-effect vector and their corresponding design matrices $\mathbf{Z}_{jk}$ into a single design matrix, to satisfy the condition of independence.

In Appendix A.1 it is shown, by using empirical Bayesian arguments, that posterior mode estimation (or maximum *a posteriori* (MAP) estimation; see Berger (1985)) for the parameter vectors $\boldsymbol{\beta}_k$ and the random-effect terms $\boldsymbol{\gamma}_{jk}$ (for fixed values of the smoothing or hyperparameters $\boldsymbol{\lambda}_{jk}$), for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$, is equivalent to penalized likelihood estimation. Hence for fixed $\boldsymbol{\lambda}_{jk}$s the $\boldsymbol{\beta}_k$s and the $\boldsymbol{\gamma}_{jk}$s are estimated within the GAMLSS framework by maximizing a penalized likelihood function l_p given by

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \boldsymbol{\gamma}_{jk}^{\mathrm{T}}\mathbf{G}_{jk}\boldsymbol{\gamma}_{jk} \tag{5}$$

where $l = \sum_{i=1}^{n} \log\{f(y_i|\boldsymbol{\theta}^i)\}$ is the log-likelihood function of the data given $\boldsymbol{\theta}^i$ for $i = 1, 2, \ldots, n$. This is equivalent to maximizing the extended or hierarchical likelihood defined by

$$l_h = l_p + \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \{\log|\mathbf{G}_{jk}| - q_{jk}\log(2\pi)\}$$

(see Pawitan (2001), page 429, and Lee and Nelder (1996)).

It is shown in Appendix C that maximizing l_p is achieved by the CG algorithm, which is described in Appendix B. Appendix C shows that the maximization of l_p leads to the shrinking (smoothing) matrix $\mathbf{S}_{jk}$, applied to partial residuals ε_{jk} to update the estimate of the additive predictor $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}$ within a backfitting algorithm, given by

$$\mathbf{S}_{jk} = \mathbf{Z}_{jk}(\mathbf{Z}_{jk}^{\mathrm{T}}\mathbf{W}_{kk}\mathbf{Z}_{jk} + \mathbf{G}_{jk})^{-1}\mathbf{Z}_{jk}^{\mathrm{T}}\mathbf{W}_{kk} \tag{6}$$

for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$, where $\mathbf{W}_{kk}$ is a diagonal matrix of iterative weights. Different forms of $\mathbf{Z}_{jk}$ and $\mathbf{G}_{jk}$ correspond to different types of additive terms in the linear predictor $\boldsymbol{\eta}_k$ for $k = 1, 2, \ldots, p$. For random-effects terms $\mathbf{G}_{jk}$ is often a simple and/or low order matrix whereas for a cubic smoothing spline term $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk}$, $\mathbf{Z}_{jk} = \mathbf{I}_n$ and $\mathbf{G}_{jk} = \lambda_{jk}\mathbf{K}_{jk}$ where $\mathbf{K}_{jk}$ is a structured matrix. Either case allows easy updating of $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}$.

The hyperparameters $\boldsymbol{\lambda}$ can be fixed or estimated. In Appendix A.2 we propose four alternative methods of estimation of $\boldsymbol{\lambda}$ which avoid integrating out the random effects.

2.3. Comparison of generalized additive models for location, scale and shape and hierarchical generalized linear models

Lee and Nelder (1996, 2001a) developed hierarchical generalized linear models. In the notation of the GAMLSS, they use, in general, extended quasi-likelihood to approximate the conditional distribution of y given $\boldsymbol{\theta} = (\mu, \phi)$, where μ and ϕ are mean and scale parameters respectively, and any conjugate distribution for the random effects $\boldsymbol{\gamma}$ (parameterized by λ). They model predictors for μ, ϕ and λ in terms of explanatory variables, and the predictor for μ also includes random-effects terms. Lee and Nelder (1996, 2001a) assumed independent random effects, whereas Lee and Nelder (2001b) relaxed this assumption to allow correlated random effects.

However, extended quasi-likelihood does not provide a proper distribution which integrates or sums to 1 (and the integral or sum cannot be obtained explicitly, varies between cases and depends on the parameters of the model). In large samples this has been found to lead to serious inaccuracies in the fitted global deviance, even for the gamma distribution (see Stasinopoulos *et al.* (2000)), resulting potentially in a misleading comparison with a proper distribution. It is also quite restrictive in the shape of distributions that are available for y given $\boldsymbol{\theta}$, particularly for continuous distributions where it is unsuitable for negatively skew data, or for platykurtic data or for leptokurtic data unless positively skewed. In addition, hierarchical generalized linear models allow neither explanatory variables nor random effects in the predictors for the shape parameters of $f(y|\boldsymbol{\theta})$.

3. The linear predictor

3.1. Parametric terms

In the GAMLSS (1) the linear predictors $\boldsymbol{\eta}_k$, for $k = 1, 2, \ldots, p$, comprise a parametric component $\mathbf{X}_k\boldsymbol{\beta}_k$ and additive components $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}$, for $j = 1, \ldots, J_k$. The parametric component can include linear and interaction terms for explanatory variables and factors, polynomials, fractional polynomials (Royston and Altman, 1994) and piecewise polynomials (with fixed knots) for variables (Smith, 1979; Stasinopoulos and Rigby, 1992).

Non-linear parameters can be incorporated into the GAMLSS (1) and fitted by either of two methods:

(a) the profile or
(b) the derivative method.

In the profile fitting method, estimation of non-linear parameters is achieved by maximizing their profile likelihood. An example of the profile method is given in Section 7.1 where the age explanatory variable is transformed to $x = \text{age}^{\xi}$ where ξ is a non-linear parameter. In the derivative fitting method, the derivatives of a predictor $\boldsymbol{\eta}_k$ with respect to non-linear parameters are included in the design matrix $\mathbf{X}_k$ in the fitting algorithm; see, for example, Benjamin *et al.* (2003). Lindsey (`http://alpha.luc.ac.be/jlindsey/`) has also considered modelling parameters of a distribution as non-linear functions of explanatory variables.

3.2. Additive terms

The additive components $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}$ in model (1) can model a variety of terms such as smoothing and random-effect terms as well as terms that are useful for time series analysis (e.g. random walks). Different additive terms that can be included in the GAMLSS will be discussed below. For simplicity of exposition we shall drop the subscripts j and k in the vectors and matrices, where appropriate.

3.2.1. Cubic smoothing splines terms

With cubic smoothing splines terms we assume in model (3) that the functions $h(t)$ are arbitrary twice continuously differentiable functions and we maximize a penalized log-likelihood, given by l subject to penalty terms of the form $\lambda \int_{-\infty}^{\infty} h''(t)^2 \, \mathrm{d}t$. Following Reinsch (1967), the maximizing functions $h(t)$ are all natural cubic splines and hence can be expressed as linear combinations of their natural cubic spline basis functions $B_i(t)$ for $i = 1, 2, \ldots, n$ (de Boor, 1978; Schumaker, 1993), i.e. $h(t) = \Sigma_{i=1}^{n} \delta_i \, B_i(t)$. Let $\mathbf{h} = h(\mathbf{x})$ be the vector of evaluations of the function $h(t)$ at the values $\mathbf{x}$ of the explanatory variable X (which is assumed to be distinct for simplicity of exposition). Let $\mathbf{N}$ be an $n \times n$ non-singular matrix containing as its columns the n-vectors of evaluations of functions $B_i(t)$, for $i = 1, 2, \ldots, n$, at $\mathbf{x}$. Then $\mathbf{h}$ can be expressed by using coefficient vector $\boldsymbol{\delta}$ as a linear combination of the columns of $\mathbf{N}$ by $\mathbf{h} = \mathbf{N}\boldsymbol{\delta}$. Let $\boldsymbol{\Omega}$ be the $n \times n$ matrix of inner products of the second derivatives of the natural cubic spline basis functions, with (r, s)th entry given by

$$\Omega_{rs} = \int B_r''(t) \, B_s''(t) \, \mathrm{d}t.$$

The penalty is then given by the quadratic form

$$Q(\mathbf{h}) = \lambda \int_{-\infty}^{\infty} h''(t)^2 \, \mathrm{d}t = \lambda \boldsymbol{\delta}^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{\delta} = \lambda \mathbf{h}^{\mathrm{T}} \mathbf{N}^{-\mathrm{T}} \boldsymbol{\Omega} \mathbf{N}^{-1} \mathbf{h} = \lambda \mathbf{h}^{\mathrm{T}} \mathbf{K} \mathbf{h},$$

where $\mathbf{K} = \mathbf{N}^{-\mathrm{T}} \boldsymbol{\Omega} \mathbf{N}^{-1}$ is a known penalty matrix that depends only on the values of the explanatory vector $\mathbf{x}$ (Hastie and Tibshirani (1990), chapter 2). The precise form of the matrix $\mathbf{K}$ can be found in Green and Silverman (1994), section 2.1.2.

The model can be formulated as a random-effects GAMLSS (1) by letting $\boldsymbol{\gamma} = \mathbf{h}$, $\mathbf{Z} = \mathbf{I}_n$, $\mathbf{K} = \mathbf{N}^{-\mathrm{T}} \boldsymbol{\Omega} \mathbf{N}^{-1}$ and $\mathbf{G} = \lambda \mathbf{K}$, so that $\mathbf{h} \sim N_n(0, \lambda^{-1} \mathbf{K}^{-})$, a partially improper prior (Silverman, 1985). This amounts to assuming complete prior uncertainty about the constant and linear functions and decreasing uncertainty about higher order functions; see Verbyla *et al.* (1999).

3.2.2. Parameter-driven time series terms and smoothness priors

First assume that an explanatory variable X has equally spaced observations x_i, $i = 1, \ldots, n$, sorted into the ordered sequence $x_{(1)} < \ldots < x_{(i)} < \ldots < x_{(n)}$ defining an equidistant grid on the

x-axis. Typically, for a parameter-driven time series term, X corresponds to time units as days, weeks, months or years. First- and second-order random walks, denoted as rw(1) and rw(2), are defined respectively by $h[x_{(i)}] = h[x_{(i-1)}] + \varepsilon_i$ and $h[x_{(i)}] = 2\,h[x_{(i-1)}] - h[x_{(i-2)}] + \varepsilon_i$ with independent errors $\varepsilon_i \sim N(0, \lambda^{-1})$ for $i > 1$ and $i > 2$ respectively, and with diffuse uniform priors for $h[x_{(1)}]$ for rw(1) and, in addition, for $h[x_{(2)}]$ for rw(2). Let $\mathbf{h} = h(\mathbf{x})$; then $\mathbf{D}_1\mathbf{h} \sim N_{n-1}(0, \lambda^{-1}\mathbf{I})$ and $\mathbf{D}_2\mathbf{h} \sim N_{n-2}(0, \lambda^{-1}\mathbf{I})$, where $\mathbf{D}_1$ and $\mathbf{D}_2$ are $(n-1) \times n$ and $(n-2) \times n$ matrices giving first and second differences respectively. The above terms can be included in the GAMLSS framework (1) by letting $\mathbf{Z} = \mathbf{I}_n$ and $\mathbf{G} = \lambda\mathbf{K}$ so that $\boldsymbol{\gamma} = \mathbf{h} \sim N(0, \lambda^{-1}\mathbf{K}^-)$, where $\mathbf{K}$ has a structured form given by $\mathbf{K} = \mathbf{D}_1^{\mathrm{T}}\mathbf{D}_1$ or $\mathbf{K} = \mathbf{D}_2^{\mathrm{T}}\mathbf{D}_2$ for rw(1) or rw(2) respectively; see Fahrmeir and Tutz (2001), pages 223–225 and 363–364. (The resulting quadratic penalty $\lambda\mathbf{h}^{\mathrm{T}}\mathbf{K}\mathbf{h}$ for rw(2) is a discretized version of the corresponding cubic spline penalty term $\lambda \int_{-\infty}^{\infty} h''(t)^2\,\mathrm{d}t$.) Hence many of the state space models of Harvey (1989) can be incorporated in the GAMLSS framework.

The more general case of a non-equally spaced variable X requires modifications to $\mathbf{K}$ (Fahrmeir and Lang, 2001), where X is any continuous variable and the prior distribution for $\mathbf{h}$ is called a smoothness prior.

3.2.3. Penalized splines terms

Smoothers in which the number of basis functions is less than the number of observations but in which their regression coefficients are penalized are referred to as penalized splines or P-splines; see Eilers and Marx (1996) and Wood (2001). Eilers and Marx (1996) used a set of q B-spline basis functions in the explanatory variable X (whose evaluations at the values $\mathbf{x}$ of X are the columns of the $n \times q$ design matrix $\mathbf{Z}$ in equation (1)). They suggested the use of a moderately large number of equal-spaced knots (i.e. between 20 and 40), at which the spline segments connect, to ensure enough flexibility in the fitted curves, but they imposed penalties on the B-spline basis function parameters $\boldsymbol{\gamma}$ to guarantee sufficient smoothness of the resulting fitted curves. In effect they assumed that $\mathbf{D}_r\boldsymbol{\gamma} \sim N_{n-r}(0, \lambda^{-1}\mathbf{I})$ where $\mathbf{D}_r$ is a $(q-r) \times q$ matrix giving rth differences of the q-dimensional vector $\boldsymbol{\gamma}$. (The same approach was used by Wood (2001) but he used instead a cubic Hermite polynomial basis rather than a B-spine. He also provided a way of estimating the hyperparameters by using generalized cross-validation (Wood, 2000).) Hence, in the GAMLSS framework (1), this corresponds to letting $\mathbf{G} = \lambda\mathbf{K}$ so that $\boldsymbol{\gamma} \sim N(\mathbf{0}, \lambda^{-1}\mathbf{K}^-)$ where $\mathbf{K} = \mathbf{D}_r^{\mathrm{T}}\mathbf{D}_r$.

3.2.4. Other smoothers

Other smoothers can be used as additive terms, e.g. the R implementation of a GAMLSS allows local regression smoothers, `loess`; Cleveland *et al.* (1993).

3.2.5. Varying-coefficient terms

Varying-coefficient models (Hastie and Tibshirani, 1993) allow a particular type of interaction between smoothing additive terms and continuous variables or factors. They are of the form $\mathbf{r}\,h(\mathbf{x})$ where $\mathbf{r}$ and $\mathbf{x}$ are vectors of fixed values of the explanatory variables R and X. It can be shown that they can be incorporated easily in the GAMLSS fitting algorithm by using a smoothing matrix in the form of equation (6) in the backfitting algorithm, with $\mathbf{Z} = \mathbf{I}_n$, $\mathbf{K} = \mathbf{N}^{-\mathrm{T}}\boldsymbol{\Omega}\mathbf{N}^{-1}$ and $\mathbf{G} = \lambda\mathbf{K}$ as in Section 3.2.1 above, but, assuming that the values of R are distinct, with the diagonal matrix of iterative weights $\mathbf{W}$ multiplied by $\mathrm{diag}(r_1^2, r_2^2, \ldots, r_n^2)$ and the partial residuals ε_i divided by r_i for $i = 1, 2, \ldots, n$.

3.2.6. Spatial (covariate) random-effect terms

Besag *et al.* (1991) and Besag and Higdon (1999) considered models for spatial random effects with singular multivariate normal distributions, whereas Breslow and Clayton (1993), Lee and Nelder (2001b) and Fahrmeir and Lang (2001) considered incorporating these spatial terms in the predictor of the mean in GLMMs. In model (1) the spatial terms can be included in the predictor of one or more of the location, scale and shape parameters. For example consider an intrinsic autoregressive model (Besag *et al.*, 1991), in which the vector of random effects for q geographical regions $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\ldots,\gamma_q)^{\mathrm{T}}$ has an improper prior density that is proportional to $\exp(-\frac{1}{2}\lambda\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{K}\boldsymbol{\gamma})$, denoted $\boldsymbol{\gamma}\sim N_q(0,\lambda^{-1}\mathbf{K}^{-})$, where the elements of the $q\times q$ matrix $\mathbf{K}$ are given by $k_{mm}=n_m$ where n_m is the total number of regions adjacent to region m and $k_{mt}=-1$ if regions m and t are adjacent, and $k_{mt}=0$ otherwise, for $m=1,2,\ldots,q$ and $t=1,2,\ldots,q$. This model has the attractive property that, conditional on λ and $\boldsymbol{\gamma}_t$ for $t\neq m$, then $\gamma_m\sim N\{\sum\gamma_t n_m^{-1},(\lambda n_m)^{-1}\}$ where the summation is over all regions which are neighbours of region m. This is incorporated in a GAMLSS by setting $\mathbf{Z}=\mathbf{I}_q$ and $\mathbf{G}=\lambda\mathbf{K}$.

3.2.7. Specific random-effects terms

Lee and Nelder (2001b) considered various random-effect terms in the predictor of the mean in GLMMs. Many specific random-effects terms can be incorporated in the predictors in model (1) including the following.

(a) An overdispersion term: in model (1) let $\mathbf{Z}=\mathbf{I}_n$ and $\boldsymbol{\gamma}\sim N_n(\mathbf{0},\lambda^{-1}\mathbf{I}_n)$; then this provides an overdispersion term for each observation (i.e. case) in the predictor.
(b) A one-factor random-effect term: in model (1) let $\mathbf{Z}$ be an $n\times q$ incidence design matrix (for a q-level factor) defined by elements $z_{it}=1$ if the ith observation belongs to the tth factor level, and otherwise $z_{it}=0$, and let $\boldsymbol{\gamma}\sim N_q(\mathbf{0},\lambda^{-1}\mathbf{I}_q)$; then this provides a one-factor random-effects model.
(c) A correlated random-effects term: in model (1), since $\boldsymbol{\gamma}\sim N(\mathbf{0},\mathbf{G}^{-})$, correlated structures can be applied to the random effects by a suitable choice of the matrix $\mathbf{G}$, e.g. first- or second-order random walks, first- or second-order autoregressive, (time-dependent) exponential decaying and compound symmetry correlation models.

3.3. Combinations of terms

Any combinations of parametric and additive terms can be combined (in the predictors of one or more of the location, scale or shape parameters) to produce more complex terms or models.

3.3.1. Combinations of random-effect terms

3.3.1.1. Two-level longitudinal repeated measurement design. Consider a two-level design with subjects as the first level, where y_{ij} for $i=1,2,\ldots,n_j$ are repeated measurements at the second level on subject j, for $j=1,2,\ldots,J$. Let $\boldsymbol{\eta}$ be a vector of predictor values, partitioned into values for each subject, i.e. $\boldsymbol{\eta}^{\mathrm{T}}=(\boldsymbol{\eta}_1^{\mathrm{T}},\boldsymbol{\eta}_2^{\mathrm{T}},\ldots,\boldsymbol{\eta}_J^{\mathrm{T}})$ of length $n=\Sigma_{j=1}^{J}n_j$. Let $\mathbf{Z}_j$ be an $n\times q_j$ design matrix (for random effects $\boldsymbol{\gamma}_j$ for subject j) having non-zero values for the n_j rows corresponding to subject j, and assume that the $\boldsymbol{\gamma}_j$ are all independent with $\boldsymbol{\gamma}_j\sim N_{q_j}(0,\mathbf{G}_j^{-1})$, for $j=1,2,\ldots,J$. (The $\mathbf{Z}_j$-matrices and random effects $\boldsymbol{\gamma}_j$ for $j=1,2,\ldots,J$ could alternatively be combined into a single design matrix $\mathbf{Z}$ and a single random vector $\boldsymbol{\gamma}$.)

3.3.1.2. Repeated measures with correlated random-effects terms. In Section 3.3.1.1, set $q_j=n_j$ and set the non-zero submatrix of $\mathbf{Z}_j$ to be the identity matrix I_{n_j}, for $j=1,2,\ldots,J$.

This allows various covariance or correlation structures in the random effects of the repeated measurements to be specified by a suitable choice of matrices $\mathbf{G}_j$, as in point (c) in Section 3.2.7.

3.3.1.3. Random- (covariate) coefficients terms. In Section 3.3.1.1 for $j = 1, 2, \ldots, J$, set $q_j = q$ and $\mathbf{G}_j = \mathbf{G}$, i.e. $\gamma_j \sim N_q(0, \mathbf{G}^{-1})$, and set the non-zero submatrix of the design matrices $\mathbf{Z}_j$ suitably by using the covariate(s). This allows the specification of random (covariate) coefficient models.

3.3.1.4. Multilevel (nested) hierarchical model terms. Let each level of the hierarchy be a one-factor random-effect term as in point (b) in Section 3.2.7.

3.3.1.5. Crossed random-effect terms. Let each of the crossed factors be a one-factor random-effect term as in point (b) in Section 3.2.7.

3.3.2. Combinations of random effects and spline terms

There are many useful combinations, e.g. combining random (covariate) coefficients and cubic smoothing spline terms in the same covariate.

3.3.3. Combinations of spline terms

For example, combining cubic smoothing spline terms in different covariates gives the additive model; Hastie and Tibshirani (1990).

4. Specific families of population distribution $f(y|\theta)$

4.1. General comments

The population probability (density) function $f(y|\boldsymbol{\theta})$ in model (1) is deliberately left general with no explicit conditional distributional form for the response variable y. The only restriction that the R implementation of a GAMLSS (Stasinopoulos *et al.*, 2004) has for specifying the distribution of y is that the function $f(y|\boldsymbol{\theta})$ and its first (and optionally expected second and cross-) derivatives with respect to each of the parameters of $\boldsymbol{\theta}$ must be computable. Explicit derivatives are preferable but numerical derivatives can be used (resulting in reduced computational speed). Table 1 shows a variety of one-, two-, three- and four-parameter distributions that the authors have successfully implemented in their software. Johnson *et al.* (1993, 1994, 1995) are the classic references on distributions and cover most of the distributions in Table 1. More information on those distributions which are not covered is provided in Section 4.2. Clearly Table 1 provides a wide selection of distributions from which to choose, but to extend the list to include other distributions is a relatively easy task. For some of the distributions that are shown in Table 1 more that one parameterization has been implemented.

We shall use notation

$$y \sim \mathcal{D}\{g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, \ldots, g_p(\theta_p) = t_p\}$$

to identify uniquely a GAMLSS, where $\mathcal{D}$ is the response variable distribution (as abbreviated in Table 1), $(\theta_1, \ldots, \theta_p)$ are the parameters of $\mathcal{D}$, $(g_1, \ldots, g_p)$ are the link functions and $(t_1, \ldots, t_p)$ are the model formulae for the explanatory terms and/or random effects in the predictors $(\eta_1, \ldots, \eta_p)$ respectively. For example

$$y \sim \mathrm{TF}\{\mu = \mathrm{cs}(x, 3),\ \log(\sigma) = x,\ \log(\nu) = 1\}$$

Table 1. Implemented GAMLSS distributions

Number of parameters	*Distribution*
Discrete, one parameter	Binomial
	Geometric
	Logarithmic
	Poisson
	Positive Poisson
Discrete, two parameters	Beta–binomial
	Generalized Poisson
	Negative binomial type I
	Negative binomial type II
	Poisson–inverse Gaussian
Discrete, three parameters	Sichel
Continuous, one parameter	Exponential
	Double exponential
	Pareto
	Rayleigh
Continuous, two parameters	Gamma
	Gumbel
	Inverse Gaussian
	Logistic
	Log-logistic
	Normal
	Reverse Gumbel
	Weibull
	Weibull (proportional hazards)
Continuous, three parameters	Box–Cox normal (Cole and Green, 1992)
	Generalized extreme family
	Generalized gamma family (Box–Cox gamma)
	Power exponential family
	t-family
Continuous, four parameters	Box–Cox *t*
	Box–Cox power exponential
	Johnson–Su original
	Reparameterized Johnson–Su

is a model where the response variable y has a t-distribution with the location parameter μ modelled, using an identity link, as a cubic smoothing spline with three effective degrees of freedom in x on top of the linear term in x, i.e. cs(x,3), the scale parameter σ modelled by using a log-linear model in x and the t-distribution degrees-of-freedom parameter ν modelled by using a constant model denoted 1 (but on the log-scale).

Quantile residuals (Section 6.2) are obtained easily provided that the cumulative distribution function (CDF) can be computed, and centile estimation is achieved easily provided that the inverse CDF can be computed. This applies to the continuous distributions in Table 1 which transform to simple standard distributions, whereas the CDF and inverse CDF of the discrete distributions can be computed numerically, if necessary.

Censoring can be incorporated easily in a GAMLSS. For example, assume that an observation is randomly right censored at value y; then its contribution to the log-likelihood l is given by $\log\{1 - F(y|\boldsymbol{\theta})\}$, where $F(y|\boldsymbol{\theta})$ is the CDF of y. Hence, the incorporation of censoring requires functions for computing $F(y|\boldsymbol{\theta})$ and also its first (and optionally expected second and cross-) derivatives with respect to each of the parameters $(\theta_1, \theta_2, \ldots, \theta_p)$ in the fitting algorithm. This has been found to be straightforward for the distributions in Table 1 for which an explicit form for the CDF exists. Similarly, truncated distributions are easily incorporated in a GAMLSS.

4.2. Specific distributions

Many three- and four-parameter families of continuous distribution for y can be defined by assuming that a transformed variable z, obtained from y, has a simple well-known distribution.

The *Box–Cox normal* family for $y>0$ which was used by Cole and Green (1992), denoted by BCN(μ,σ,ν), reparameterized from Box and Cox (1964), assumes that z has a standard normal distribution $N(0,1)$ with mean 0 and variance 1 where

$$z=\begin{cases}\dfrac{1}{\sigma\nu}\left\{\left(\dfrac{y}{\mu}\right)^{\nu}-1\right\}, & \text{if } \nu\neq 0,\\ \dfrac{1}{\sigma}\log\left(\dfrac{y}{\mu}\right), & \text{if } \nu=0.\end{cases} \tag{7}$$

Cole and Green (1992) were the first to model all three parameters of a distribution as nonparametric smooth functions of a single explanatory variable.

The *generalized gamma* family for $y>0$, as parameterized by Lopatatzidis and Green (2000), denoted by GG(μ,σ,ν), assumes that z has a gamma GA$(1,\sigma^2\nu^2)$ distribution with mean 1 and variance $\sigma^2\nu^2$, where $z=(y/\mu)^{\nu}$, for $\nu>0$.

The *power exponential* family for $-\infty<y<\infty$ which was used by Nelson (1991), denoted by PE(μ,σ,ν), a reparameterization of that of Box and Tiao (1973), assumes that z has a gamma GA$(1,\nu)$ distribution with mean 1 and variance ν, where

$$z=\frac{\nu}{2}\left|\frac{y-\mu}{\sigma\, c(\nu)}\right|^{\nu},$$

and the function

$$c(\nu)=\left\{2^{-2/\nu}\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}\right\}^{1/2},$$

from Nelson (1991), where $\nu>0$. For this parameterization μ and σ are the mean and standard deviation of y respectively.

The *Student* t-family for $-\infty<y<\infty$ (e.g. Lange *et al.* (1989)), denoted by TF(μ,σ,ν), assumes that z has a standard t-distribution with ν degrees of freedom, where $z=(y-\mu)/\sigma$.

The four-parameter *Box–Cox* t-family for $y>0$, denoted by BCT(μ,σ,ν,τ), is defined by assuming that z given by expression (7) has a standard t-distribution with τ degrees of freedom; Rigby and Stasinopoulos (2004a).

The *Box–Cox power exponential* family for $y>0$, denoted BCPE(μ,σ,ν,τ), is defined by assuming that z given by expression (7) has a standard power exponential distribution; Rigby and Stasinopoulos (2004b). This distribution is useful for modelling (positive or negative) skewness combined with (lepto or platy) kurtosis in continuous data.

The *Johnson–Su* family for $-\infty<y<\infty$, denoted by JSU$_0(\mu,\sigma,\nu,\tau)$ (Johnson, 1949), is defined by assuming that $z=\nu+\tau\,\sinh^{-1}\{(y-\mu)/\sigma\}$ has a standard normal distribution. The *reparameterized Johnson–Su* family, denoted by JSU(μ,σ,ν,τ), has mean μ and standard deviation σ for all values of ν and τ.

5. The algorithms

Two basic algorithms are used for maximizing the penalized likelihood that is given in equation (5). The first, the CG algorithm, is a generalization of the Cole and Green (1992) algorithm (and uses the first and (expected or approximated) second and cross-derivatives of the likelihood function with respect to the parameters $\boldsymbol{\theta}$). However, for many population probability (density)

functions $f(y|\boldsymbol{\theta})$ the parameters $\boldsymbol{\theta}$ are information orthogonal (since the expected values of the cross-derivatives of the likelihood function are 0), e.g. location and scale models and dispersion family models, or approximately so. In this case the simpler RS algorithm, which is a generalization of the algorithm that was used by Rigby and Stasinopoulos (1996a, b) for fitting mean and dispersion additive models (and does not use the cross-derivatives), is more suited. The parameters $\boldsymbol{\theta}$ are fully information orthogonal for only the negative binomial, gamma, inverse Gaussian, logistic and normal distributions in Table 1. Nevertheless, the RS algorithm has been successfully used for fitting all the distributions in Table 1, although occasionally it can be slow to converge. Note also that the RS algorithm is not a special case of the CG algorithm, as explained in Appendix B.

The object of the algorithms is to maximize the penalized likelihood function l_p, given by equation (5), for fixed hyperparameters $\boldsymbol{\lambda}$. The details of the algorithms are given in Appendix B, whereas the justification that the CG algorithm maximizes the penalized likelihood l_p, given by equation (5), is provided in Appendix C. The justification for the RS algorithm is similar.

The algorithms are implemented in the option `method` in the function `gamlss()` within the R package GAMLSS (Stasinopoulos *et al.*, 2004), where a combination of both algorithms is also allowed. The major advantages of the two algorithms are

(a) the modular fitting procedure (allowing different model diagnostics for each distribution parameter),
(b) easy addition of extra distributions,
(c) easy addition of extra additive terms and
(d) easily found starting values since they only require initial values for the $\boldsymbol{\theta}$- rather than for the $\boldsymbol{\beta}$-parameters.

The algorithms have generally been found to be stable and fast using very simple starting values (e.g. constants) for the $\boldsymbol{\theta}$-parameters.

Clearly, for a specific data set and model, the (penalized) likelihood can potentially have multiple local maxima. This is investigated by using different starting values and has generally not been found to be a problem in the data sets that were analysed, possibly because of the relatively large sample sizes that were used.

Singularities in the likelihood function that are similar to those that were reported by Crisp and Burridge (1994) can potentially occur in specific cases within the GAMLSS framework, especially when the sample size is small. The problem can be alleviated by appropriate restrictions on the scale parameter (penalizing it for going close to 0).

6. Model selection

6.1. Statistical modelling

Let $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}$ represent the GAMLSS, where

(a) $\mathcal{D}$ specifies the distribution of the response variable,
(b) $\mathcal{G}$ specifies the set of link functions $(g_1, \ldots, g_p)$ for parameters $(\theta_1, \ldots, \theta_p)$,
(c) $\mathcal{T}$ specifies the set of predictor terms $(t_1, \ldots, t_p)$ for predictors $(\eta_1, \ldots, \eta_p)$ and
(d) $\boldsymbol{\lambda}$ specifies the set of hyperparameters.

For a specific data set, the GAMLSS model building process consists of comparing many different competing models for which different combinations of components $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}$ are tried.

Inference about quantities of interest can be made either conditionally on a single selected 'final' model or by averaging between selected models. Conditioning on a single final model was criticized by Draper (1995) and Madigan and Raftery (1994) since it ignores model uncertainty and generally leads to an underestimation of the uncertainty about quantities of interest. Averaging between selected models can reduce this underestimation; Hjort and Claeskens (2003).

As with all scientific inferences the determination of the adequacy of any model depends on the substantive question of interest and requires subject-specific knowledge.

6.2. Model selection, inference and diagnostics

For parametric GAMLSS models each model $\mathcal{M}$ of the form (2) can be assessed by its fitted global deviance GD given by $\mathrm{GD} = -2\,l(\hat{\boldsymbol{\theta}})$ where $l(\hat{\boldsymbol{\theta}}) = \Sigma_{i=1}^{n} l(\hat{\boldsymbol{\theta}}^i)$. Two nested parametric GAMLSS models, $\mathcal{M}_0$ and $\mathcal{M}_1$, with fitted global deviances GD_0 and GD_1 and error degrees of freedom $\mathrm{df}_{\mathrm{e}0}$ and $\mathrm{df}_{\mathrm{e}1}$ respectively may be compared by using the (generalized likelihood ratio) test statistic $\Lambda = \mathrm{GD}_0 - \mathrm{GD}_1$ which has an asymptotic χ^2-distribution under $\mathcal{M}_0$, with degrees of freedom $d = \mathrm{df}_{\mathrm{e}0} - \mathrm{df}_{\mathrm{e}1}$ (given that the regularity conditions are satisfied). For each model $\mathcal{M}$ the error degrees of freedom parameter df_{e} is defined by $\mathrm{df}_{\mathrm{e}} = n - \Sigma_{k=1}^{p} \mathrm{df}_{\theta_k}$, where df_{θ_k} are the degrees of freedom that are used in the predictor model for parameter θ_k for $k = 1, \ldots, p$.

For comparing non-nested GAMLSSs (including models with smoothing terms), to penalize overfitting the generalized Akaike information criterion GAIC (Akaike, 1983) can be used. This is obtained by adding to the fitted global deviance a fixed penalty # for each effective degree of freedom that is used in a model, i.e. $\mathrm{GAIC}(\#) = \mathrm{GD} + \#\mathrm{df}$, where df denotes the total effective degrees of freedom used in the model and GD is the fitted global deviance. The model with the smallest value of the criterion GAIC(#) is then selected. The Akaike information criterion AIC (Akaike, 1974) and the Schwarz Bayesian criterion SBC (Schwarz, 1978) are special cases of the GAIC(#) criterion corresponding to $\# = 2$ and $\# = \log(n)$ respectively. The two criteria, AIC and SBC, are asymptotically justified as predicting the degree of fit in a new data set, i.e. approximations to the average predictive error. A justification for the use of SBC comes also as a crude approximation to Bayes factors; Raftery (1996, 1999). Claeskens and Hjort (2003) considered a focused information criterion in which the criterion for model selection depends on the objective of the study, in particular on the specific parameter of interest. Using GAIC(#) allows different penalties # to be tried for different modelling purposes. The sensitivity of the selected model to the choice of # can also be investigated.

For GAMLSSs with hyperparameters $\boldsymbol{\lambda}$, the hyperparameters can be estimated by one of the methods that are described in Appendix A.2. Different random-effect models (for the same fixed effects models) can be compared by using their maximized (Laplace approximated) profile marginal likelihood of $\boldsymbol{\lambda}$ (eliminating both fixed and random effects), $l(\hat{\boldsymbol{\lambda}})$, given by equation (14) in Appendix A.2.3 in the way that Lee and Nelder (1996, 2001a, b) used their adjusted profile h-likelihood. Different fixed effects models (for the same random-effects models) can be compared by using their approximate maximized (Laplace approximated) marginal likelihood of $\boldsymbol{\beta}$ (eliminating the random effects $\boldsymbol{\gamma}$), i.e. $l(\hat{\boldsymbol{\beta}}) \approx l_h(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) - \frac{1}{2}\log|\hat{\mathbf{H}}/2\pi|$, where $\hat{\mathbf{H}} = -E(\partial^2 l_h/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}})$ evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ and l_h is defined in Section 2.2, conditional on chosen hyperparameters.

To test whether a specific fixed effect predictor parameter is different from 0, a χ^2-test is used, comparing the change in global deviance Λ for parametric models (or the change in the approximate marginal deviance (eliminating the random effects) for random-effects models) when the parameter is set to 0 with a χ_1^2 critical value. Profile (marginal) likelihood for fixed effect model parameters can be used for the construction of confidence intervals. The above test and confidence intervals are conditional on any hyperparameters being fixed at selected values.

An alternative approach, which is suitable for very large data sets, is to split the data into

(a) training,
(b) validation and
(c) test data sets

and to use them for model fitting, selection and assessment respectively; Ripley (1996) and Hastie *et al.* (2001).

For each $\mathcal{M}$ the (normalized randomized quantile) residuals of Dunn and Smyth (1996) are used to check the adequacy of $\mathcal{M}$ and, in particular, the distribution component $\mathcal{D}$. The (normalized randomized quantile) residuals are given by $\hat{r}_i = \Phi^{-1}(u_i)$ where Φ^{-1} is the inverse CDF of a standard normal variate and $u_i = F(y_i|\hat{\boldsymbol{\theta}}^i)$ if y_i is an observation from a continuous response, whereas u_i is a random value from the uniform distribution on the interval $[F(y_i - 1|\hat{\boldsymbol{\theta}}^i), F(y_i|\hat{\boldsymbol{\theta}}^i)]$ if y_i is an observation from a discrete integer response, where $F(y|\boldsymbol{\theta})$ is the CDF. For a right-censored continuous response u_i is defined as a random value from a uniform distribution on the interval $[F(y_i|\hat{\boldsymbol{\theta}}^i), 1]$. Note that, when randomization is used, several randomized sets of residuals (or a median set from them) should be studied before a decision about the adequacy of model $\mathcal{M}$ is taken. The true residuals r_i have a standard normal distribution if the model is correct.

7. Examples

The following five examples are used primarily to demonstrate the power and flexibility of GAMLSSs.

7.1. Dutch girls' body mass index data example

The variables body mass index BMI and age were recorded for 20243 Dutch girls in a cross-sectional study of growth and development in the Dutch population in 1980; Cole and

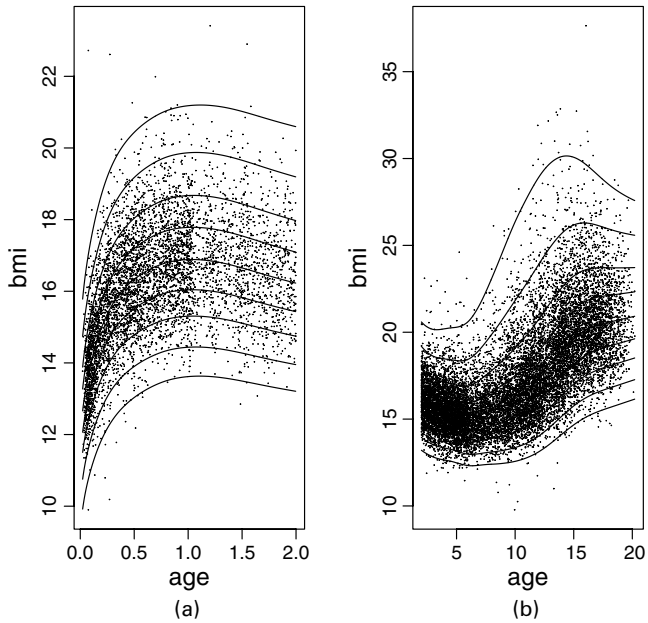

Fig. 1. Body mass index data: BMI against age with fitted centile curves

Roede (1999). The objective here is to obtain smooth reference centile curves for BMI against age.

Figs 1(a) and 1(b) provide plots of BMI against age, separately for age ranges 0–2 years and 2–21 years respectively for clarity of presentation, indicating a positively skew (and possibly leptokurtic) distribution for BMI given age and also a non-linear relationship between the location (and possibly also the scale, skewness and kurtosis) of BMI with age. Previous modelling of the variable BMI (e.g. Cole *et al.* (1998)) using the *LMS* method of Cole and Green (1992), has found significant kurtosis in the residuals after fitting the model, indicating that the kurtosis was not adequately modelled. It has also previously been found (e.g. Rigby and Stasinopoulos (2004a)) that a power transformation of age to explanatory variable $X = \text{age}^{\xi}$ improves the model fit substantively in similar data analysis.

Hence, given $X = x$, the dependent variable BMI, denoted y, was modelled by using a Box–Cox t-distribution BCT(μ, σ, ν, τ) from Section 4.2, where the parameters μ, σ, ν and τ are modelled as smooth nonparametric functions of x, i.e. assume, given $X = x_i$, that $y_i \sim \text{BCT}(\mu_i, \sigma_i, \nu_i, \tau_i)$, independently for $i = 1, 2, \ldots, n$, where

$$\left.\begin{aligned} \mu_i &= h_1(x_i), \\ \log(\sigma_i) &= h_2(x_i), \\ \nu_i &= h_3(x_i), \\ \log(\tau_i) &= h_4(x_i). \end{aligned}\right\} \tag{8}$$

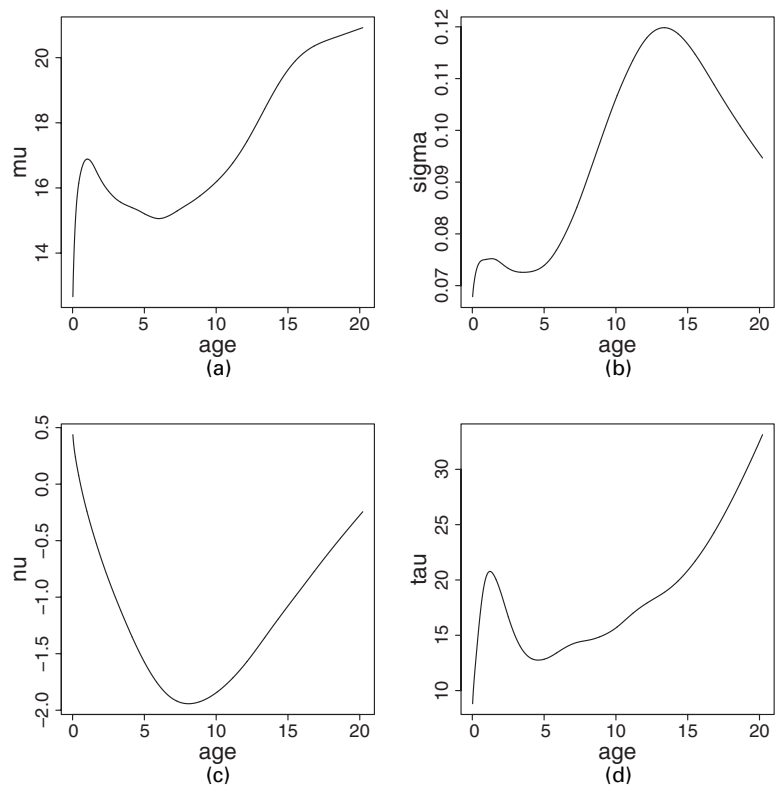

Fig. 2. Body mass index data: fitted parameters (a) μ, (b) σ, (c) ν and (d) τ against age

Here $h_k(x)$ are arbitrary smooth functions of x for $k = 1, 2, 3, 4$ as in Section 3.2.1, and $x_i = \text{age}_i^{\xi}$ for $i = 1, 2, \ldots, n$, where ξ is a non-linear parameter in the model. Log-link functions were used for σ and τ in expression (8) to ensure that $\sigma > 0$ and $\tau > 0$.

In the model fitting, the above model is denoted $y \sim \text{BCT}\{\mu = \text{cs}(x, \text{df}'_\mu), \log(\sigma) = \text{cs}(x, \text{df}'_\sigma), \nu = \text{cs}(x, \text{df}'_\nu), \log(\tau) = \text{cs}(x, \text{df}'_\tau)\}$ where df′ indicates the extra degrees of freedom on top of a linear term in x. For example, in the model for μ, the total degrees of freedom used are $\text{df}_\mu = 2 + \text{df}'_\mu$. Hence x or $\text{cs}(x, 0)$ refers to a linear model in x.

Model selection was achieved by minimizing the generalized Akaike information criterion GAIC(#), which is discussed in Section 6.2 and Appendix A.2.1, with penalty $\# = 2.4$, over the parameters $\text{df}_\mu, \text{df}_\sigma, \text{df}_\nu, \text{df}_\tau$ and ξ using the numerical optimization algorithm L-BFGS-B in function `optim` (from the R package; Ihaka and Gentleman (1996)), which is incorporated in the GAMLSS package. The algorithm converged to the values $(\text{df}_\mu, \text{df}_\sigma, \text{df}_\nu, \text{df}_\tau, \xi) = (16.2, 8.5, 4.7, 6.1, 0.50)$, correct to the decimal places given, with total effective degrees of freedom equal to 36.5 (including one for the parameter ξ), global deviance $\text{GD} = 76454.5$ and $\text{GAIC}(2.4) = 76542.1$, and this was the model selected. (The choice of penalty, which was selected here to demonstrate flexible modelling of the parameters, affects particularly the fitted τ model for this data set. For example, a penalty of $\# = 2.5$ led to a model with $(\text{df}_\mu, \text{df}_\sigma, \text{df}_\nu, \text{df}_\tau, \xi) = (16.0, 8.0, 4.8, 1, 0.52)$ with a constant τ model, $\text{GD} = 76468.1$ and $\text{GAIC}(2.5) = 76545.1$).

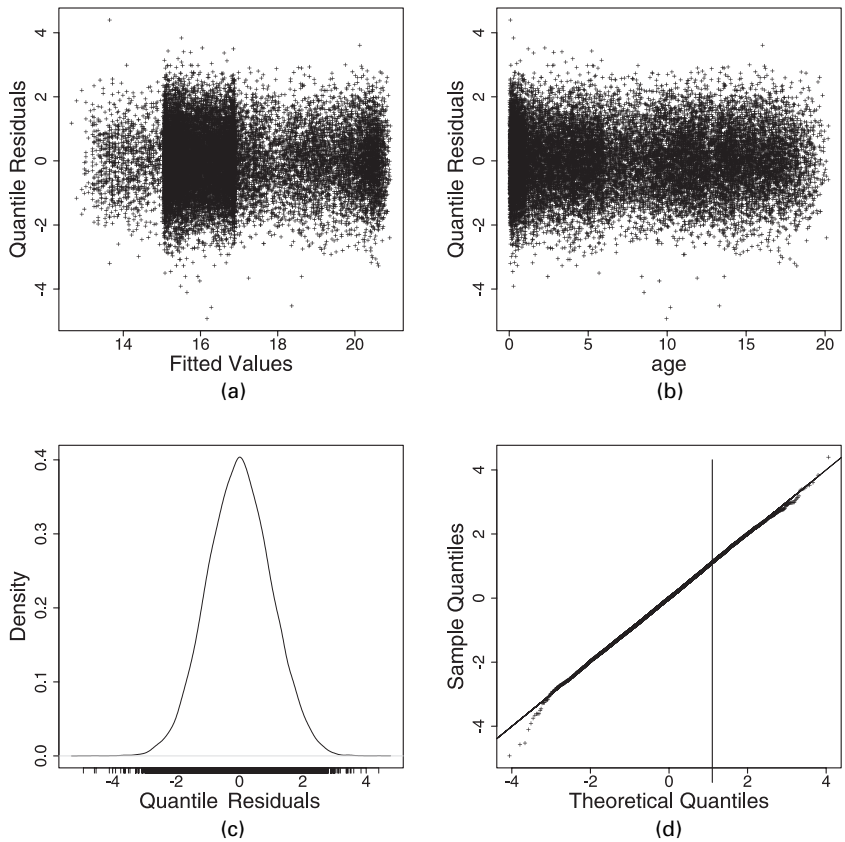

Fig. 3. Body mass index data: (a) residuals against fitted values of μ, (b) residuals against age, (c) kernel density estimate and (d) *QQ*-plot

The fitted models for μ, σ, ν and τ for the selected model are displayed in Fig. 2. The fitted ν indicates positive skewness in BMI for all ages (since $\hat{\nu} < 1$), whereas the fitted τ indicates modest leptokurtosis particularly at the lower ages. Fig. 3 displays the (normalized quantile) residuals, which were defined in Section 6.2, from the fitted model. Figs 3(a) and 3(b) plot the residuals against the fitted values of μ and against age respectively, whereas Figs 3(c) and 3(d) provide a kernel density estimate and normal QQ-plot for them respectively. The residuals appear random, although the QQ-plot shows a possible single outlier in the upper tail and a slightly longer extreme (0.06%) lower tail than the Box–Cox t-distribution. Nevertheless the model provides a good fit to the data. The fitted model centile curves for BMI for centiles $100\alpha = 0.4$, 2.3, 10, 25, 50, 75, 90, 97.7, 99.6 (chosen to be two-thirds of a z-score apart) are displayed in Figs 1(a) and 1(b) for age ranges 0–2 years and 2–21 years respectively.

7.2. Hodges's health maintenance organization data example

Here we consider a one-factor random-effects model for response variable health insurance premium (prind) with state as the random factor. The data were analysed in Hodges (1998).

Hodges (1998) modelled the data by using a normal conditional model for y_{ij} given γ_j the random effect in the mean for state j, and a normal distribution for γ_j, i.e. his model can be expressed by $y_{ij}|\mu_{ij}, \sigma \sim N(\mu_{ij}, \sigma^2)$, $\mu_{ij} = \beta_1 + \gamma_j$, $\log(\sigma) = \beta_2$ and $\gamma_j \sim N(0, \sigma_1^2)$, independently for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, J$, where i indexes the observations within states.

Fig. 4 provides box plots of prind against state, showing the variation in the location and scale

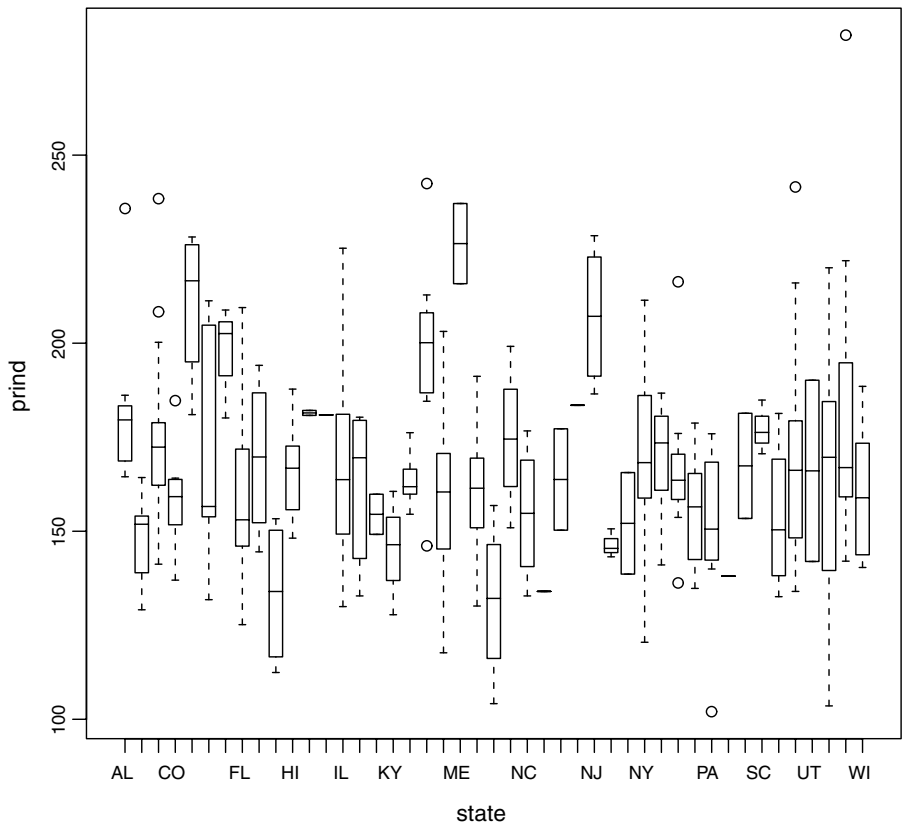

Fig. 4. Health maintenance organization data: box plots of prind against state

of prind between states and a positively skewed (and possible leptokurtic) distribution of prind within states. Although Hodges (1998) used an added variable diagnostic plot to identify the need for a Box–Cox transformation of y, he did not model the data by using a transformation of y.

In the discussion of Hodges (1998), Wakefield commented as follows.

> 'If it were believed that there were different within-state variances then one possibility would be to assume a hierarchy for these also.'

Hodges, in his reply, also suggested treating the 'within-state precisions or variances as draws from some distribution'.

Hence we consider a Box–Cox t-model which allows for both skewness and kurtosis in the conditional distribution of y given the parameters $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$. We allow for possible differences between states in the location, scale and shape of the conditional distribution of y, by including a random-effect term in each of the models for the parameters μ, σ, ν and τ, i.e. we assume a general model where, independently for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, J$,

$$y_{ij} | \mu_{ij}, \sigma_{ij}, \nu_{ij}, \tau_{ij} \sim \mathrm{BCT}(\mu_{ij}, \sigma_{ij}, \nu_{ij}, \tau_{ij})$$

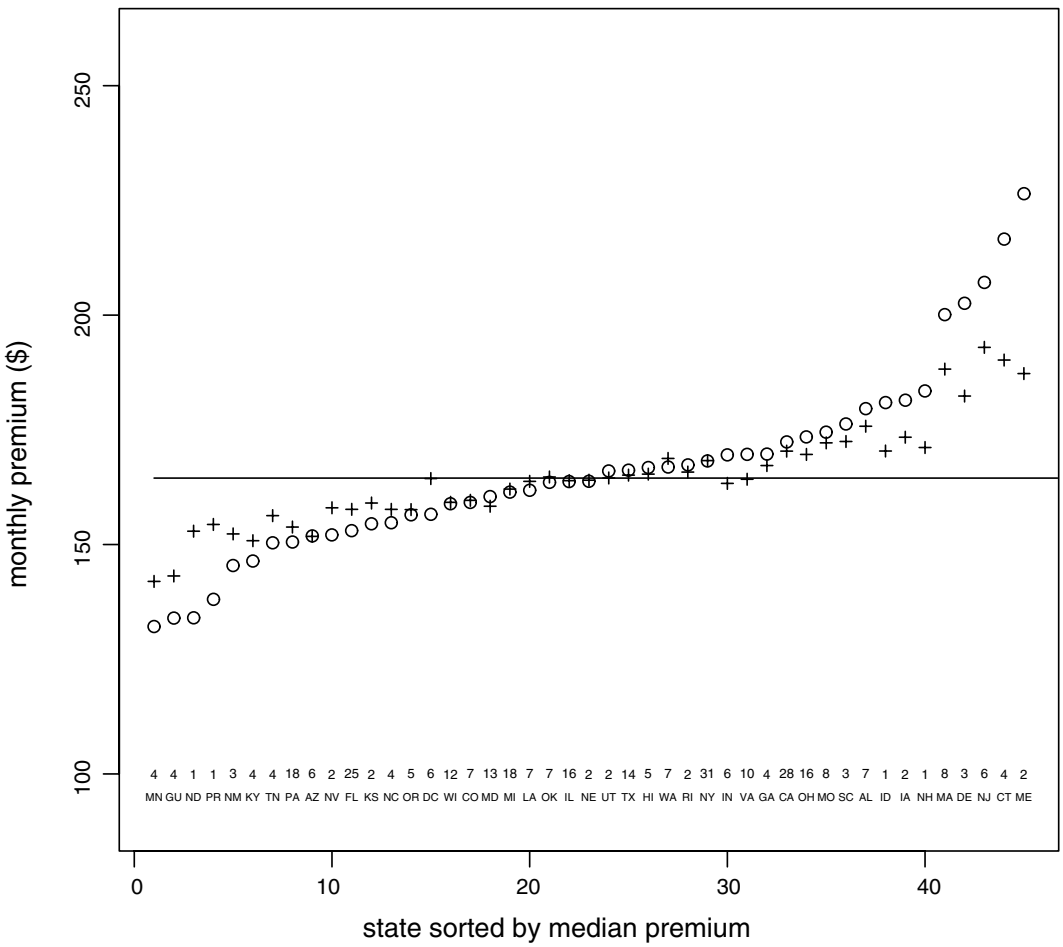

Fig. 5. Health maintenance organization data: sample (○) and fitted (+) medians of prind against state

where $\mu_{ij} = \beta_1 + \gamma_{j1}$, $\log(\sigma_{ij}) = \beta_2 + \gamma_{j2}$, $\nu_{ij} = \beta_3 + \gamma_{j3}$ and $\log(\tau_{ij}) = \beta_4 + \gamma_{j4}$, and where $\gamma_{jk} \sim N(0, \sigma_k^2)$ independently for $j = 1, 2, \ldots, J$ and $k = 1, 2, 3, 4$.

Using an Akaike information criterion, i.e. GAIC(2), for hyperparameter selection, as discussed in Section 6.2 and Appendix A.2.1, led to the conclusion that the random-effect parameters for ν and τ are not needed, i.e. $\sigma_3 = \sigma_4 = 0$. The remaining random-effect parameters were estimated by using the approximate marginal likelihood approach, which is described in Appendix A.2.3, giving fitted parameter values $\hat{\sigma}_1 = 13.14$ and $\hat{\sigma}_2 = 0.0848$ with corresponding fixed effects parameter values $\hat{\beta}_1 = 164.8$, $\hat{\beta}_2 = -2.213$, $\hat{\beta}_3 = -0.0697$ and $\hat{\beta}_4 = 2.148$ and an approximate marginal deviance of 3118.62 obtained from equation (14) in Appendix A.2.3. This was the chosen fitted model.

Since $\hat{\nu} = \hat{\beta}_3 = -0.0697$ is close to 0, the fitted conditional distribution of y_{ij} is approximately defined by $\hat{\sigma}_{ij}^{-1} \log(y_{ij}/\hat{\mu}_{ij}) \sim t_{\hat{\tau}}$, a t-distribution with $\hat{\tau} = \exp(\hat{\beta}_4) = 8.57$ degrees of freedom, for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, J$.

Fig. 5 plots the sample and fitted medians ($\boldsymbol{\mu}$) of prind against state (ordered by the sample median). The fitted values of σ (which are not shown here) vary very little. The heterogeneity in the sample variances of prind between the states (in Fig. 4) seems to be primarily due to sampling variation caused by the high skewness and kurtosis in the conditional distribution of y (rather than either the variance–mean relationship or the random effect in σ). Fig. 6 provides marginal (Laplace-approximated) profile deviance plots, as described in Section 6.2, for each of ν and τ, for fixed hyperparameters, giving 95% intervals $(-0.866, 0.788)$ for ν and $(4.6, 196.9)$ for τ, indicating considerable uncertainty about these parameters. (The fitted model suggests a log-transformation for y, whereas the added variable plot that was used by Hodges (1998) suggested a Box–Cox transformation parameter $\nu = 0.67$ which, although rather different, still lies within the 95% interval for ν. Furthermore the wide interval for τ suggests that a conditional distribution model for y_{ij} defined by $\sigma_{ij}^{-1} \log(y_{ij}/\mu_{ij}) \sim N(0, 1)$ may provide a reasonable model. This model has $\hat{\sigma}_1 = 13.07$ and $\hat{\sigma}_2 = 0.105$.)

Fig. 7(a) provides a normal QQ-plot for the (normalized quantile) residuals, which were defined in Section 6.2, for the chosen model. Fig. 7(a) indicates an adequate model for the conditional distribution of y. The outlier case for Washington state, identified by Hodges (1998), does not appear to be an outlier in this analysis. Figs 7(b) and 7(c) provide respectively normal QQ-plots for the fitted random effects γ_{j1} for μ and γ_{j2} for $\log(\sigma)$, for $j = 1, 2, \ldots, J$. Fig. 7(b)

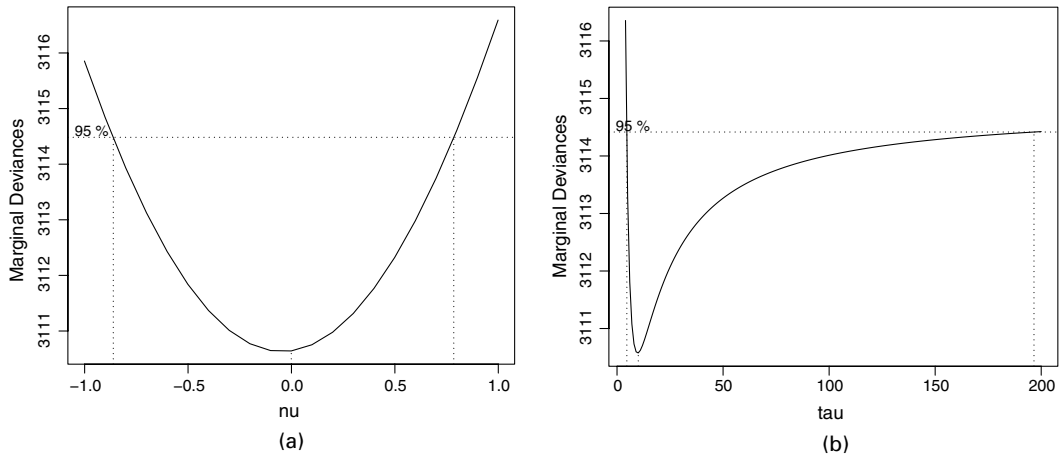

Fig. 6. Health maintenance organization data: profile approximate marginal deviances for (a) ν and (b) τ

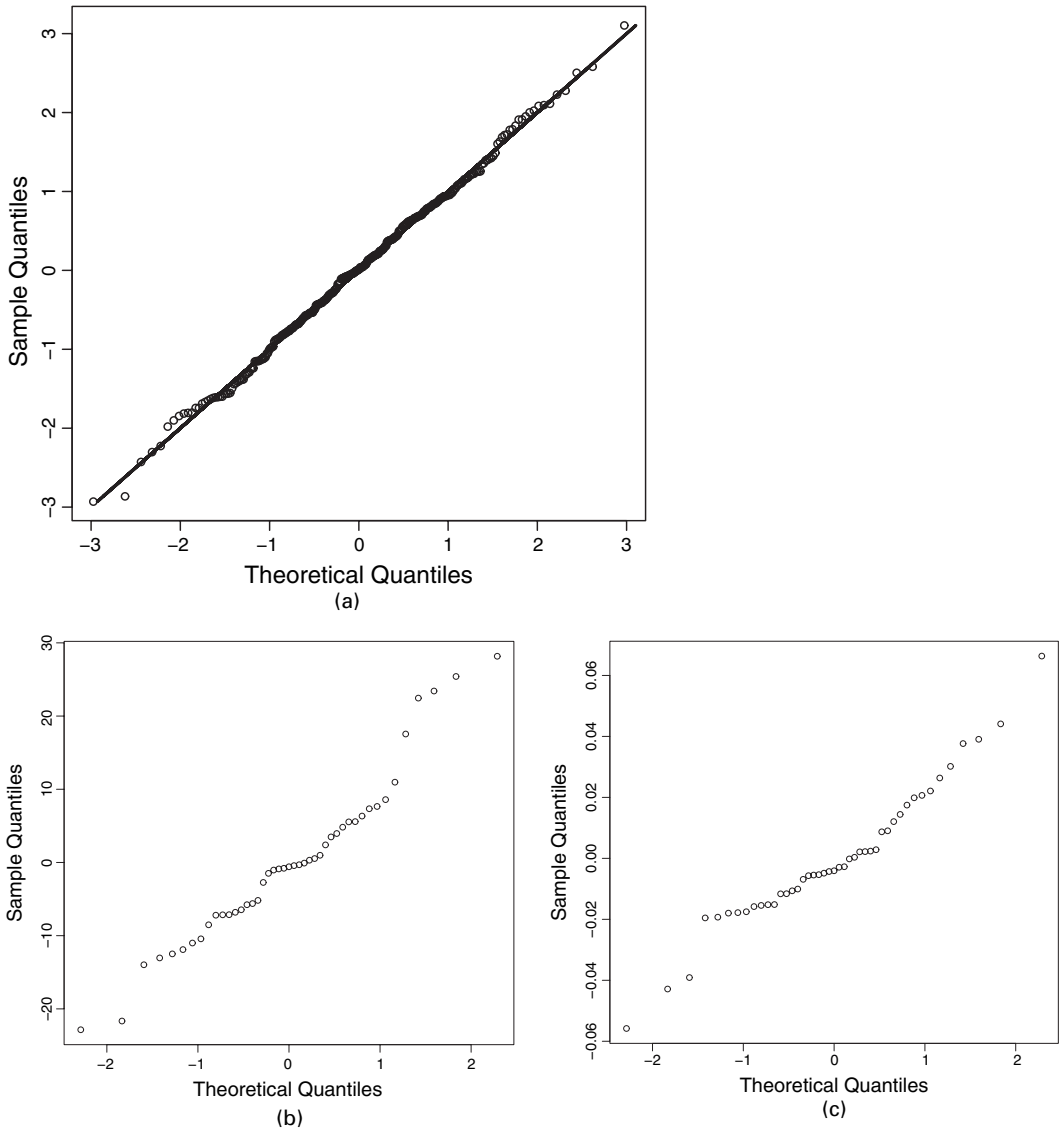

Fig. 7. Health maintenance organization data: *QQ*-plots for (a) the residuals, (b) the random effects in μ and (c) the random effects in $\log(\sigma)$

indicates that the normal distribution for the random effects in the model for μ may be adequate, although there appear to be five outlier states with high prind medians, i.e. states CT, DE, MA, ME and NJ, and also possibly two outlier states with low prind medians, GU and MN. Fig. 7(c) indicates some departure from the assumption of normal random effects in the model for $\log(\sigma)$.

7.3. The hospital stay data

The hospital stay data, 1383 observations, are from a study at the Hospital del Mar, Barcelona, during the years 1988 and 1990; see Gange *et al.* (1996). The response variable is the number of

Table 2. Models for the hospital stay data

Model	*Link*	*Terms*	*GD*	*AIC*	*SBC*
I	logit(μ)	ward + loglos + year	4519.4	4533.4	4570.1
	log(σ)	year			
II	logit(μ)	ward + loglos + year	4483.0	4501.0	4548.1
	log(σ)	year + ward			
III	logit(μ)	ward + cs(loglos,1) + year	4459.4	4479.4	4531.8
	log(σ)	year + ward			
IV	logit(μ)	ward + cs(loglos,1) + year + cs(age,1)	4454.4	4478.4	4541.2
	log(σ)	year + ward			

inappropriate days (noinap) out of the total number of days (los) that patients spent in hospital. The following variables were used as explanatory variables:

(a) age, the age of the patient;
(b) ward, the type of ward in the hospital (medical, surgical or other);
(c) year, the year (1988 or 1990);
(d) loglos, log(los/10).

Gange *et al.* (1996) used a logistic regression model for the number of inappropriate days, with binomial and beta–binomial errors, and found that the latter provided a better fit to the data. They modelled both the mean and the dispersion of the beta–binomial distribution as functions of explanatory variables by using the epidemiological package EGRET (Cytel Software Corporation, 2001), which allowed them to fit a parametric model using a logit link for the mean and an identity link for the dispersion $\phi = \sigma$. Their final model was $\mathrm{BB}\{\mathrm{logit}(\mu) = \mathrm{ward} + \mathrm{year} + \mathrm{loglos}, \sigma = \mathrm{year}\}$.

First we fit their final model, which is equivalent to model I in Table 2. Although we use a log-link for the dispersion σ in Table 2, this does not affect model I since year is a factor. Table 2 shows GD, AIC and SBC, which were defined in Section 6.2, for model I, to be 4519.4, 4533.4 and 4570.1 respectively. Here we are interested in whether we can improve the model by using the flexibility of a GAMLSS. For the dispersion parameter model we found that the addition of ward improves the fit (see model II in Table 2 with AIC = 4501.0 and SBC = 4548.1) but no other term was found to be significant. Non-linearities in the mean model for the terms loglos and age were investigated by using cubic smoothing splines in models III and IV. There is strong support for including a smoothing term for loglos as indicated by the reduction in AIC and SBC for model III compared with model II. The inclusion of a smoothing term for age is not so clear cut since, although there is some marginal support from AIC, it is clearly not supported by SBC, when comparing model III with model IV.

The fitted smoothing functions for loglos and age from model IV are shown in Fig. 8. Fig. 9 displays a set of the (normalized randomized quantile) residuals (see Section 6.2) from model IV. The residuals seem to be satisfactory. Other sets of (normalized randomized quantile) residuals were very similar.

7.4. The epileptic seizure data

The epileptic seizure data, which were obtained from Thall and Vail (1990), comprise four repeated measurements of seizure counts (each over a 2-week period preceding a clinical visit)

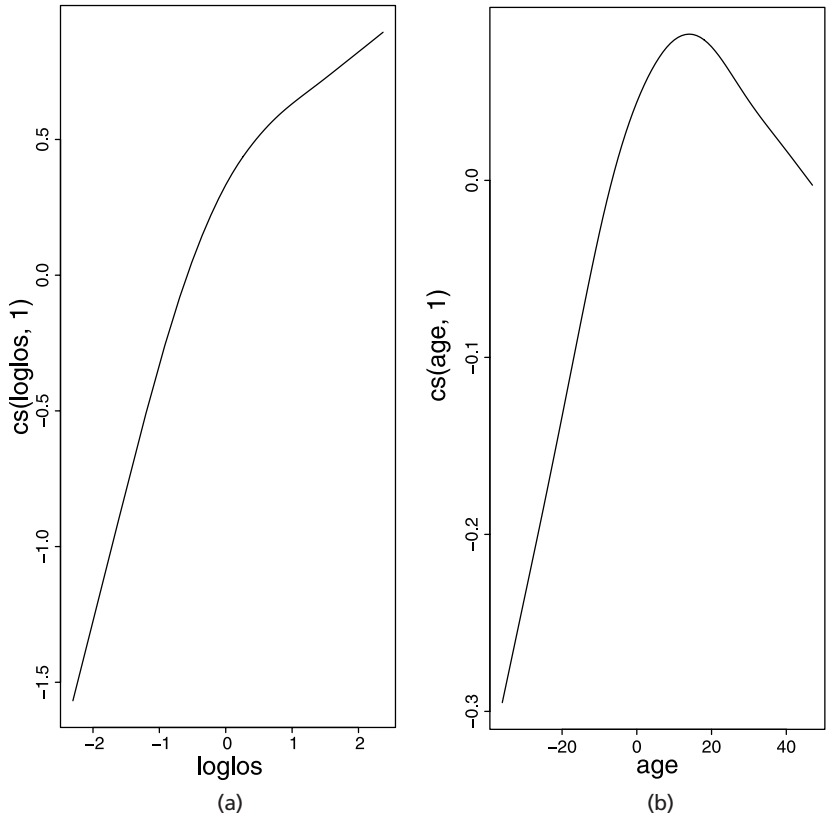

Fig. 8. Hospital stay data: fitted smoothing curves for (a) loglos and (b) age from model IV

for 59 epileptics: a total of 236 cases. Breslow and Clayton (1993) and Lee and Nelder (1996) identified casewise overdispersion in the counts which they modelled by using a random effect for cases in the predictor for the mean in a Poisson GLMM, whereas Lee and Nelder (2000) additionally considered an overdispersed Poisson GLMM (using extended quasi-likelihood). They also identified random effects for subjects in the predictor for the mean.

Here we directly model the casewise overdispersion in the counts by using a negative binomial (type I) model and consider random effects for subjects in the predictors for both the mean and the dispersion. Specifically we assume that, conditional on the mean μ_i and σ_i (i.e. conditional on the random effects), the seizure counts y_{ij} are independent over subjects $i = 1, 2, \ldots, 59$ and repeated measurements $j = 1, 2, 3, 4$ with a negative binomial (type I) distribution, $y_{ij}|\mu_{ij}, \sigma_{ij} \sim \mathrm{NBI}(\mu_{ij}, \sigma_{ij})$ where the logarithm of the mean is modelled by using explanatory terms and the logarithms of both the mean and the dispersion include a random-effects term for subjects. (Note that the conditional variance of y_{ij} is given by $V(y_{ij}|\mu_{ij}, \sigma_{ij}) = \mu_{ij} + \sigma_{ij}\mu_{ij}^2$.)

The model is denoted by NBI{$\log(\mu) =$ lbase $*$ trt + visit + lage + random(subjects), $\log(\sigma) =$ random(subjects)}, where, equivalently to Breslow and Clayton (1993), lbase is the logarithm of a quarter of the number of base-line seizures, trt is a treatment factor (coded 0 for placebo and 1 for drug), visit is a covariate for the clinic visits (coded -0.3, -0.1, 0.1, 0.3 for the four visits), lage is the logarithm of the age of the subject, lbase $*$ trt indicates an interaction term and random(subjects) indicates a random-effect term for subjects with distribution $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ in the log-mean and log-dispersion models respectively.

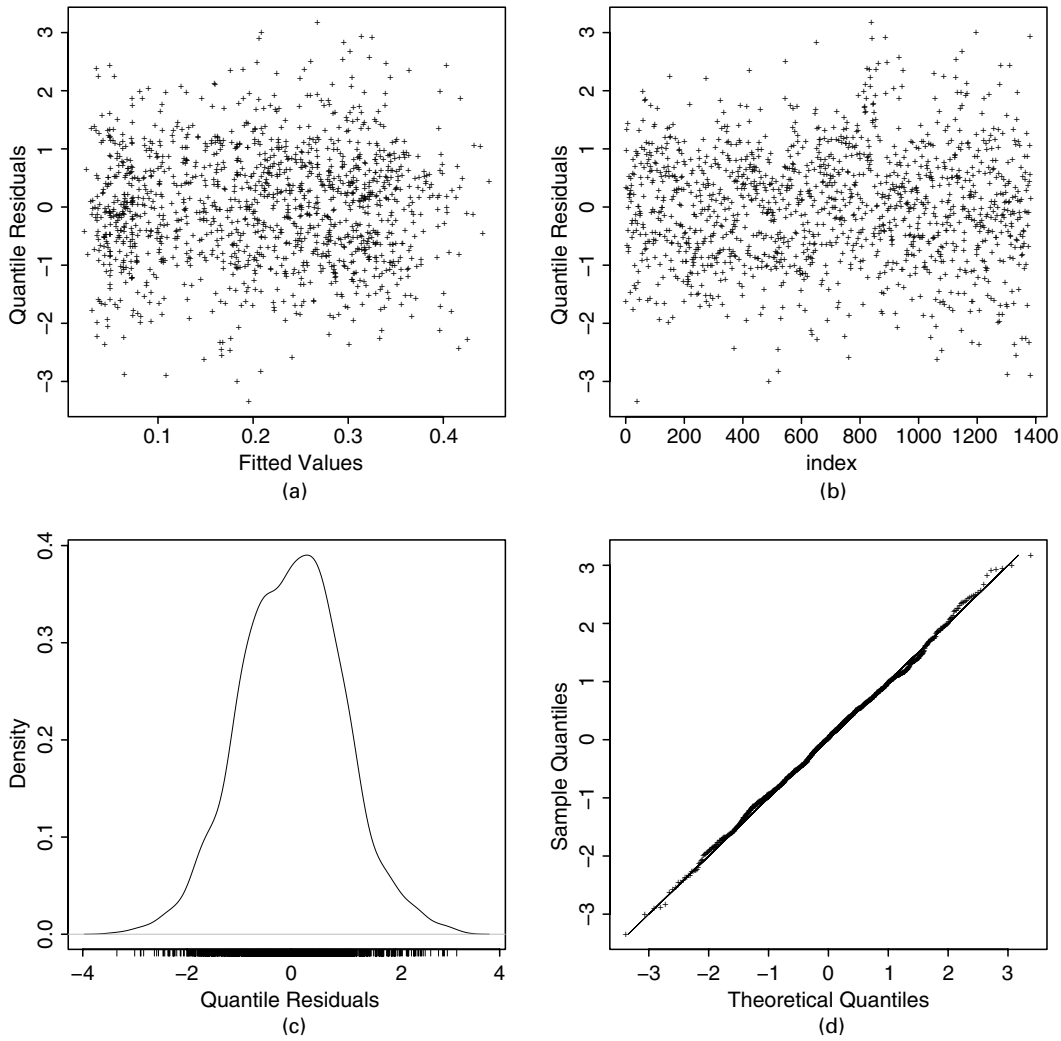

Fig. 9. Hospital stay data: (a) residuals against fitted values, (b) residuals against index, (c) kernel density estimate and (d) *QQ*-plot

The approximate marginal likelihood approach that is described in Appendix A.2.3 led to the fitted random-effects parameters $\hat{\sigma}_1 = 0.465$ and $\hat{\sigma}_2 = 1.056$ with an approximate marginal deviance of 1250.84, obtained from equation (14). (Alternatively, using a generalized Akaike information criterion with penalty 3, i.e. GAIC(3), for hyperparameter selection, as discussed in Appendix A.2.1, led to $\hat{\sigma}_1 = 0.414$ and $\hat{\sigma}_2 = 1.202$ (corresponding to $\mathrm{df}_\mu = 39.9$ and $\mathrm{df}_\sigma = 9.99$ respectively) with GAIC(3) = 1255.7.) Hence it appears that there are random effects for subjects in both the log-mean and the log-dispersion models of the negative binomial distribution of the seizure count. The fitted parameters for $\log(\mu)$ are the intercept $\hat{\beta}_1 = 0.2786$, $\hat{\beta}_{\mathrm{trt}} = -0.3345$, $\hat{\beta}_{\mathrm{lbase}} = 0.9034$, $\hat{\beta}_{\mathrm{visit}} = -0.2907$, $\hat{\beta}_{\mathrm{lage}} = 0.4657$ and $\hat{\beta}_{\mathrm{trt}*\mathrm{lbase}} = 0.3081$, and for $\log(\sigma)$ the intercept $\hat{\beta}_2 = -2.515$.

Breslow and Clayton (1993) considered including in the mean model random slopes in the covariate visit for subjects; however, this was not found to improve the model. Lee and Nelder

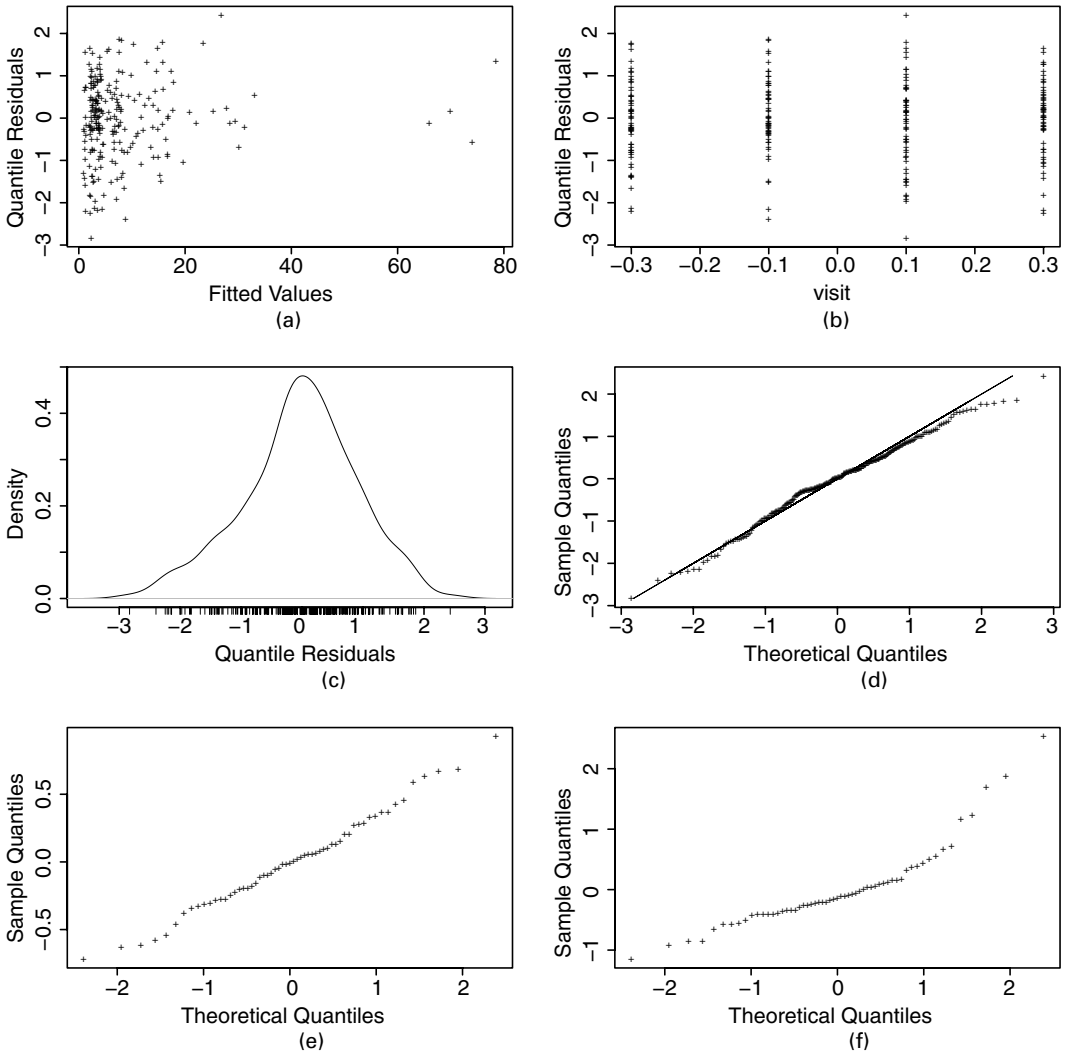

Fig. 10. Diagnostic plots for the epileptic seizures data: (a) residuals against fitted values, (b) residuals against visit, (c) residual kernel density estimate, (d) *QQ*-plot of the residuals, (e) *QQ*-plot of the random effects in $\log(\mu)$ and (f) *QQ*-plot of the random effects in $\log(\sigma)$

(1996, 2000) suggested that the casewise overdispersion may depend on an indicator variable for the fourth visit, denoted here by V4. In our model this is equivalent to replacing the dispersion model by $\log(\sigma) = \mathrm{V4}$. This model led to $\hat{\sigma}_1 = 0.333$ with $\mathrm{GAIC}(3) = 1268.5$.

Fig. 10 provides diagnostic plots for our model. Figs 10(a) and 10(b) plot the (normalized randomized quantile) residuals, which were defined in Section 6.2, against the fitted values and the covariate visit respectively and appear random. Figs 10(c) and 10(d) provide a kernel density estimate and normal QQ-plot for the residuals respectively and indicate some departure from the conditional negative binomial distribution for y. Figs 10(e) and 10(f) provide normal QQ-plots for the subject random effects in the log-mean and log-dispersion models respectively, indicating that the normal distribution for the random effects is adequate for $\log(\mu)$ but not for $\log(\sigma)$.

7.5. The river flow data

The river flow data, which were obtained from Tong (1990), comprise 1096 consecutive observations of the daily river flow r, of the river Vatnsdalsa in Iceland, measured in cubic metres per second, the daily precipitation p in millimetres and the mean daily temperature t in degrees centigrade at the meteorological station at Hveravellir in north-west Iceland. The data span the period of 1972, 1973 and 1974 and are shown in Fig. 11. The task is to build a stochastic model to predict the river flow by using the temperature and precipitation. Tong (1990) used a heavily parameterized self-exciting threshold autoregressive model with normal errors (conditional on current and past values of the explanatory variables p and t and past values of r). Here we investigate a variety of (conditional) distributions to model the river flow. We include the following explanatory variables, which were computed from r, p and t:

(a) lr, $\log(r)$;
(b) lp, $\log(p+1)$;
(c) lp90, the logarithm of the average precipitation for the last 90 days;
(d) tp, an indicator variable for positive t (i.e. $t>0$);
(e) t7, the average temperature over the last week (i.e. 7 days);
(f) t90, the average temperature for the last 90 days.

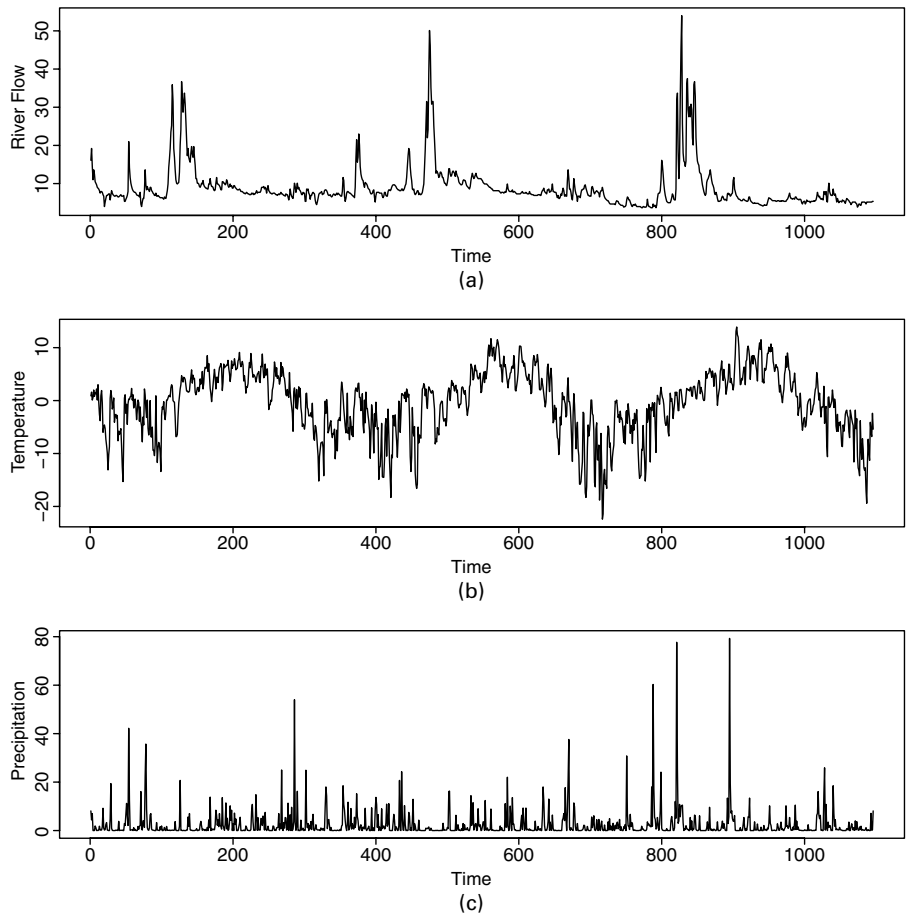

Fig. 11. River flow data: (a) riverflow, (b) temperature and (c) precipitation against time in days

Table 3. Models for river flow data

Distribution	*I*		*II*		*III*	
	GD	*SBC*	*GD*	*SBC*	*GD*	*SBC*
Gumbel	4903	4986	2346	2505	3503	3787
Reverse Gumbel	3343	3426	2060	2219	2581	2865
Normal	4077	4160	1982	2141	3041	3325
Gamma	2578	2661	1944	2103	2443	2726
Inverse Gaussian	2333	2416	1932	2091	2343	2626
Logistic	3267	3350	1872	2031	2426	2709
Box–Cox normal	2440	2529	1923	2089	2277	2568
t-family	2361	2451	1831	1997	2081	2371
Johnson–Su	2354	2451	1816	1988	2031	2238
Box–Cox *t*	2096	2192	1805	1978	1950	2247

Lag variables for river flow at lags 1, 2 and 3 (i.e. r1, r2 and r3 respectively), and at lag 1 for log-river-flow (lr1), precipitation (p1) and log-precipitation (lp1) were also used as explanatory variables. The first 90 observations and the last observation were weighted out from the analysis, leaving 1005 observations for model fitting.

Initially an inverse Gaussian distribution was assumed, owing to the skewed distribution of the river flow, with a constant dispersion model. After some initial search an adequate location model was found. Column I of Table 3 shows the global deviance GD and SBC for the resulting model I given by $\{\mu = \text{poly(r1, 2)} + \text{r2} + \text{r3} + (\text{tw} + p + \text{p1} * \text{t90}) * \text{tp} - (\text{tw} + p + \text{p1} * \text{t90})\}$, for a variety of distribution families, where any scale and shape parameters (e.g. σ, ν and τ) were modelled as constants. Note that poly(r1, 2) refers to a polynomial of order 2 in r1, i.e. a quadratic in r1. The conclusion, by looking at the SBC values, is that the Box–Cox *t*-distribution family fits best, indicating that the distribution of the river flow is both skew and leptokurtic.

Selecting now the Box–Cox *t*-distribution BCT, a search for an adequate dispersion model was made. Column II of Table 3 shows the resulting model II with μ as in model I and $\{\log(\sigma) = \text{poly(lr1, 3)} + \text{lp} + \text{lp1} + \text{lp90} + (t + \text{t90}) * \text{tp}\}$ fitted with different distribution families, again using constant shape parameters. BCT again fits best and the dispersion model has dramatically improved the fit, since SBC is reduced from 2192 to 1978. Constant models were found to be adequate for both the shape parameters of BCT.

For comparison, column III of Table 3 shows the Tong (1990) model, fitted with different distribution families. Tong (1990) used a normal distribution model with a heavily parameterized threshold mean model including many lags of r, t and p and a simple threshold dispersion model (using the optimum common threshold cut-off at $r = 13$). This model resulted in an SBC of 3325.

The final BCT model has μ and σ given by model II and ν and τ constant. From Section 4.2, the final fitted BCT model is given by $y = \hat{\mu}(\hat{\nu}\hat{\sigma}Z + 1)^{1/\hat{\nu}}$ where $Z \sim t_{\hat{\tau}}$ has a *t*-distribution with fitted degrees of freedom parameter $\hat{\tau} = 3.414$, shape parameter $\hat{\nu} = -0.6413$ and

$$\hat{\mu} = -0.019 + 1.219\,\text{r1} - 0.0043\,\text{r1}^2 - 0.243\,\text{r2} + 0.042\,\text{r3} + (0.152 - 0.0079\,\text{t7} + 0.0181\,p + 0.0640\,\text{p1} - 0.017\,\text{t90} - 0.0069\,\text{p1} * \text{t90}) \quad (\text{if } t > 0),$$

$$\log(\hat{\sigma}) = 6.850 - 13.845\,\text{lr1} + 5.679\,\text{lr1}^2 - 0.715\,\text{lr1}^3 + 0.169\,\text{lp} + 0.286\,\text{lp1} + 0.71029\,\text{lp90} - 0.0157t + 0.0821\,\text{t90} + (0.0764 - 0.0619t - 0.1268\,\text{t90}) \quad (\text{if } t > 0).$$

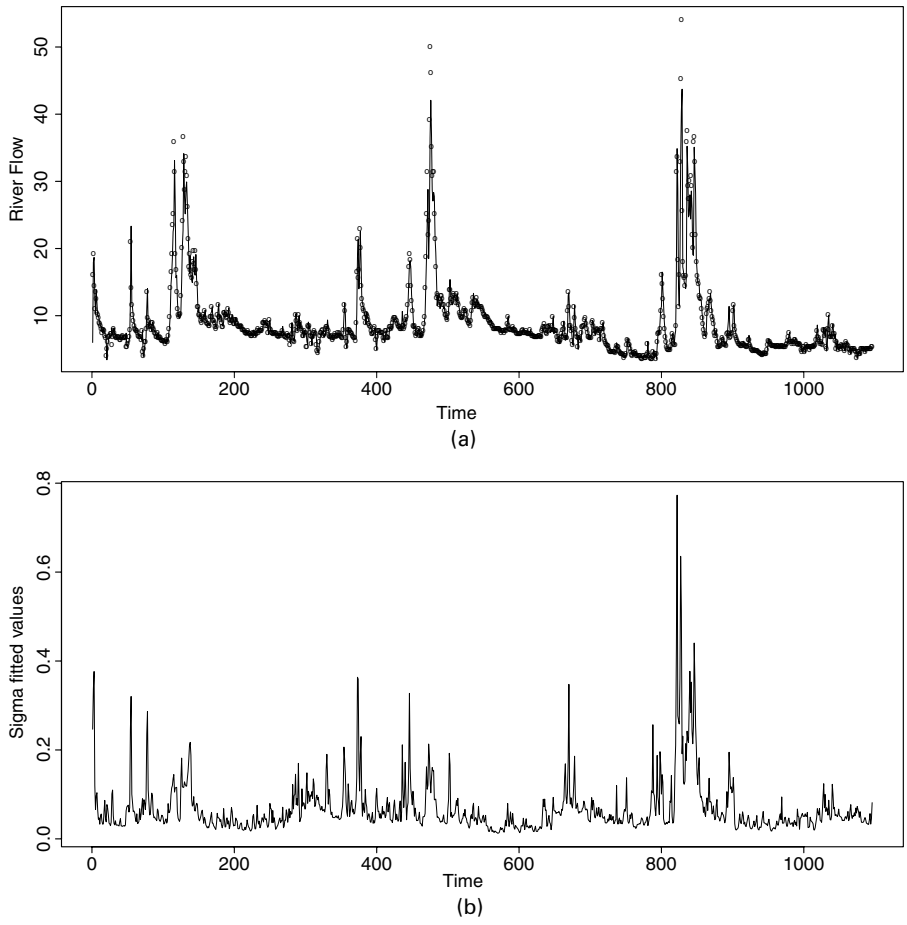

Fig. 12. River flow data: fitted values of (a) μ and (b) σ against time in days

Note that, for the BCT model, the location parameter μ is approximately the median of y. Fig. 12 displays the fitted values of μ and σ plotted against time in days. The (normalized quantile) residuals (see Section 6.2) for the final BCT model are shown in Fig. 13. Figs 13(a) and 13(b) plot their autocorrelation and partial autocorrelation functions respectively, whereas Figs 13(c) and 13(d) provide a kernel density estimate and QQ-plot for them respectively. The residuals appear satisfactory. In particular the assumptions of (conditional) independence and a BCT distribution for the river flow observations appear to be reasonable.

8. Conclusions

The GAMLSS is a very general class of models for a univariate response variable. It provides a common coherent framework for regression-type models, uniting models that are often considered as different in the statistical literature. It is therefore highly suited to educational objectives. It allows a very wide family of distributions for the response variable to be fitted, reducing the danger of distributional misspecification. It allows all the parameters of the distribution of the dependent variable to be modelled, so that location, scale, skewness and kurtosis parameters

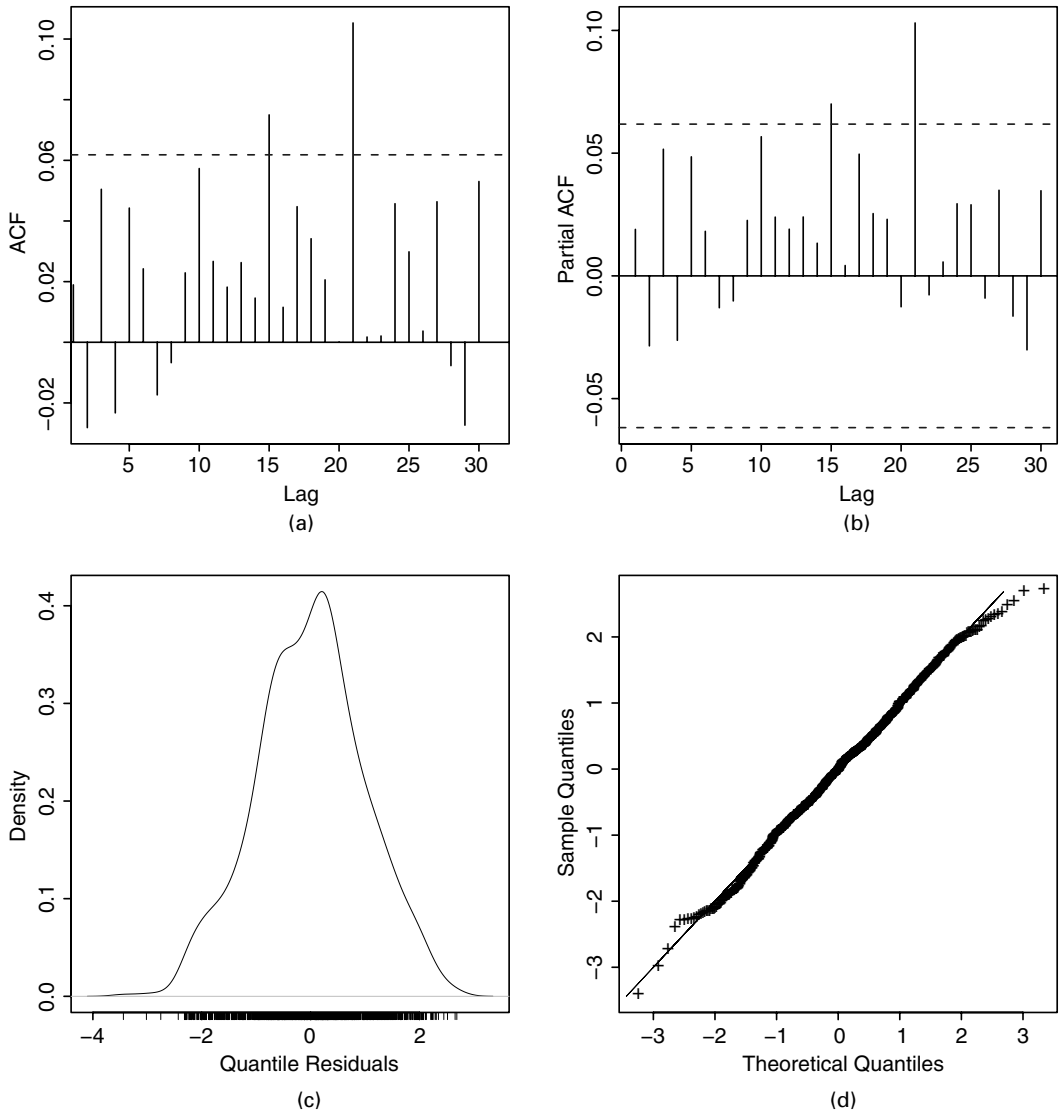

Fig. 13. River flow data residual plots (a) autocorrelation function ACF, (b) partial autocorrelation function, (c) kernel density estimate and (d) *QQ*-plot

can each be modelled explicitly if required. Different terms can be included in the predictor for each parameter, including splines and random effects, providing extra flexibility. For fixed hyperparameters the fitting algorithm of the GAMLSS model is very fast, so many alternative models can be fitted and explored before a final selection of a model or a combination of models is made. The hyperparameters can be estimated if required. The GAMLSS is implemented as a package (which is available free of charge from the authors) in the statistical environment R. The modular nature of the fitting algorithm allows additional alternative distributions and additive terms to be incorporated easily. The GAMLSS can also be used as an exploratory tool to select potential models for a subsequent fully Bayesian analysis.

Acknowledgements

The authors thank Calliope Akantziliotou for her help in the R implementation of the GAMLSS, Bob Gilchrist and Brian Francis for their comments and their encouragement during this work, Tim Cole for suggesting the body mass index data set, Jim Hodges, S. Gange and Howell Tong for providing the health maintenance organization, the hospital stay and river flow data sets respectively, the R Development Core Team for the package R (which is free of charge) and finally the four referees for comments that helped to improve the paper.

Appendix A: Inferential framework for the generalized additive model for location, scale and shape

A.1. Posterior mode estimation of the parameters β and random effects γ

For the GAMLSS (1) we use an empirical Bayesian argument, to obtain MAP, or posterior mode, estimation (see Berger (1985)) of both the $\boldsymbol{\beta}_k$s and the $\boldsymbol{\gamma}_{jk}$s assuming normal, possibly improper, priors for the $\boldsymbol{\gamma}_{jk}$s. We show below that this is equivalent to maximizing the penalized likelihood l_p, which is given by equation (5). To show this we shall use arguments that have been developed in the statistical literature by Wahba (1978), Silverman (1985), Green (1985), Kohn and Ansley (1988), Speed (1991), Green and Silverman (1994), Verbyla *et al.* (1999), Hastie and Tibshirani (2000) and Fahrmeir and Lang (2001).

The components of a GAMLSS (1) are

(a) $\mathbf{y}$, the response vector of length n,
(b) $\mathbf{X}=(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$, design matrices,
(c) $\boldsymbol{\beta}^{\mathrm{T}}=(\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_p^{\mathrm{T}})$, linear parameters,
(d) $\mathbf{Z}=(\mathbf{Z}_{11}, \mathbf{Z}_{21}, \ldots, \mathbf{Z}_{J_1 1}, \ldots, \mathbf{Z}_{1p}, \mathbf{Z}_{2p}, \ldots, \mathbf{Z}_{J_p p})$, design matrices,
(e) $\boldsymbol{\gamma}^{\mathrm{T}}=(\boldsymbol{\gamma}_{11}^{\mathrm{T}}, \boldsymbol{\gamma}_{21}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{J_1 1}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{1p}^{\mathrm{T}}, \boldsymbol{\gamma}_{2p}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{J_p p}^{\mathrm{T}})$, random effects, and
(f) $\boldsymbol{\lambda}^{\mathrm{T}}=(\boldsymbol{\lambda}_{11}^{\mathrm{T}}, \boldsymbol{\lambda}_{21}^{\mathrm{T}}, \ldots, \boldsymbol{\lambda}_{J_1 1}^{\mathrm{T}}, \ldots, \boldsymbol{\lambda}_{1p}^{\mathrm{T}}, \boldsymbol{\lambda}_{2p}^{\mathrm{T}}, \ldots, \boldsymbol{\lambda}_{J_p p}^{\mathrm{T}})$, hyperparameters.

Assume that the joint distribution of all the components in model (1) is given by

$$f(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})\ f(\boldsymbol{\gamma}|\boldsymbol{\lambda})\ f(\boldsymbol{\lambda})\ f(\boldsymbol{\beta}) \tag{9}$$

where $f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$ and $f(\boldsymbol{\gamma}|\boldsymbol{\lambda})$ are conditional distributions for $\mathbf{y}$ and $\boldsymbol{\gamma}$ and $f(\boldsymbol{\lambda})$ and $f(\boldsymbol{\beta})$ are appropriate priors for $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ respectively and $\mathbf{X}$ and $\mathbf{Z}$ are assumed fixed and known throughout. Assuming now that the hyperparameters $\boldsymbol{\lambda}$ are fixed and, assuming a constant improper prior for $\boldsymbol{\beta}$, then the posterior distribution for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ given $\mathbf{y}$ and $\boldsymbol{\lambda}$ is given by

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\lambda}) \propto f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})\ f(\boldsymbol{\gamma}|\boldsymbol{\lambda}). \tag{10}$$

Model (1) assumes conditionally independent y_i for $i = 1, 2, \ldots, n$ given $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and assumes that the $\boldsymbol{\gamma}_{jk}$s have independent normal, possibly improper, prior distributions, $\boldsymbol{\gamma}_{jk} \sim N(0, \mathbf{G}_{jk}^{-})$. Hence, from expression (10),

$$\log\{f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{y}, \boldsymbol{\lambda})\} = l_p + c(\mathbf{y}, \boldsymbol{\lambda})$$

where l_p is given in equation (5) and $l = \log\{f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})\} = \Sigma_{i=1}^{n} \log\{f(y_i|\boldsymbol{\theta}^i)\}$, and $c(\mathbf{y}, \boldsymbol{\lambda})$ is a function of $\mathbf{y}$ and $\boldsymbol{\lambda}$. Note that, for a GAMLSS, l_p is equivalent, with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, to the h-likelihood of Lee and Nelder (1996, 2001a, b).

Hence l_p is maximized over $(\boldsymbol{\beta},\boldsymbol{\gamma})$, giving posterior mode (or MAP) estimation of $(\boldsymbol{\beta},\boldsymbol{\gamma})$ and, for fixed hyperparameters $\boldsymbol{\lambda}$, MAP estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is equivalent to maximizing the penalized likelihood l_p that is given by equation (5).

The details of the RS and CG algorithms for maximizing the penalized likelihood l_p, over both the parameters $\boldsymbol{\beta}$ and the random-effects terms $\boldsymbol{\gamma}$ (for fixed hyperparameters $\boldsymbol{\lambda}$), are given in Appendix B. The justification of the CG algorithm is given in Appendix C.

A.2. Hyperparameter estimation

The hyperparameters $\boldsymbol{\lambda}$ can be estimated within a classical likelihood framework for random effects by maximizing the marginal likelihood for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ given $\mathbf{y}$, i.e.

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})\ f(\boldsymbol{\gamma}|\boldsymbol{\lambda})\ \mathrm{d}\boldsymbol{\gamma}.$$

The maximization of $L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y})$ over $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ involves high dimensional integration so any approach to maximizing it will be computer intensive. Note that the maximum likelihood estimator for $\boldsymbol{\beta}$ from this approach will not in general be the same as the MAP estimator for $\boldsymbol{\beta}$ that was described in the previous section.

In restricted maximum likelihood (REML) estimation, effectively a non-informative (constant) prior is assumed for $\boldsymbol{\beta}$ and both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are integrated out of the joint density $f(\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\beta}|\boldsymbol{\lambda})$ to give the marginal likelihood $L(\boldsymbol{\lambda}|\mathbf{y})$, which is maximized over $\boldsymbol{\lambda}$.

In a fully Bayesian inference for the GAMLSS, the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ is obtained from equation (9), e.g. by using Markov chain Monte Carlo sampling; see Fahrmeir and Tutz (2001) or Fahrmeir and Lang (2001).

The above methods of estimation of the hyperparameters $\boldsymbol{\lambda}$ are in general highly computationally intensive: the maximum likelihood and REML methods require high dimensional integration, whereas the fully Bayes method requires Markov chain Monte Carlo sampling.

The following four methods, which do not require such computational intensity, are considered for hyperparameter estimation in GAMLSSs.

The methods are summarized in the following algorithm.

(a) *Procedure 1*: estimate the hyperparameters $\boldsymbol{\lambda}$ by one of the methods
  (i) minimizing a profile generalized Akaike information criterion GAIC over $\boldsymbol{\lambda}$,
  (ii) minimizing a profile generalized cross-validation criterion over $\boldsymbol{\lambda}$,
  (iii) maximizing the approximate marginal density (or profile marginal likelihood) for $\boldsymbol{\lambda}$ by using a Laplace approximation or
  (iv) approximately maximizing the marginal likelihood for $\boldsymbol{\lambda}$ by using an (approximate) EM algorithm.

(b) *Procedure 2*: for fixed current hyperparameters $\boldsymbol{\lambda}$, use the GAMLSS (RS or CG) algorithm to obtain posterior mode (MAP) estimates of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

Procedure 2 is nested within procedure 1 and a numerical algorithm is used to estimate $\boldsymbol{\lambda}$.

We now consider the methods in more detail.

A.2.1. Minimizing a profile generalized Akaike information criterion over $\boldsymbol{\lambda}$

GAIC (Akaike, 1983) was considered by Hastie and Tibshirani (1990), pages 160 and 261, for hyperparameter estimation in GAMs. In GAMs a cubic smoothing spline function $h(x)$ is used to model the dependence of a predictor on explanatory variable x. For a single smoothing spline term, since λ is related to the smoothing degrees of freedom $\mathrm{df} = \mathrm{tr}(\mathbf{S})$ through equation (6), selection (or estimation) of λ may be achieved by minimizing GAIC(#), which is defined in Section 6.2, over λ.

When the model contains p cubic smoothing spline functions in different explanatory variables, then the corresponding p smoothing hyperparameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ can be jointly estimated by minimizing GAIC(#) over $\boldsymbol{\lambda}$. However, with multiple smoothing splines $\Sigma_{j=1}^{p} \mathrm{tr}(\mathbf{S}_j)$ is only an approximation to the full model complexity degrees of freedom.

The GAIC(#) criterion can be applied more generally to estimate hyperparameters $\boldsymbol{\lambda}$ in the distribution of random-effects terms. The (model complexity) degrees of freedom df need to be obtained for models with random-effects terms. This has been considered by Hodges and Sargent (2001). The degrees of freedom of a model with a single random-effects term can be defined as the trace of the random-effect (shrinkage) smoother $\mathbf{S}$, i.e. $\mathrm{df} = \mathrm{tr}(\mathbf{S})$, where $\mathbf{S}$ is given by equation (6). As with smoothing terms, when there are other terms in the model $\Sigma_{j=1}^{p} \mathrm{tr}(\mathbf{S}_j)$ is only an approximation to the full model complexity degrees of freedom. The full model complexity degrees of freedom for model (1) are given by $\mathrm{df} = \mathrm{tr}(\mathbf{A}^{-1}\mathbf{B})$ where $\mathbf{A}$ is defined in Appendix C and $\mathbf{B}$ is obtained from $\mathbf{A}$ by omitting the matrices $\mathbf{G}_{jk}$ for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$.

A.2.2. Minimizing a generalized cross-validation over $\boldsymbol{\lambda}$

The generalized cross-validation criterion was considered by Hastie and Tibshirani (1990), pages 259–263, for hyperparameter estimation in GAMs. The criterion GAIC in Appendix A.2.1 is replaced by the generalized cross-validation criterion, which is minimized over $\boldsymbol{\lambda}$. Verbyla *et al.* (1999) considered the

approximate equivalence of generalized cross-validation and REML methods of estimating $\boldsymbol{\lambda}$ in smoothing splines models, which was considered in more detail by Wahba (1985) and Kohn *et al.* (1991).

A.2.3. Maximizing the approximate marginal density (or profile marginal likelihood) of $\boldsymbol{\lambda}$ *by using a Laplace approximation*

For GLMMs, Breslow and Clayton (1993) used a first-order Laplace integral approximation to integrate out the random effects $\boldsymbol{\gamma}$ and to approximate the marginal likelihood, leading to estimating equations based on penalized quasi-likelihood for the mean model parameters and pseudonormal (REML) likelihood for the dispersion components. Breslow and Lin (1995) extended this to a second-order Laplace approximation.

Lee and Nelder (1996) took a similar approach, estimating the dispersion components by using a first-order approximation to the Cox and Read (1987) profile likelihood which eliminates the nuisance parameters $\boldsymbol{\beta}$ from the marginal likelihood, which they called an adjusted profile h-likelihood. Lee and Nelder (2001a) extended this to a second-order approximation.

Here we consider an approximate Bayesian approach. Assuming a uniform improper prior for both $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ then from equation (9) the posterior marginal of $\boldsymbol{\lambda}$ is given by

$$f(\boldsymbol{\lambda}|\mathbf{y}) = \int\int \frac{\exp(l_h)}{f(\mathbf{y})}\, \mathrm{d}\boldsymbol{\gamma}\, \mathrm{d}\boldsymbol{\beta} \tag{11}$$

where

$$l_h = l_h(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \log\{f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})\} + \log\{f(\boldsymbol{\gamma}|\boldsymbol{\lambda})\} = l_p + \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\{\log|\mathbf{G}_{jk}| - q_{jk}\log(2\pi)\}$$

was first defined in Section 2.2 and where l_p is given by equation (5). Using a first-order Laplace approximation (Tierney and Kadane, 1986) to the integral (11) gives

$$f(\boldsymbol{\lambda}|\mathbf{y}) \approx \frac{\exp(\hat{l}_h)}{f(\mathbf{y})}\left|\frac{\hat{\mathbf{D}}}{2\pi}\right|^{-1/2} \tag{12}$$

where $\hat{l}_h = l_h(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ and

$$\hat{\mathbf{D}} = \mathbf{D}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = -\begin{pmatrix} \dfrac{\partial^2 l_h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}} & \dfrac{\partial^2 l_h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^{\mathrm{T}}} \\ \dfrac{\partial^2 l_h}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}^{\mathrm{T}}} & \dfrac{\partial^2 l_h}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}} \end{pmatrix}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\, \boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} \tag{13}$$

is the observed information matrix, evaluated at $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ and $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})$, the MAP estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ given each fixed $\boldsymbol{\lambda}$. (Note that matrix $\mathbf{D}$ is a rearrangement of matrix $\mathbf{A}$ from Appendix C.) Estimation of $\boldsymbol{\lambda}$ can be achieved by maximizing approximation (12) over $\boldsymbol{\lambda}$ (e.g. by using a numerical maximization algorithm). Alternatively, this can be considered as a generalization of REML estimation of $\boldsymbol{\lambda}$, maximizing an approximate profile log-likelihood for $\boldsymbol{\lambda}$, denoted here as $l(\boldsymbol{\lambda})$, given by replacing $D(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ by the expected information $\hat{\mathbf{H}} = \mathbf{H}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, giving

$$l(\boldsymbol{\lambda}) = \hat{l}_h - \tfrac{1}{2}\log|\hat{\mathbf{H}}/2\pi|. \tag{14}$$

This is closely related to the adjusted profile h-likelihood of Lee and Nelder (1996, 2001a, b).

A.2.4. Approximately maximizing the marginal likelihood for $\boldsymbol{\lambda}$ *by using an (approximate) EM algorithm*

An approximate EM algorithm was used by Fahrmeir and Tutz (2001), pages 298–303, and by Diggle *et al.* (2002), pages 172–175, to estimate hyperparameters in GLMMs and is similarly applied here to maximize approximately over $\boldsymbol{\lambda}$ the marginal likelihood of $\boldsymbol{\lambda}$, $L(\boldsymbol{\lambda})$ (or equivalently the posterior marginal distribution of $\boldsymbol{\lambda}$ for a non-informative uniform prior).

In the E-step of the EM algorithm, $M(\boldsymbol{\lambda}|\hat{\boldsymbol{\lambda}}) = E[\log\{f(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\lambda})\}]$, is approximated, where the expectation is over the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ given $\mathbf{y}$ and $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$, i.e. $f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{y}, \hat{\boldsymbol{\lambda}})$, where $\hat{\boldsymbol{\lambda}}$ is the current estimate of $\boldsymbol{\lambda}$, giving, apart from a function of $\mathbf{y}$,

$$M(\boldsymbol{\lambda}|\hat{\boldsymbol{\lambda}}) = -\frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}(\mathrm{tr}[\mathbf{G}_{jk}\{\hat{\gamma}_{jk}\hat{\gamma}_{jk}^{\mathrm{T}} + \hat{V}(\hat{\gamma}_{jk})\}] - \log|\mathbf{G}_{jk}|) \tag{15}$$

where $\hat{\gamma}_{jk}$ and $\hat{V}(\hat{\gamma}_{jk})$ are the posterior mode and curvature (i.e. submatix of $\mathbf{A}^{-1}$) of $\boldsymbol{\gamma}_{jk}$ from the MAP estimation in Appendix C.

In the M-step of the EM algorithm, $M(\boldsymbol{\lambda}|\hat{\boldsymbol{\lambda}})$ is maximized over $\boldsymbol{\lambda}$ by a numerical maximization algorithm (e.g. the function `optim` in the R package). If $\mathbf{G}_{jk} = \mathbf{G}_k$ for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$, and the $\mathbf{G}_k$ are unconstrained positive definite symmetric matrices (e.g. in a random-coefficients model), then equation (15) can be maximized explicitly giving, for $k = 1, 2, \ldots, p$,

$$\hat{\mathbf{G}}_k^{-1} = \frac{1}{J_k}\sum_{j=1}^{J_k}\{\hat{\gamma}_{jk}\hat{\gamma}_{jk}^{\mathrm{T}} + \hat{V}(\hat{\gamma}_{jk})\}. \tag{16}$$

Appendix B: The algorithms

B.1. Introduction

Let $\mathbf{u}_k = \partial l/\partial \boldsymbol{\eta}_k$ be the score functions, $\mathbf{z}_k = \boldsymbol{\eta}_k + \mathbf{W}_{kk}^{-1}\mathbf{u}_k$ be the adjusted dependent variables and $\mathbf{W}_{ks}$ be diagonal matrices of iterative weights, for $k = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, p$, which can have one of the forms

$$-\frac{\partial^2 l}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}_s^{\mathrm{T}}},$$

$$-E\left(\frac{\partial^2 l}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}_s^{\mathrm{T}}}\right)$$

or

$$\mathrm{diag}\left\{\left(\frac{\partial l_i}{\partial \eta_{ik}}\frac{\partial l_i}{\partial \eta_{is}}\right)\right\},$$

over $i = 1, 2, \ldots, n$, i.e. the observed information, expected information or product score function, depending respectively on whether a Newton–Raphson, Fisher scoring or quasi-Newton–Raphson algorithm is used (see Lange (1999), chapter 11, for a definition of the techniques), in the RS and CG algorithms below.

Let r be the outer cycle iteration index, k the parameter index, i the inner cycle iteration index, m the backfitting index and j the random-effects (or nonparametric) term index. Also, for example, let $\boldsymbol{\gamma}_{jk}^{(r,i,m)}$ denote the current value of the vector $\boldsymbol{\gamma}_{jk}$ in the rth outer, ith inner and mth backfitting cycle iteration and let $\boldsymbol{\gamma}_{jk}^{(r,i,.)}$ denote the value of $\boldsymbol{\gamma}_{jk}$ at the convergence of the backfitting cycle for the ith inner cycle of the rth outer cycle, which is also the starting value $\boldsymbol{\gamma}_{jk}^{(r,i+1,1)}$ for the $(i+1)$th inner cycle of the rth outer cycle, for $j = 1, 2, \ldots, J_k$ and $k = 1, \ldots, p$. Note also, for example, that $\boldsymbol{\gamma}_{jk}^{(r,i,c)}$ means the current (i.e. most recently) updated estimate of $\boldsymbol{\gamma}_{jk}$ and the algorithm operates in the backfitting cycle of the ith inner cycle of the rth outer cycle.

B.2. The RS algorithm

Essentially the RS algorithm has an outer cycle which maximizes the penalized likelihood with respect to $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_{jk}$, for $j = 1, \ldots, J_k$, in the model successively for each θ_k in turn, for $k = 1, \ldots, p$. At each calculation in the algorithm the current updated values of all the quantities are used.

The RS algorithm is not a special case of the CG algorithm because in the RS algorithm the diagonal weight matrix $\mathbf{W}_{kk}$ is evaluated (i.e. updated) *within* the fitting of each parameter $\boldsymbol{\theta}_k$, whereas in the CG algorithm all weight matrices $\mathbf{W}_{ks}$ for $k = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, p$ are evaluated *after* fitting *all* $\boldsymbol{\theta}_k$ for $k = 1, 2, \ldots, p$.

The RS algorithm is as follows.

Step 1: start—initialize fitted values $\boldsymbol{\theta}_k^{(1,1)}$ and random effects $\boldsymbol{\gamma}_{jk}^{(1,1,1)}$, for $j = 1, \ldots, J_k$ and $k = 1, 2, \ldots, p$. Evaluate the initial linear predictors $\boldsymbol{\eta}_k^{(1,1)} = g_k(\boldsymbol{\theta}_k^{(1,1)})$, for $k = 1, 2, \ldots, p$.

Step 2: start the outer cycle $r = 1, 2, \ldots$ until convergence. For $k = 1, 2, \ldots, p$:

(a) start the inner cycle $i = 1, 2, \ldots$ until convergence—
  (i) evaluate the current $\mathbf{u}_k^{(r,i)}$, $\mathbf{W}_{kk}^{(r,i)}$ and $\mathbf{z}_k^{(r,i)}$;
  (ii) start the backfitting cycle $m = 1, 2, \ldots$ until convergence;
  (iii) regress the current partial residuals $\varepsilon_{0k}^{(r,i,m)} = \mathbf{z}_k^{(r,i)} - \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}^{(r,i,m)}$ against design matrix $\mathbf{X}_k$, using the iterative weights $\mathbf{W}_{kk}^{(r,i)}$ to obtain the updated parameter estimates $\boldsymbol{\beta}_k^{(r,i,m+1)}$;
  (iv) for $j = 1, 2, \ldots, J_k$ smooth the partial residuals $\varepsilon_{jk}^{(r,i,m)} = \mathbf{z}_k^{(r,i)} - \mathbf{X}_k\boldsymbol{\beta}_k^{(r,i,m+1)} - \sum_{t=1,\, t\neq j}^{J_k} \mathbf{Z}_{tk}\boldsymbol{\gamma}_{tk}^{(r,i,c)}$, using the shrinking (smoothing) matrix S_{jk} given by equation (6) to obtain the updated (and current) additive predictor term $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}^{(r,i,m+1)}$;
  (v) end the backfitting cycle, on convergence of $\beta_k^{(r,i,.)}$and $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}^{(r,i,.)}$ and set $\boldsymbol{\beta}_k^{(r,i+1)} = \boldsymbol{\beta}_k^{(r,i,.)}$ and $\boldsymbol{\gamma}_{jk}^{(r,i+1)} = \boldsymbol{\gamma}_{jk}^{(r,i,.)}$ for $j = 1, 2, \ldots, J_k$ and otherwise update m and continue the backfitting cycle;
  (vi) calculate the updated $\boldsymbol{\eta}_k^{(r,i+1)}$ and $\boldsymbol{\theta}_k^{(r,i+1)}$;
(b) end the inner cycle on convergence of $\boldsymbol{\beta}_k^{(r,.)}$ and the additive predictor terms $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}^{(r,.)}$ and set $\boldsymbol{\beta}_k^{(r+1,1)} = \boldsymbol{\beta}_k^{(r,.)}$, $\boldsymbol{\gamma}_{jk}^{(r+1,1)} = \boldsymbol{\gamma}_{jk}^{(r,.)}$, for $j = 1, 2, \ldots, J_k$, $\boldsymbol{\eta}_k^{(r+1,1)} = \boldsymbol{\eta}_k^{(r,.)}$ and $\boldsymbol{\theta}_k^{(r+1,1)} = \boldsymbol{\theta}_k^{(r,.)}$; otherwise update i and continue the inner cycle.

Step 3: update the value of k.
Step 4: end the outer cycle—if the change in the (penalized) likelihood is sufficiently small; otherwise update r and continue the outer cycle.

B.3. The CG algorithm

Algorithm CG, based on Cole and Green (1992) is as follows.

Step 1: start—initialize $\boldsymbol{\theta}_k^{(1,1)}$ and $\boldsymbol{\gamma}_{jk}^{(1,1,1)}$ for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$. Evaluate $\boldsymbol{\eta}_k^{(1)} = \boldsymbol{\eta}_k^{(1,1)} = g_k(\boldsymbol{\theta}_k^{(1,1)})$ for $k = 1, 2, \ldots, p$.
Step 2: start the outer cycle $r = 1, 2, \ldots$ until convergence.
Step 3: evaluate and *fix* the current $\mathbf{u}_k^{(r)}$, $\mathbf{W}_{ks}^{(r)}$ and $\mathbf{z}_k^{(r)}$ for $k = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, p$. Perform a single rth step of the Newton–Raphson algorithm by

(a) starting the inner cycle $i = 1, 2, \ldots$ until convergence—for $k = 1, 2, \ldots, p$,
  (i) start the backfitting cycle $m = 1, 2, \ldots$ until convergence

$$X_k\boldsymbol{\beta}_k^{(r,i,m+1)} = \mathbf{H}_k^{(r)}\boldsymbol{\varepsilon}_{0k}^{(r,i,m)},$$

  and for $j = 1, 2, \ldots, J_k$

$$\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}^{(r,i,m+1)} = \mathbf{S}_{jk}^{(r)}\boldsymbol{\varepsilon}_{jk}^{(r,i,m)},$$

  (ii) end the backfitting cycle, on convergence of $\beta_k^{(r,i,.)}$ and $z_{jk}\boldsymbol{\gamma}_{jk}^{(r,i,.)}$ and set $\boldsymbol{\beta}_k^{(r,i+1)} = \boldsymbol{\beta}_k^{(r,i,.)}$ and $\boldsymbol{\gamma}_{jk}^{(r,i+1)} = \boldsymbol{\gamma}_{jk}^{(r,i,.)}$ for $j = 1, 2, \ldots, J_k$ and otherwise update m and continue the backfitting cycle, and
  (iii) calculate the updated $\boldsymbol{\eta}_k^{(r,i+1)}$ and $\boldsymbol{\theta}_k^{(r,i+1)}$ and then update k;
(b) end the inner cycle on convergence of $\boldsymbol{\beta}_k^{(r,.)}$ and the additive predictor terms $\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}^{(r,.)}$ and set $\boldsymbol{\beta}_k^{(r+1,1)} = \boldsymbol{\beta}_k^{(r,.)}$, $\boldsymbol{\gamma}_{jk}^{(r+1,1)} = \boldsymbol{\gamma}_{jk}^{(r,.)}$, $\boldsymbol{\eta}_k^{(r+1)} = \boldsymbol{\eta}_k^{(r+1,1)} = \boldsymbol{\eta}_k^{(r,.)}$ and $\boldsymbol{\theta}_k^{(r+1,1)} = \boldsymbol{\theta}_k^{(r,.)}$, for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$; otherwise update i and continue the inner cycle.

Step 4: end the outer cycle if the change in the (penalized) likelihood is sufficiently small; otherwise update r and continue the outer cycle.

The matrices $\mathbf{H}_k^{(r)}$ and $\mathbf{S}_{jk}^{(r)}$, which are defined in Appendix C, are the projection matrices and the shrinking matrices, for the parametric and additive components of the model respectively, at the rth iteration, for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$.

The partial residuals $\varepsilon_{0k}^{(r,i,m)}$ and $\boldsymbol{\varepsilon}_{jk}^{(r,i,m)}$ are the current working variables for fitting the parametric and the additive (random-effects or smoothing) components of the model respectively and are defined as

$$\varepsilon_{0k}^{(r,i,m)} = \mathbf{z}_k^{(r)} - \sum_{t=1}^{J_k} \mathbf{Z}_{tk}\boldsymbol{\gamma}_{tk}^{(r,i,c)} - \mathbf{W}_{kk}^{(r)\,-1} \sum_{s=1,s\neq k}^{p} \mathbf{W}_{ks}^{(r)}(\boldsymbol{\eta}_s^{(r,c)} - \boldsymbol{\eta}_s^{(r)}),$$

$$\boldsymbol{\varepsilon}_{jk}^{(r,i,m)} = \mathbf{z}_k^{(r)} - \mathbf{X}_k\boldsymbol{\beta}_k^{(r,i,m+1)} - \sum_{t=1,t\neq j}^{J_k} \mathbf{Z}_{tk}\boldsymbol{\gamma}_{tk}^{(r,i,c)} - \mathbf{W}_{kk}^{(r)\,-1} \sum_{s=1,s\neq k}^{p} \mathbf{W}_{ks}^{(r)}(\boldsymbol{\eta}_s^{(r,c)} - \boldsymbol{\eta}_s^{(r)}).$$

The full Newton–Raphson step length in the algorithm can be replaced by a step of size α, by updating the linear predictors as

$$\boldsymbol{\eta}_k^{(r+1)}(\alpha) = \alpha\boldsymbol{\eta}_k^{(r+1)} + (1-\alpha)\boldsymbol{\eta}_k^{(r)}$$

rather than $\boldsymbol{\eta}_k^{(r+1)}$ for $k = 1, 2, \ldots, p$, at the end of the inner cycle for the rth outer cycle and then evaluating $\mathbf{u}_k^{(r+1)}$, $\mathbf{W}_{ks}^{(r+1)}$ and $\mathbf{z}_k^{(r+1)}$, for $k = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, p$, using the $\boldsymbol{\eta}_k^{(r+1)}(\alpha)$ for $k = 1, 2, \ldots, p$. The optimum step length for a particular iteration r can be obtained by maximizing $l_p(\alpha)$ over α.

The inner (backfitting) cycle of the algorithm can be shown to converge (for cubic smoothing splines and similar linear smoothers); Hastie and Tibshirani (1990), chapter 5. The outer cycle is simply a Newton–Raphson algorithm. Thus, if step size optimization is performed, the outer loop will converge as well. Standard general results on the Newton–Raphson algorithm ensure convergence (Ortega and Rheinboldt, 1970). Step optimization is rarely needed in practice in our experience.

Appendix C: Maximization of the penalized likelihood

In this appendix it is shown that maximization of the penalized log-likelihood function l_p that is given by equation (5) over the parameters $\boldsymbol{\beta}_k$ and terms $\boldsymbol{\gamma}_{jk}$ for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$ leads to the algorithm that is described in Appendix B.

This is achieved by the following two steps.

(a) The first and second derivatives of equation (5) are obtained to give a Newton–Raphson step for maximizing equation (5) with respect to $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_{jk}$ for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$.
(b) Each step of the Newton–Raphson algorithm is achieved by using a backfitting procedure cycling through the parameters and through the additive terms of the k linear predictors.

C.1. Step (a)

The algorithm maximizes the penalized likelihood function l_p, given by equation (5), using a Newton–Raphson algorithm. The first derivative (score function) and the second derivatives of l_p with respect to $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_{jk}$ for all $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$ are evaluated at iteration r at the current predictors $\boldsymbol{\eta}_k^{(r)}$ for $k = 1, 2, \ldots, p$.

Let $\boldsymbol{\alpha}_k^{\mathrm{T}} = (\boldsymbol{\beta}_k^{\mathrm{T}}, \boldsymbol{\gamma}_{1k}^{\mathrm{T}}, \boldsymbol{\gamma}_{2k}^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{J_k k}^{\mathrm{T}})$, $\mathbf{a}_k = \partial l_p/\partial\boldsymbol{\alpha}_k$ and $\mathbf{A}_{ks} = -\partial^2 l_p/\partial\boldsymbol{\alpha}_k\,\partial\boldsymbol{\alpha}_s^{\mathrm{T}}$ for $k = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, p$, and let $\boldsymbol{\alpha}^{\mathrm{T}} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \boldsymbol{\alpha}_2^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_p^{\mathrm{T}})$, $\mathbf{a} = \partial l_p/\partial\boldsymbol{\alpha}$ and $\mathbf{A} = -\partial^2 l_p/\partial\boldsymbol{\alpha}\,\partial\boldsymbol{\alpha}^{\mathrm{T}}$.

The Newton–Raphson step is given by $\mathbf{A}^{(r)}(\boldsymbol{\alpha}^{(r+1)} - \boldsymbol{\alpha}^{(r)}) = \mathbf{a}^{(r)}$, i.e.

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1p} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2p} \\ \vdots & \cdots & \cdots & \vdots \\ \mathbf{A}_{p1} & \mathbf{A}_{p2} & \cdots & \mathbf{A}_{pp} \end{pmatrix}^{(r)} \begin{pmatrix} \boldsymbol{\alpha}_1^{(r+1)} - \boldsymbol{\alpha}_1^{(r)} \\ \boldsymbol{\alpha}_2^{(r+1)} - \boldsymbol{\alpha}_2^{(r)} \\ \vdots \\ \boldsymbol{\alpha}_p^{(r+1)} - \boldsymbol{\alpha}_p^{(r)} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_p \end{pmatrix}^{(r)}$$

where the matrix $\mathbf{A}_{ks}$ is given by

$$\begin{pmatrix} \mathbf{X}_k^{\mathrm{T}}\mathbf{W}_{ks}\mathbf{X}_s & \mathbf{X}_k^{\mathrm{T}}\mathbf{W}_{ks}\mathbf{Z}_{1s} & \cdots & \mathbf{X}_k^{\mathrm{T}}\mathbf{W}_{ks}\mathbf{Z}_{J_s s} \\ \mathbf{Z}_{1k}^{\mathrm{T}}\mathbf{W}_{ks}\mathbf{X}_s & \mathbf{Z}_{1k}^{\mathrm{T}}\mathbf{W}_{ks}\mathbf{Z}_{1s} + \mathbf{G}_{1k}\ (\text{if } s = k) & \cdots & \mathbf{Z}_{1k}^{\mathrm{T}}\mathbf{W}_{ks}\mathbf{Z}_{J_s s} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{Z}_{J_k k}^{\mathrm{T}}\mathbf{W}_{ks}\mathbf{X}_s & \mathbf{Z}_{J_k k}^{\mathrm{T}}\mathbf{W}_{ks}\mathbf{Z}_{1s} & \cdots & \mathbf{Z}_{J_k k}^{\mathrm{T}}\mathbf{W}_{ks}\mathbf{Z}_{J_s s} + \mathbf{G}_{J_k k}\ (\text{if } s = k) \end{pmatrix}$$

and the vector

$$\mathbf{a}_k^{(r)} = \begin{pmatrix} \mathbf{X}_k^{\mathrm{T}}\mathbf{u}_k^{(r)} \\ \mathbf{Z}_{1k}^{\mathrm{T}}\mathbf{u}_k^{(r)} - \mathbf{G}_{1k}\boldsymbol{\gamma}_{1k}^{(r)} \\ \vdots \\ \mathbf{Z}_{J_k k}^{\mathrm{T}}\mathbf{u}_k^{(r)} - \mathbf{G}_{J_k k}\boldsymbol{\gamma}_{J_k k}^{(r)} \end{pmatrix}$$

where $\mathbf{u}_k = \partial l/\partial\boldsymbol{\eta}_k$ and $\mathbf{W}_{ks} = -\partial^2 l/\partial\boldsymbol{\eta}_k\,\partial\boldsymbol{\eta}_s^{\mathrm{T}} = -\mathrm{diag}\{\partial^2 l_i/\partial\eta_{ik}\,\partial\eta_{is}\}$ over $i = 1, 2, \ldots, n$, for $k = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, p$ (see Appendix B for alternative weight matrices).

C.2. Step (b)

Now considering the row corresponding to updating $\boldsymbol{\gamma}_{jk}$ gives

$$\mathbf{G}_{jk}(\boldsymbol{\gamma}_{jk}^{(r+1)} - \boldsymbol{\gamma}_{jk}^{(r)}) + \mathbf{Z}_{jk}^{\mathrm{T}} \sum_{s=1}^{p} \mathbf{W}_{ks}^{(r)}(\boldsymbol{\eta}_s^{(r+1)} - \boldsymbol{\eta}_s^{(r)}) = \mathbf{Z}_{jk}^{\mathrm{T}} \mathbf{u}_k^{(r)} - \mathbf{G}_{jk}\boldsymbol{\gamma}_{jk}^{(r)}.$$

Expanding and rearranging this gives

$$\mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}^{(r+1)} = \mathbf{S}_{jk}^{(r)}\boldsymbol{\varepsilon}_{jk}^{(r)} \tag{17}$$

where, for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, p$,

$$\mathbf{S}_{jk}^{(r)} = \mathbf{Z}_{jk}(\mathbf{Z}_{jk}^{\mathrm{T}}\mathbf{W}_{kk}^{(r)}\mathbf{Z}_{jk} + \mathbf{G}_{jk})^{-1}\mathbf{Z}_{jk}^{\mathrm{T}}\mathbf{W}_{kk}^{(r)}$$

is a shrinking (smoothing) matrix and where

$$\boldsymbol{\varepsilon}_{jk}^{(r)} = \mathbf{z}_k^{(r)} - \mathbf{X}_k\boldsymbol{\beta}_k^{(r+1)} - \sum_{t=1, t\neq j}^{J_k} \mathbf{Z}_{tk}\boldsymbol{\gamma}_{tk}^{(r+1)} - \mathbf{W}_{kk}^{(r)^{-1}} \sum_{s=1, s\neq k}^{p} \mathbf{W}_{ks}^{(r)}(\boldsymbol{\eta}_s^{(r+1)} - \boldsymbol{\eta}_s^{(r)})$$

are the partial residuals and $\mathbf{z}_k^{(r)} = \boldsymbol{\eta}_k^{(r)} + \mathbf{W}_{kk}^{(r)^{-1}}\mathbf{u}_k^{(r)}$ is the adjusted dependent variable.

(A device for obtaining updated estimate $\boldsymbol{\gamma}_{jk}^{(r+1)}$ in equation (17) is to apply weighted least squares estimation to an augmented data model given by

$$\begin{pmatrix} \boldsymbol{\varepsilon}_{jk}^{(r)} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_{jk} \\ -\mathbf{D}_{jk} \end{pmatrix} \boldsymbol{\gamma}_{jk} + \begin{pmatrix} \mathbf{e}_{0k} \\ \mathbf{e}_{jk} \end{pmatrix} \tag{18}$$

where $\mathbf{0}$ is a vector of 0s of length q_{jk}, $\mathbf{D}_{jk}^{\mathrm{T}}\mathbf{D}_{jk} = \mathbf{G}_{jk}$, $\mathbf{e}_{0k} \sim N(0, \mathbf{W}_{kk}^{(r)-1})$ and $\mathbf{e}_{jk} \sim N(0, \mathbf{I})$. This device can be generalized to estimate $\boldsymbol{\alpha}_k$ and even $\boldsymbol{\alpha}$.)

Similarly, taking the row corresponding to $\boldsymbol{\beta}_k$ and rearranging gives

$$\mathbf{X}_k\boldsymbol{\beta}_k^{(r+1)} = \mathbf{H}_k^{(r)}\boldsymbol{\varepsilon}_{0k}^{(r)} \tag{19}$$

where, for $k = 1, 2, \ldots, p$,

$$\mathbf{H}_k^{(r)} = \mathbf{X}_k(\mathbf{X}_k^{\mathrm{T}}\mathbf{W}_{kk}^{(r)}\mathbf{X}_k)^{-1}\mathbf{X}_k^{\mathrm{T}}\mathbf{W}_{kk}^{(r)}$$

and

$$\boldsymbol{\varepsilon}_{0k}^{(r)} = \mathbf{z}_k^{(r)} - \sum_{t=1}^{J_k} \mathbf{Z}_{tk}\boldsymbol{\gamma}_{tk}^{(r+1)} - \mathbf{W}_{kk}^{(r)^{-1}} \sum_{s=1, s\neq k}^{p} \mathbf{W}_{ks}^{(r)}(\boldsymbol{\eta}_s^{(r+1)} - \boldsymbol{\eta}_s^{(r)})$$

for $k = 1, 2, \ldots, p$.

A single rth Newton–Raphson step is achieved by using a backfitting procedure for each k, cycling through equation (19) and then equation (17) for $j = 1, 2, \ldots, J_k$ and cycling over $k = 1, 2, \ldots, p$ until convergence of the set of updated values $\boldsymbol{\alpha}_k^{(r+1)}$ for $k = 1, 2, \ldots, p$. The updated predictors $\boldsymbol{\eta}_k^{(r+1)}$, first derivatives $\mathbf{u}_k^{(r+1)}$, diagonal weighted matrices $\mathbf{W}_{ks}^{(r+1)}$ and adjusted dependent variables $\mathbf{z}_k^{(r+1)}$, for $k = 1, 2, \ldots, p$ and $s = 1, 2, \ldots, p$, are then calculated and the $(r+1)$th Newton–Raphson step is performed, until convergence of the Newton–Raphson algorithm.

References

Aitkin, M. (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–128.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE. Trans. Autom. Control*, **19**, 716–723.

Akaike, H. (1983) Information measures and model selection. *Bull. Int. Statist. Inst.*, **50**, 277–290.

Benjamin, M., Rigby, R. A. and Stasinopoulos, D. M. (2003) Generalized Autoregressive Moving Average Models. *J. Am. Statist. Ass.*, **98**, 214–223.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Besag, J. and Higdon, D. (1999) Bayesian analysis of agriculture field experiments (with discussion). *J. R. Statist. Soc*. B, **61**, 691–746.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.*, **43**, 1–59.
de Boor, C. (1978) *A Practical Guide to Splines*. New York: Springer.
Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc*. B, **26**, 211–252.
Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. New York: Wiley.
Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9–25.
Breslow, N. E. and Lin, X. (1995) Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.
Claeskens, G. and Hjort, N. L. (2003) The focused information criterion. *J. Am. Statist. Ass.*, **98**, 900–916.
Cleveland, W. S., Grosse, E. and Shyu, M. (1993) Local regression models. In *Statistical Modelling in S* (eds I. Chambers and T. Hastie), pp. 309–376. New York: Chapman and Hall.
Cole, T. J., Freeman, J. V. and Preece, M. A. (1998) British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Statist. Med.*, **17**, 407–429.
Cole, T. J. and Green, P. J. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statist. Med.*, **11**, 1305–1319.
Cole, T. J. and Roede, M. J. (1999) Centiles of body mass index for Dutch children age 0-20 years in 1980—a baseline to assess recent trends in obesity. *Ann. Hum. Biol.*, **26**, 303–308.
Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B, **49**, 1–39.
Crisp, A. and Burridge, J. (1994) A note on nonregular likelihood functions in heteroscedastic regression models. *Biometrika*, **81**, 585–587.
CYTEL Software Corporation (2001) *EGRET for Windows*. Cambridge: CYTEL Software Corporation.
Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*, 2nd edn. Oxford: Oxford University Press.
Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. R. Statist. Soc*. B, **57**, 45–97.
Dunn, P. K. and Smyth, G. K. (1996) Randomised quantile residuals. *J. Comput. Graph. Statist.*, **5**, 236–244.
Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statist. Sci.*, **11**, 89–121.
Fahrmeir, L. and Lang, S. (2001) Bayesian inference for generalized additive mixed models based on Markov random field priors. *Appl. Statist.*, **50**, 201–220.
Fahrmeir, L. and Tutz, G. (2001) *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd edn. New York: Springer.
Gange, S. J., Muñoz, A., Sáez, M. and Alonso, J. (1996) Use of the beta–binomial distribution to model the effect of policy changes on appropriateness of hospital stays. *Appl. Statist.*, **45**, 371–382.
Green, P. J. (1985) Linear models for field trials, smoothing and cross-validation. *Biometrika*, **72**, 527–537.
Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.
Harvey, A. C. (1989) *Forecasting Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Hastie, T. and Tibshirani, R. (1993) Varying-coefficient models (with discussion). *J. R. Statist. Soc*. B, **55**, 757–796.
Hastie, T. J. and Tibshiran, R. J. (2000) Bayesian backfitting. *Statist. Sci.*, **15**, 213–223.
Hastie, T. J., Tibshirani, R. J. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.
Hjort, N. L. and Claeskens, G. (2003) Frequentist model average estimation. *J. Am. Statist. Ass.*, **98**, 879–899.
Hodges, J. S. (1998) Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). *J. R. Statist. Soc*. B, **60**, 497–536.
Hodges, J. S. and Sargent, D. J. (2001) Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, **88**, 367–379.
Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Computnl Graph. Statist.*, **5**, 299–314.
Johnson, N. L. (1949) Systems of frequency curves generated by methods of translation. *Biometrika*, **36**, 149–176.
Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994) *Continuous Univariate Distributions*, vol. I, 2nd edn. New York: Wiley.
Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions*, vol. II, 2nd edn. New York: Wiley.
Johnson, N. L., Kotz, S. and Kemp, A. W. (1993) *Univariate Discrete Distributions*, 2nd edn. New York: Wiley.
Kohn, R. and Ansley, C. F. (1998) Equivalence between Bayesian smoothness prior and optimal smoothing for function estimation. In *Bayesian Analysis of Time Series and Dynamic Models* (ed. J. C. Spall), pp. 393–430. New York: Dekker.

Kohn, R., Ansley, C. F. and Tharm, D. (1991) The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *J. Am. Statist. Ass.*, **86**, 1042–1050.
Lange, K. (1999) *Numerical Analysis for Statisticians*. New York: Springer.
Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modelling using the *t* distribution. *J. Am. Statist. Ass.*, **84**, 881–896.
Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc*. B, **58**, 619–678.
Lee, Y. and Nelder, J. A. (2000) Two ways of modelling overdispersion in non-normal data. *Appl. Statist.*, **49**, 591–598.
Lee, Y. and Nelder, J. A. (2001a) Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.
Lee, Y. and Nelder, J. A. (2001b) Modelling and analysing correlated non-normal data. *Statist. Modllng*, **1**, 3–16.
Lin, X. and Zhang, D. (1999) Inference in generalized additive mixed models by using smoothing splines. *J. R. Statist. Soc*. B, **61**, 381–400.
Lopatatzidis, A. and Green, P. J. (2000) Nonparametric quantile regression using the gamma distribution. To be published.
Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Statist. Ass.*, **89**, 1535–1546.
McCulloch, C. E. (1997) Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Ass.*, **92**, 162–170.
Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc*. A, **135**, 370–384.
Nelson, D. B. (1991) Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, **59**, 347–370.
Ortega, J. M. and Rheinboldt, W. C. (1970) *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic Press.
Pawitan, Y. (2001) *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford: Oxford University Press.
Raftery, A. E. (1996) Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, **83**, 251–266.
Raftery, A. E. (1999) Bayes Factors and BIC: comment on 'A critique of the Bayesian Information Criterion for model selection'. *Sociol. Meth. Res.*, **27**, 411–427.
Reinsch, C. (1967) Smoothing by spline functions. *Numer. Math.*, **10**, 177–183.
Rigby, R. A. and Stasinopoulos, D. M. (1996a) A semi-parametric additive model for variance heterogeneity. *Statist. Comput.*, **6**, 57–65.
Rigby, R. A. and Stasinopoulos, D. M. (1996b) Mean and dispersion additive models. In *Statistical Theory and Computational Aspects of Smoothing* (eds W. Härdle and M. G. Schimek), pp. 215–230. Heidelberg: Physica.
Rigby, R. A. and Stasinopoulos, D. M. (2004a) Box-Cox *t* distribution for modelling skew and leptokurtotic data. *Technical Report 01/04*. STORM Research Centre, London Metropolitan University, London.
Rigby, R. A. and Stasinopoulos, D. M. (2004b) Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. *Statist. Med.*, **23**, 3053–3076.
Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
Royston, P. and Altman, D. G. (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Appl. Statist.*, **43**, 429–467.
Schumaker, L. L. (1993) *Spline Functions: Basic Theory*. Melbourne: Krieger.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Silverman, B. W. (1985) Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. R. Statist. Soc*. B, **47**, 1–52.
Smith, P. L. (1979) Splines as a useful and convenient statistical tool. *Am. Statistn*, **33**, 57–62.
Speed, T. P. (1991) Comment on 'That BLUP is a good thing: the estimation of random effects' (by G. K. Robinson). *Statist. Sci.*, **6**, 42–44.
Stasinopoulos, D. M. and Rigby, R. A. (1992) Detecting break points in generalised linear models. *Comput. Statist. Data Anal.*, **13**, 461–471.
Stasinopoulos, D. M., Rigby, R. A. and Akantziliotou, C. (2004) Instructions on how to use the GAMLSS package in R. *Technical Report 02/04*. STORM Research Centre, London Metropolitan University, London.
Stasinopoulos, D. M., Rigby, R. A. and Fahrmeir, L. (2000) Modelling rental guide data using mean and dispersion additive models. *Statistician*, **49**, 479–493.
Thall, P. F. and Vail, S. C. (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657–671.
Tierney, L. and Kadane, J. B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Ass.*, **81**, 82–86.
Tong, H. (1990) *Non-linear Time Series*. Oxford: Oxford University Press.
Verbyla, A. P., Cullis, B. R., Kenward, M. G. and Welham, S. J. (1999) The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Appl. Statist.*, **48**, 269–311.

Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc*. B, **40**, 364–372.
Wahba, G. (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, **4**, 1378–1402.
Wood, S. N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Statist. Soc*. B, **62**, 413–428.
Wood, S. N. (2001) mgcv: GAMs and Generalised Ridge Regression for R. *R News*, **1**, 20–25.
Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *J. Am. Statist. Ass.*, **86**, 79–95.

Discussion on the paper by Rigby and Stasinopoulos

Peter W. Lane (*GlaxoSmithKline, Harlow*)

I congratulate Robert Rigby and Mikis Stasinopoulous on their addition to the toolbox for analytical statistics. They have clearly been working towards the present generality of the generalized additive model for location, scale and shape for several years and have developed the supporting theory in conjunction with a software package in the public domain R system. The model includes many of the modelling extensions that have been introduced by researchers in the past few decades and provides a unifying framework for estimation and inference. Moreover, they have found other directions in which to extend it themselves, allowing for modelling of further parameters beyond the mean and variance and with a much wider class of distributions.

This is a very extensive paper, and it would take much longer than the time that is available today to get to grips with the many ideas and issues that are covered. Two particular aspects encourage me to go away to experiment with the new tool. One is the inclusion of facilities for smooth terms, which have much potential for practial use in handling relationships that must be adjusted for, without the need for a parametric model. I am particularly glad to see facilities for smoothing made available as an integrated part of a general model, unlike the approach that is taken in some statistical software. The other aspect is the provision for non-linear hyperparameters, which I experimented with myself in a class I called generalized non-linear models and made available in GenStat (Lane, 1996). The structure of the generalized additive model for location, scale and shape allows such parameters to be estimated by non-linear algorithms, involving the inevitable concerns over details of the search process, without having to sacrifice the benefits of not having these concerns within the main generalized additive parts of the model.

I am surprised not to see the beta distribution included in the very extensive list of available distributions. In fact, none of the distributions that are listed there are suitable for the analysis of continuous variables observed in a restricted range. In pharmaceutical trials in several therapeutic areas, responses from patients are gathered in the form of a visual analogue scale. This requires patients to mark a point on a line in the range [0,1] to represent some aspect under study, such as their perception of pain. Some of my colleagues (Wu *et al.*, 2003) have investigated the analysis of such data by using the beta distribution, and it would be useful to see how to fit this into the general scheme.

I am very pleased to see that facilities for model checking are also provided and feature prominently in the illustrative examples in this paper. These are invaluable in helping to understand the fit of a model, and in highlighting potential problems.

I would like to raise three concerns with the paper. The main one is with the use of maximum likehood for fitting models with random effects. I am under the impression that such an approach in general leads to biased estimators, and that it is preferable to use restricted maximum likehood. This strikes me as indicating that the generalized linear mixed model and hierarchical generalized linear model approaches are more appropriate for those problems that come within their scope.

My experience with general tools for complex regression models has given me a sceptical outlook when presented with a new one. All too often, I have found that models cannot be applied in practice without extensive knowledge of the underlying algorithms to cope with difficulties in start-up or convergence. As a result, the apparent flexibility of a tool cannot actually be used, and I have to make do with a simpler model than I would like because of difficulties that I cannot overcome. I fear that the disclaimer in Section 5 about potential problems with the likelihood approach for these very general models may signal similar difficulties here. It is noticeable that three of the illustrative examples involve large numbers of observations (over 1000) and the other two, still with over 200 observations, have few parameters.

I am also concerned by the arbitrary nature of the generalized Akaike information criterion that is suggested for comparing models. The examples use three different values, 2.0, 2.4 and 3.0, for what I can

only describe as a 'fudge factor', and they include no comment on why these values are used rather than any others. I am aware that, with large data sets, automatic methods of model selection tend to lead to the inclusion of more model terms than are needed for a reasonable explanation; we need a better approach than is offered by these information criteria.

However, I appreciate that most of my concerns can probably be levelled at any scheme for fitting a wide class of complex models. So I am happy to conclude by proposing a vote of thanks to the authors for a stimulating paper and a new modelling tool to experiment with.

Simon Wood (*University of Glasgow*)

I would like to start by congratulating the authors on a very interesting paper, reporting an impressive piece of work. It is good to see sophisticated approaches to the modelling of the mean being extended to other moments.

The paper is thought provoking in many ways, but I am particularly interested in picking up on Section 3.2.3 and considering how the use of penalized regression spline terms might lead to some simplifications, and perhaps improvements, with respect to fitting algorithms and inference for at least some models in the generalized additive model for location, scale and shape class. For example, if the body mass index model of Section 7.1 is represented by using relatively low rank cubic regression spline bases for h_1–h_4 then equation (8) can be rewritten as

$$\begin{aligned}\mu_i &= \mathbf{A}^{[1]}\boldsymbol{\theta}^{[1]},\\ \log(\sigma_i) &= \mathbf{A}^{[2]}\boldsymbol{\theta}^{[2]},\\ \nu_i &= \mathbf{A}^{[3]}\boldsymbol{\theta}^{[3]},\\ \log(\tau_i) &= \mathbf{A}^{[4]}\boldsymbol{\theta}^{[4]},\end{aligned}$$

where the $\mathbf{A}^{[j]}$ are model matrices and the $\boldsymbol{\theta}^{[j]}$ are coefficients to be estimated. If $\boldsymbol{\theta}' = (\boldsymbol{\theta}^{[1]\prime}, \boldsymbol{\theta}^{[2]\prime}, \boldsymbol{\theta}^{[3]\prime}, \boldsymbol{\theta}^{[4]\prime})$, then the associated penalty on each h_j can be written as $\boldsymbol{\theta}'\mathbf{S}_j\boldsymbol{\theta}$ $(= \int h_j''(x)^2\,\mathrm{d}x)$. Given smoothing parameters λ_j, model estimation can then proceed by direct maximization of the penalized likelihood of the model

$$l(\boldsymbol{\theta}) - \frac{1}{2}\sum_{j=1}^{4}\lambda_j\boldsymbol{\theta}'\mathbf{S}_j\boldsymbol{\theta}$$

by using Newton's method, for example. Following the authors, smoothing parameter estimation by the generalized Akaike information criterion (GAIC) is also straightforward (if computationally time consuming), given that the estimated degrees of freedom for each model parameter are

$$\mathrm{diag}\{(\mathbf{H}_{\boldsymbol{\theta}} + \textstyle\sum \lambda_j\mathbf{S}_j)^{-1}\mathbf{H}_{\boldsymbol{\theta}}\},$$

where $\mathbf{H}_{\boldsymbol{\theta}}$ is the negative Hessian of the unpenalized likelihood with respect to $\boldsymbol{\theta}$. Furthermore, an approximate posterior covariance matrix for the model coefficients can be derived:

$$(\mathbf{H}_{\boldsymbol{\theta}} + \textstyle\sum \lambda_j\mathbf{S}_j)^{-1}.$$

Fig. 14 illustrates the results of applying this approach and should be compared with Fig. 2 of Rigby and Stasinopoulos. All computations were performed using R 2.0.0 (R Development Core Team, 2004). For this example, h_1 was represented by using a rank 20 cubic regression spline whereas h_3 and h_4 were each represented by using rank 10 cubic regression splines (class `cr` smooth constructor functions from R library `mgcv` were used to set up the model matrices and penalty matrices). Given smoothing parameters, the penalized likelihood was maximized by Newton's method with step halving, backed up by steepest descent with line searching, if the Hessian of the penalized log-likelihood was not negative definite. Constants were used as starting values for the functions, these being obtained by fitting a model in which h_1–h_4 were each assumed to be constant. Rapid convergence is facilitated by first conditioning on a moderate constant value for h_4 and optimizing only h_1–h_3. The resulting h_1–h_3-estimates were used as starting values in a subsequent optimization with respect to all the functions. The two-stage optimization helps because of the flatness of the log-likelihood with respect to changes in h_4 corresponding to $\tau > 30$. This penalized likelihood maximization was performed using 'exact' first and second derivatives. The smoothing parameters were estimated by GAIC minimization, with $\# = 2$. The GAIC was optimized by using a quasi-Newton method with finite differenced derivatives (R routine `optim`). The estimated degrees of freedom for the smooth functions were 20, 9.2, 5.9 and 7.4, which are higher than those which were obtained in Section 7.1, since I used $\# = 2$ rather than $\# = 2.4$. Fitting required around a fifth of the time of the `gamlss`

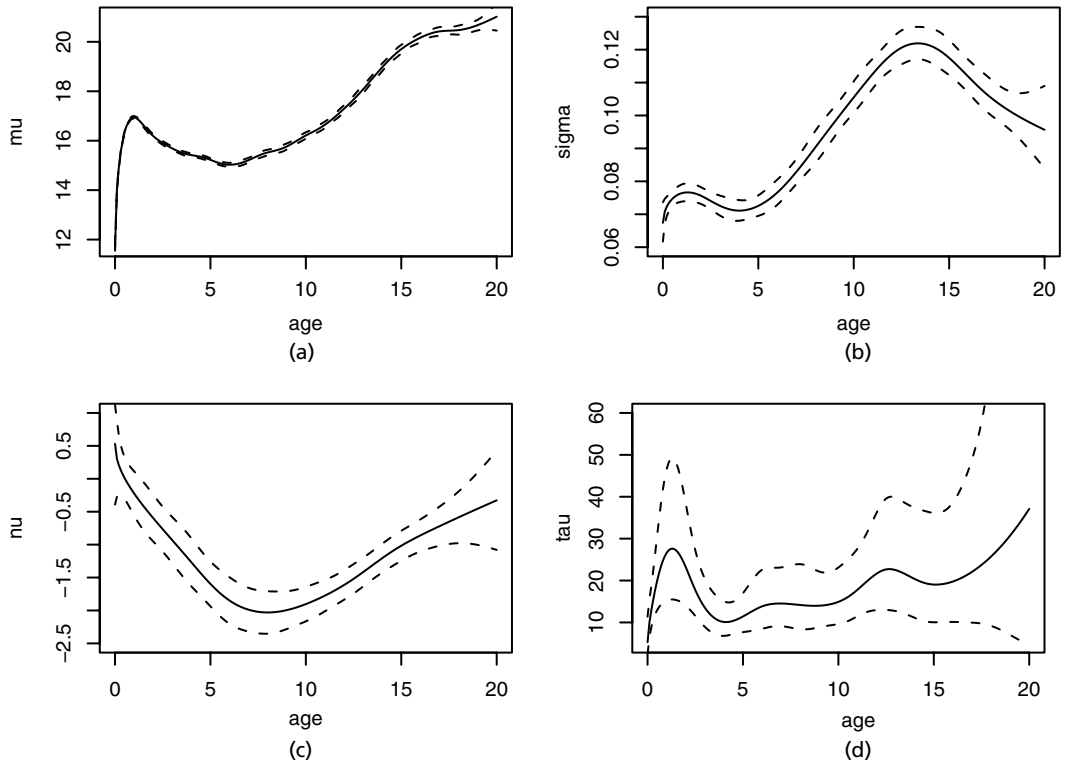

Fig. 14. Equivalent figure to Fig. 2, showing estimates that were achieved by using penalized cubic regression splines to represent the smooth terms in the model (note the wide bands in (d); clearly the data provide only limited information about τ): ———, estimated functions; – – –, limits of twice the standard error bands

package, and with some optimization and judicious use of compiled code a somewhat greater speed up might be expected.

So the direct penalized regression approach to the generalized additive model for location, scale and shape class may have the potential to offer some computational benefits, as well as making approximate inference about the uncertainty of the model components quite straightforward. Clearly, then, this is a paper which not only presents a substantial body of work but also suggests many further areas for exploration, and it is therefore a pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

M. C. Jones (*The Open University, Milton Keynes*)

It is excellent to see three- and four-parameter distributional families being employed for continuous response variables in the authors' general models. My comments on this fine and important paper focus on Section 4.2.

First, there are three-parameter symmetric families on $\Re$, the third parameter, in addition to location and scale which are here set to 0 and 1, controlling the tail weight. The Student t-family, of course, ranges from the normal distribution to distributions with very heavy, power, tails, such as the Cauchy distribution. The power exponential family, in contrast, ranges from as light tailed a distribution as the uniform through the normal and double-exponential distributions, but retaining an exponential nature in the tails. Rider's (1958) rather overlooked distribution with density

$$\frac{(\nu+1)\sin\{\pi/(\nu+1)\}}{2\pi(1+|x|^{\nu+1})}, \qquad \nu>0,$$

might be considered as it ranges all the way from uniform to power tails (including the Cauchy but not the normal).

Second, there are four-parameter families on $\Re$, additionally allowing skewness. Johnson's S_U-family is cited: its simple skewness device, which works fine for the inverse sinh transformation, does not always accommodate skewness attractively, because for other transformations skewness can increase and then decrease again as ν (now denoting the skewness parameter) increases. The symmetric S_U-distribution, which has the normal as least heavy-tailed member, can be extended much of the way towards uniformity by what I call the 'sinh–arcsinh' distribution:

$$z = \sinh\{\tau \sinh^{-1}(y)\}.$$

This seamlessly unites the best aspects of the Johnson S_U- and Rieck and Nedelman (1991) sinh–normal distributions. It can readily be 'skewed' through

$$z = \tfrac{1}{2}[\exp\{\tau_1 \sinh^{-1}(y)\} - \exp\{-\tau_2 \sinh^{-1}(y)\}].$$

The other three-parameter distributions that were mentioned earlier can be 'skewed' either by special means such as Jones and Faddy (2003) for the t-distribution or by

$$\left\{\Gamma\left(1+\frac{1}{\nu_1}\right)+\Gamma\left(1+\frac{1}{\nu_2}\right)\right\}^{-1}\{\exp(-|y|^{\nu_1})\ I(y<0)+\exp(-y^{\nu_2})\ I(y>0)\}$$

(similar to Nandi and Mämpel (1995)) for the exponential power (with a natural analogue for Rider's distribution). Alternatively, general skewing methods such as Azzalini's (1985) $g(y) = 2\,f(y)\,F(\lambda y)$ and the two-piece approach

$$g(y) = 2(1+\lambda^2)^{-1}\lambda\{f(\lambda y)\,I(y<0) + f(y/\lambda)\,I(y>0)\}$$

(Fernandez and Steel, 1998) can be considered.

The remainder of the distributions in Section 4.2 live on $\Re^+$. Three of the four employ the much overrated Box–Cox transformation. A big disadvantage, at least for the purist, is that the Box–Cox transformation requires messy truncated distributions for z with the truncation point depending on the parameters of the transformation. The authors recognize this elsewhere (Rigby and Stasinopoulos, 2004a, b). A better alternative, if one must take the transformation approach, might be to take logarithms and then to employ the wider families of distributions genuinely on $\Re$, such as those above.

But there are many distributions on $\Re^+$ directly including, finally, the generalized gamma family. My only comment here is that this is a well-known family with a long history before an unpublished 2000 report, e.g. Amoroso (1925), Stacy (1962) and Johnson *et al.* (1994), section 8.7.

John A. Nelder (*Imperial College London*)

In my view the use of penalized likelihood (PL) for estimating (σ, ν, τ) as well as μ in expression (4) cannot be justified. The use of restricted maximum likelihood (REML) shows that different functions must be maximized to estimate mean and dispersion parameters. For another misunderstanding of this point see Little and Rubin (2002), section 6.3. The generalization of REML by Lee and Nelder (1996, 2001) shows that an adjusted profile h-likelihood (APHL) should be used for estimation of dispersion parameters. Consider the random-effect model: for $i = 1, \ldots, n$ and $j = 1, 2$

$$y_{ij} = \beta + u_i + e_{ij},$$

where $u_i \sim N(0, \lambda)$ with a known λ and $e_{ij} \sim N(0, \sigma^2)$ are independent. Here their PL estimator gives $\hat{\sigma}^2 = \Sigma_{ij}\, d_{ij}/2n$ with $\hat{\beta} = \bar{y}_{..}$ and $\hat{u}_i = w(\bar{y}_{i.} - \bar{y}_{..})$, $w = \lambda/(\lambda + \sigma^2/2)$ and $d_{ij} = (y_{ij} - \hat{\beta} - \hat{u}_i)^2$. This has a serious bias, e.g. $E(\hat{\sigma}^2) = \sigma^2/2$ when $\lambda = \infty$ (i.e. $w = 1$). Lee and Nelder (2001) showed that the use of APHL in equation (15) gives a consistent REML estimator. PL has been proposed for fitting smooth terms such as occur in generalized additive models. However, in random-effect models the number of random effects can increase with the sample size, so the use of the appropriate APHL is important. If appropriate profiling is used the algebra for fitting dispersion is fairly complicated; I predict that for fitting kurtosis it will be enormously complicated.

Lee and Nelder use extended quasi-likelihood for more general models, where no likelihood is available: for its good performances see Lee (2004). When the model allows exact likelihoods they use them in forming the h-likelihood; even with binary data the h-likelihood method often produces the least bias compared with other methods, including Markov chain Monte Carlo sampling (Noh and Lee, 2004).

Youngjo Lee (*Seoul National University*)
I am unsure by how much the generalized additive model for location, scale and shape class is more general than the hierarchical generalized linear model (HGLM) class of models. Recently, the latter class has been extended to allow random effects in both the mean and the dispersion (Lee and Nelder, 2004). This class enables models with various heavy-tailed distributions to be explored, some of which may be new. Various forms of skewness can also be generated. Although this approach uses a combination of interlinked generalized linear models, it does not mean that we are restricted to the variance function, and higher cumulants, of exponential families.

For example some models in Table 1 can be easily written as instances of the HGLM class; their beta–binomial distribution becomes the binomial–beta HGLM, their negative binomial distribution the Poisson–gamma HGLM, their Pareto distribution the exponential–inverse gamma HGLM etc. For further examples, consider a model

$$y_i = \mu_i + \varepsilon_i,$$

where $\varepsilon_i = \sigma_i e_i$, $e_i \sim N(0, 1)$, and

$$\log(\sigma_i^2) = \alpha + b_i.$$

For b_i there are various possible distributions, For example, if $a_i = \exp(b_i) \sim \alpha/\chi^2_\alpha$ where χ^2_α is a random variable with the χ^2-distribution with α degrees of freedom, then marginally the ε_i follow the t-distribution with α degrees of freedom. Alternatively we may assume that

$$b_i \sim N(0, \delta).$$

An advantage of the normality assumption for b_i is that it allows correlations between b_i and v_i, giving an asymmetric distribution; further complicated random-effect models can be considered. For a more detailed discussion of this and the use of other distributions see Lee and Nelder (2004). In this way we can generate new models which have various forms of marginal skewness and kurtosis. It is not clear, however, that ν and τ in Rigby and Stasinopoulos's equation(4) can be called the skewness and kurtosis.

In summary, models in this paper can generate potentially useful new models, but these will require the proper use of h-likelihood if they are to be useful for inferences.

Mario Cortina Borja (*Institute of Child Health, London*)
I congratulate the authors for a very clear exposition of the foundations of a large class of models, and especially for providing a flexible computational tool to fit and analyse these models. One of the most appealing aspects of the R library that has been written by the authors is how easy it is to incorporate new distributions. I considered the von Mises distribution with density

$$f(\theta; \mu, \kappa) = \frac{\exp\{\kappa \cos(\theta - \mu)\}}{2\pi I_0(\kappa)},$$

where I_0 is the modified Bessel function of the first kind and order 0, $0 < \theta \leqslant 2\pi$, $0 < \mu \leqslant 2\pi$, is the location parameter and $\kappa \geqslant 0$ is the scale parameter; for κ large the distribution has a narrow peak, whereas if $\kappa = 0$ the distribution is the uniform distribution on $(0, 2\pi]$. This distribution is widely used to model seasonal patterns; it is a member of the exponential family and may be modelled in the context of the generalized additive model for location, scale and shape (GAMLSS) by using the link functions $\mu = 2\tan^{-1}(\text{LP})$ and $\kappa = \exp(\text{LP})$, where LP is a linear predictor.

As an example of using the GAMLSS to model circular responses, I have analysed the number of cases of sudden infant death syndrome (SIDS) in the UK by month of death between 1983 and 1998; these data appear in Mooney *et al.* (2003) and were corrected to 31-day months. Though it is not easy to decide on an optimal model, one strong contender, based on the Schwarz Bayesian criterion, fits a constant mean μ (indicating a peak incidence in January) and a natural cubic spline with three effective degrees of freedom as a function of year of death for the scale parameter κ. The fitted smooth curve for this parameter (Fig. 15) may reflect the effect of the 'back-to-sleep' campaign that was implemented in the early 1990s which reduced the number of SIDS cases by 70% in the UK; it corresponds to a dampening of the seasonal effect on SIDS.

Non-symmetric circular distributions and zero-inflated distributions can be modelled as mixtures, and I wonder whether it would be easy to implement these in a GAMLSS.

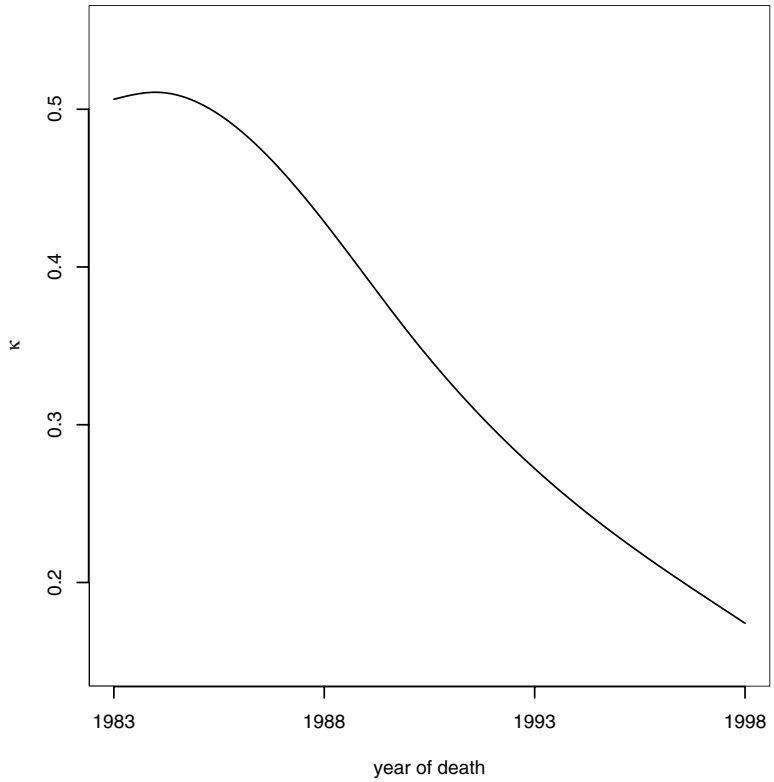

Fig. 15. Effect of year of death on the scale parameter of the von Mises distribution for the number of SIDS cases in the UK, 1983–1998

N. T. Longford (*SNTL, Leicester*)

This paper competes with Lee and Nelder (1996) and their extensions, conveying the message that for any data structure and associations that we could possibly think of there are models and algorithms to fit them. But now models are introduced even for some structures that we would not have thought of .... I want to rephrase my comment on Lee and Nelder (1996) which I regard equally applicable to this paper. The new models are top of the range mathematical Ferraris, but the model selection that is used with them is like a sequence of tollbooths at which partially sighted operators inspect driver's licences and road worthiness certificates.

Putting the simile aside, let the alternative models that are considered in either of the examples be $1, \ldots, M$, and the estimators that would be applied, if model m were selected, $\hat{\theta}_1, \ldots, \hat{\theta}_M$, each of them unbiased for the parameter of interest θ, and having sampling variance s_m^2 estimated without bias by $\hat{s}_m^2$, *if* model m is appropriate: not when it is selected, but when it is valid! Model selection, by whichever criterion and sequence of model comparisons, leads to the estimator

$$\tilde{\theta} = \sum_m I_m \hat{\theta}_m,$$

where I_m indicates whether model m is selected ($I_m = 1$) or not ($I_m = 0$). This is a mixture of the single-model-based estimators; in all but some degenerate cases it is biased for θ. Further, $\mathrm{var}(\tilde{\theta})$ is conventionally estimated by

$$\tilde{s}^2 = \sum_m I_m \hat{s}_m^2,$$

assuming that whichever model is selected is done so with certainty. This can grossly underestimate the

mean-squared error of $\tilde{\theta}$, and does so not only because $\tilde{\theta}$ is biased. The distribution of the mixture $\tilde{\theta}$ is difficult to establish because the indicators I_m are correlated with $\hat{\theta}_m$.

A misconception underlying all attempts to find *the* model is that the maximum likelihood assuming the most parsimonious valid model is efficient. This is only asymptotically so. For some parameters (and finite samples), maximum likelihood under some invalid submodels of this model is more efficient because the squared bias that is incurred is smaller than the reduction of the variance. Proximity to asymptotics is not indicated well by the sample size because information about the parameters for the distributional tail behaviour is relatively modest in the complex models engaged.

Longford (2003, 2005) discusses the problem and proposes a solution.

Adrian Bowman (*University of Glasgow*)

I congratulate the authors on a further substantial advance in flexible modelling. The generalized linear model represented a major synthesis of regression models by allowing a wide range of types of response data and explanatory variables to be handled in a single unifying framework. The generalized additive model approach considerably extended this by allowing smooth nonparametric effects to be added to the list of available model components. The authors have gone substantially further by incorporating the rich set of tools that has been created by recent advances in mixed models and, in addition, by allowing models to describe the structure of parameters beyond the mean. The end result is an array of models of astonishing variety.

One major issue which this complexity raises is what tools can be used to navigate such an array of models? The authors rightly comment that particular applications provide contexts which can give guidance on the structure of individual components. Where the aim is one of prediction, as is the case in several of the examples of the paper, criteria such as Akaike's information criterion and the Schwarz Bayesian criterion are appropriate. However, where interest lies in more specific aspects of model components, such as the identification of whether an individual variable enters the model in a linear or nonparametric manner, or indeed may have no effect, then prediction-based methods are less appropriate. Even with the usual form of generalized additive model, likelihood ratio test statistics do not have the usual χ^2 null distributions and the problem seems likely to be exacerbated in the more complex setting of a generalized additive model for location, scale and shape.

In view of this, any further guidance which the authors could provide on how to interpret the global deviance column of Table 2, or more generally on appropriate reference distributions when comparing models, would be very welcome.

The following contribution was received in writing after the meeting.

T. J. Cole (*Institute of Child Health, London*)

I congratulate the authors on their development of the generalized additive model for location, scale and shape (GAMLSS). Its flexible approach to the modelling of higher moments of the distribution is very powerful and works particularly well with age-related reference ranges.

In my experience with the *LMS* method (Cole and Green, 1992), which is a special case of the GAMLSS, it is difficult to choose the effective degrees of freedom (EDFs) for the cubic smoothing spline curves as there is no clear criterion for goodness of fit (see Pan and Cole (2004)). In theory the authors' generalized Akaike information criterion GAIC(#) (Section 6.2) provides such a criterion, but in practice it can be supremely sensitive to the choice of the hyperparameter #. I am glad that the authors chose to highlight this in their first example (Section 7.1). With $\# = 2.4$ the shape parameter τ was modelled as a cubic smoothing spline with 6.1 EDFs (Fig. 2), whereas with $\# = 2.5$ it was modelled as a constant. The two most well-known cases of the GAIC are the AIC itself (where $\# = 2$) and the Schwarz Bayesian criterion (SBC) (where $\# = \log(n) = 9.9$ here), so the distinction between 2.4 and 2.5 is clearly tiny on this scale. The use of the SBC in the example would have led to a much more parsimonious model than for GAIC(2.5).

The take-home message is that, although optimal GAMLSSs are simple to fit conditional on #, the choice of # is largely subjective on the scale from 2 to $\log(n)$ and can affect the model dramatically. In my view # should reflect the sample size in some way, so I prefer the SBC to the AIC. In addition it is good practice to reduce the EDFs as far as possible (Pan and Cole, 2004), which comes to the same thing. I also wonder whether a different GAIC might be applied to the different parameters of the distribution, so that for example an extra EDF used to model the shape parameter should be penalized more heavily than an extra EDF for the mean.

The **authors** replied later, in writing, as follows.

We thank all the disscusants for their constructive comments and reply below to the issues that were raised.

Distributions

An important advantage of the generalized additive model for location, scale and shape (GAMLSS) is that the model allows any distribution for the response variable y. In reply to Dr Borja, mixture distributions (including zero-inflated distributions) are easily implemented in a GAMLSS. For example, a zero-inflated negative binomial distribution (a mixture of zero with probability ν and a negative binomial NB(μ, σ) distribution with probability $1-\nu$) is easily implemented as a three-parameter distribution (e.g. with log-links for μ and σ and a logit link for ν). The beta distribution BE(μ, σ), which was suggested by Dr Lane, has now been implemented, as has an inflated beta distribution with additional point probabilities for y at 0 and 1.

The exponential family distribution that is used in generalized linear, additive and mixed models usually has at most two parameters: a mean parameter μ and a scale parameter ϕ ($=\sigma$ in our notation). Having only two parameters it cannot model skewness and kurtosis. The exponential family distribution has been approximated by using extended quasi-likelihood (see McCullagh and Nelder (1989)) and used in hierarchical generalized linear models (HGLMs) by Lee and Nelder (1996, 2001). However, extended quasi-likelihood is not a proper distribution, as discussed in Section 2.3 of the paper, and suffers from the same skewness and kurtosis restrictions as the exponential family. The range of distributions that are available in HGLMs is extended via a random-effect term. However, the GAMLSS allows any distribution for y and is conceptually simpler because it models the distribution of y directly, rather than via a random-effect term. The level of generality of the double HGLM will be clearer on publication of Lee and Nelder (2004).

The four-parameter distributions on $\Re$ that were discussed by Professor Jones can be implemented in a GAMLSS. The Box–Cox t- and Box–Cox power exponential distributions in the paper are four-parameter distributions on $\Re^+$ for which there are fewer direct contenders. They are easy to fit in our experience and provide generalizations of the Box–Cox normal distribution (Cole and Green, 1992), which is widely used in centile estimation, allowing the modelling of kurtosis as well as skewness. Users are also welcome to implement other distributions.

Restricted maximum likelihood

Dr Lane and Professor Nelder highlight the use of restricted maximum likelihood (REML) estimation for reducing bias in parameter estimation. In the paper, the random-effects hyperparameters $\boldsymbol{\lambda}$ are estimated by REML estimation, whereas the fixed effects parameters $\boldsymbol{\beta}$ and random-effects parameters $\boldsymbol{\gamma}$ are estimated by posterior mode estimation, conditional on the estimated $\boldsymbol{\lambda}$. If the total (*effective*) degrees of freedom for estimating the random effects $\boldsymbol{\gamma}$ and the fixed effects $\boldsymbol{\beta}_1$ for the distribution parameter $\boldsymbol{\mu}$ are substantial relative to the total degrees of freedom (i.e. the sample size), then REML estimation of the hyperparameters $\boldsymbol{\lambda}$ *and* the fixed effects $(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$ for parameters $(\boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ respectively may be preferred. This is achieved in a GAMLSS by treating $(\boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$ in the same way as $\boldsymbol{\lambda}$ in Appendix A.2.3 and obtaining the approximate marginal likelihood $l(\boldsymbol{\zeta}_1)$ for $\boldsymbol{\zeta}_1 = (\boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \boldsymbol{\lambda})$ obtained by integrating out $\boldsymbol{\zeta}_2 = (\boldsymbol{\beta}_1, \boldsymbol{\gamma})$ from the joint posterior density of $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$, giving $l(\boldsymbol{\zeta}_1) = \hat{l}_h - \frac{1}{2}\log|\hat{H}/2\pi|$, where $\hat{l}_h = l_h(\boldsymbol{\zeta}_1, \hat{\boldsymbol{\zeta}}_2)$ and $\hat{H} = H(\hat{\boldsymbol{\zeta}}_2) = -E(\partial^2 l_h / \partial \boldsymbol{\zeta}_2 \partial \boldsymbol{\zeta}_2^{\mathrm{T}})$, evaluated at $\hat{\boldsymbol{\zeta}}_2$, the posterior mode estimate of $\boldsymbol{\zeta}_2$ given $\boldsymbol{\zeta}_1$. Hence REML estimation of $\boldsymbol{\zeta}_1$ is achieved by maximizing $l(\boldsymbol{\zeta}_1)$ over $\boldsymbol{\zeta}_1$. This procedure leads to REML estimation of the scale and shape parameters and the random-effects hyperparameters.

For example, in Hodges's data from Section 7.2 of the paper, the above procedure gives the following REML estimates (with the original estimates given in parentheses): $\hat{\beta}_2 = -2.14$ (-2.21), $\hat{\beta}_3 = -0.0222$ (-0.0697), $\hat{\beta}_4 = -2.19$ (2.15), $\hat{\sigma}_1 = 14.0$ (13.1) and $\hat{\sigma}_2 = 0.0590$ (0.0848).

Bias in the estimators may be further reduced by use of a second-order Laplace approximation to the integrated joint posterior density above; see Breslow and Lin (1995) and Lee and Nelder (2001).

Alternatively, other methods of bias reduction, e.g. bootstrapping, could be considered.

Model selection

Dr Lane and Professor Cole highlight the issue of the choice of penalty # in the generalized Akaike information criterion GAIC(#) that is used in the paper for model selection. The use of criterion GAIC(#) allows investigation of the sensitivity of the selected model to the choice of penalty #. This is well illustrated in the Dutch girls' body mass index (BMI) data example from Section 7.1. The resulting optimal effective degrees of freedom that were selected for μ, σ, ν and τ and the estimated parameter ξ in the transformation $x = \mathrm{age}^{\xi}$ are given in Table 4 for each of the penalties $\# = 2, 2.4, 2.5, 9.9$.

Table 4. Model selected by criterion GAIC(#) with penalty #

# (criterion)	df_μ	df_σ	df_ν	df_τ	ξ
2 (AIC)	16.9	8.7	5.0	9.5	0.51
2.4 (GAIC)	16.2	8.5	4.7	6.1	0.50
2.5 (GAIC)	16.0	8.0	4.8	1	0.52
9.9 (SBC)	12.3	6.3	3.7	1	0.53

The apparent sensitivity of df_τ to # is due to the existence of two local optima. The value $\# = 2.4$ that is used in the paper is the critical value of # above which the optimization switches from one local optimum to the other. Reducing # below 2.4, or increasing # above 2.5, changes the selected degrees of freedom smoothly. Hence there are two clearly different models for the BMI, one corresponding to $\# \leqslant 2.4$ and the other corresponding to $\# \geqslant 2.5$.

A sensitivity analysis of the chosen model to outliers shows that the non-constant τ-function for $\# = 2.4$ in Fig. 2(d), and in particular the minima in τ at 0 and 4 years (with corresponding peaks in the kurtosis) are due to substantial numbers of outliers in the age ranges 0–0.5 and 3–5 years respectively. Consequently we believe that these peaks in kurtosis may be genuine, requiring physiological explanation. We therefore recommend the chosen model for $\# = 2.4$ as in the paper.

In our opinion the Schwarz Bayesian criterion (SBC) is too conservative (i.e. restrictive) in its model selection, leading to bias in the selected functions for μ, σ, ν and τ (particularly at turning-points), whereas the AIC is too liberal, leading to rough (or erratic) selected functions. Fig. 16 gives the selected parameter functions using the AIC. Compare this with Fig. 14 of Simon Wood. The standard errors in Fig. 16 are conditional on the chosen degrees of freedom and ξ, and on the other selected parameter functions. The final selection of model(s) should be made with the expert prior knowledge of specialists in the field.

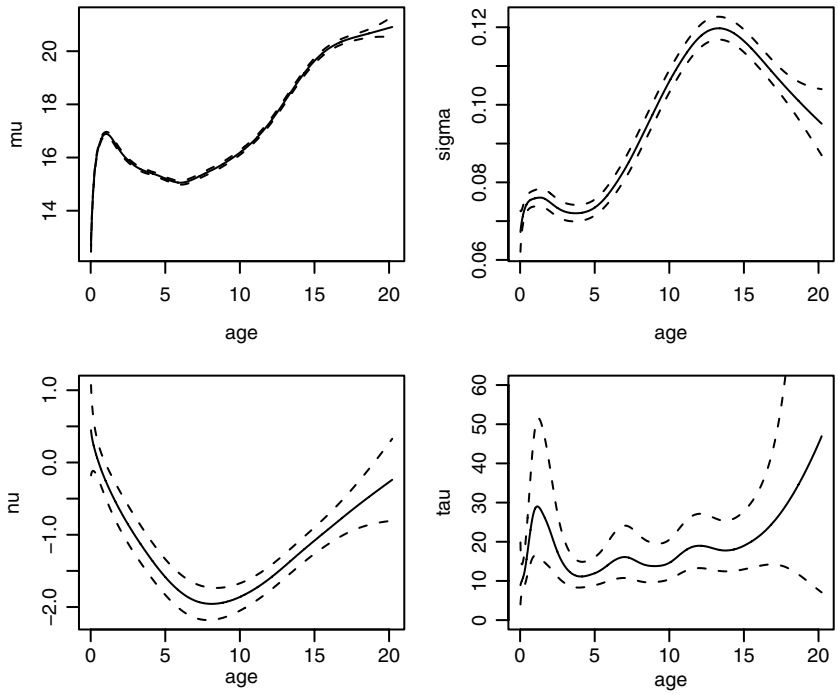

Fig. 16. BMI data: fitted parameters against age by using the AIC for model selection

Conditioning on a single selected model ignores model uncertainty and generally leads to an underestimation of the uncertainty about quantities of interest, as discussed in Section 6.1. This issue was also raised by Dr Longford. Clearly it is an important issue, but not the focus of the current paper.

Where the focus is on whether an explanatory variable, say x, has a significant effect (rather than on prediction), then for a parametric GAMLSS this can be tested by using the generalized likelihood ratio test statistic Λ, as discussed in Section 6.2. The inadequacy of a linear function in x can be established by testing a linear against a polynomial function in x using Λ. The statistic Λ may be used as a guide to comparing a linear with a nonparametric smooth function in x, although, as pointed out by Professor Bowman, the asymptotic χ^2-distribution no longer applies, and so a formal test is not available.

Algorithm convergence

Dr Lane highlights the issue of possible convergence problems. Occasional problems with convergence may be due to one of the following reasons: using a highly inappropriate distribution for the response variable y (e.g. a symmetric distribution when y is highly skewed), using an unnecessarily complicated model (especially for σ, ν or τ), using extremely poor starting values (which is usually overcome by fitting a related model and using its fitted values as starting values for the current model) or overshooting in the Fisher scoring (or quasi-Newton) algorithm (which is usually overcome for parametric models by using a reduced step length). Hence any convergence problems are usually easily resolved. The possibility of multiple maxima is investigated by using different starting values.

Extensions to generalized additive models for location, scale and shape

The GAMLSS has been extended to allow for non-linear parametric terms, non-normal random-effects terms, correlations between random effects for different distribution parameters and incorporating priors for $\boldsymbol{\beta}$ and/or $\boldsymbol{\lambda}$.

Conclusion

The GAMLSS provides a very general class of models for a univariate response variable, presented in a unified and coherent framework. The GAMLSS allows any distribution for the response variable and allows modelling of all the parameters of the distribution. The GAMLSS is highly suited to flexible data analysis and provides a framework that is suitable for educational objectives.

References in the discussion

Amoroso, L. (1925) Ricerche intorno alla curve dei redditi. *Ann. Mat. Pura Appl.* IV, **2**, 123–159.

Azzalini, A. (1985) A class of distributions which includes the normal ones. *Scand. J. Statist.*, **123**, 171–178.

Breslow, N. E. and Lin, X. (1995) Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Cole, T. J. and Green, P. J. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statist. Med.*, **11**, 1305–1319.

Fernandez, C. and Steel, M. F. J. (1998) On Bayesian modeling of fat tails and skewness. *J. Am. Statist. Ass.*, **93**, 359–371.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994) *Continuous Univariate Distributions*, vol. I, 2nd edn. New York: Wiley.

Jones, M. C. and Faddy, M. J. (2003) A skew extension of the t-distribution, with applications. *J. R. Statist. Soc.* B, **65**, 159–174.

Lane, P. W. (1996) Generalized nonlinear models. In *Compstat Proceedings in Computational Statistics* (ed. A. Prat), pp. 331–336. Heidelberg: Physica.

Lee, Y. (2004) Estimating intraclass correlation for binary data using extended quasi-likelihood. *Statist. Modllng*, **4**, 113–126.

Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619–678.

Lee, Y. and Nelder, J. A. (2001) Hierarchical generalised linear models: a synthesis of generalised linear models, random effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y. and Nelder, J. A. (2004) Double hierarchical generalized linear models. To be published.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. New York: Wiley.

Longford, N. T. (2003) An alternative to model selection in ordinary regression. *Statist. Comput.*, **13**, 67–80.

Longford, N. T. (2005) *Modern Analytical Equipment for the Survey Statistician: Missing Data and Small-area Estimation*. New York: Springer. To be published.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Mooney, J. A., Helms, P. J. and Jolliffe, I. T. (2003) Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. *Computnl Statist. Data Anal.*, **41**, 505–513.

Nandi, A. K. and Mämpel, D. (1995) An extension of the generalized Gaussian distribution to include asymmetry. *J. Franklin Inst.*, **332**, 67–75.
Noh, M. and Lee, Y. (2004) REML estimation for binary data in GLMMs. To be published.
Pan, H. and Cole, T. J. (2004) A comparison of goodness of fit tests for age-related reference ranges. *Statist. Med.*, **23**, 1749–1765.
R Development Core Team (2004) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. (Available from `http://www.R-project.org`.)
Rider, P. R. (1958) Generalized Cauchy distributions. *Ann. Inst. Statist. Math.*, **9**, 215–223.
Rieck, J. R. and Nedelman, J. R. (1991) A log-linear model for the Birnbaum–Saunders distribution. *Technometrics*, **33**, 51–60.
Rigby, R. A. and Stasinopoulos, D. M. (2004a) Box-Cox *t* distribution for modelling skew and leptokurtotic data. *Technical Report 01/04*. STORM Research Centre, London Metropolitan University, London.
Rigby, R. A. and Stasinopoulos, D. M. (2004b) Smooth centile curves for skew and kurtotic data modelled using the Box-Cox Power Exponential distribution. *Statist. Med.*, **23**, 3053–3076.
Stacy, E. W. (1962) A generalization of the gamma distribution. *Ann. Math. Statist.*, **33**, 1187–1192.
Wu, Y., Fedorov, V. V. and Propert, K. J. (2003) Optimal design for beta distributed responses. *Technical Report 2004-1*. GlaxoSmithKline, Collegeville.