Penalised likelihood in non-standard regression problems

Peter J. Green, University of Bristol[1]

1 Introduction

One of the beauties of the roughness penalty approach to nonparametric regression is its conceptual versatility. Applications of roughness penalties in a variety of 'standard' situations, almost all of them naturally arising in linear modelling or generalized linear modelling, are now fairly routine. In this paper we explore a number of extensions. Some of these are to more general problems, while others are to specific problems not easily or naturally dealt with by other methods.

The various problems discussed here are of course intended to be of intrinsic interest, but perhaps more importantly they provide 'templates' for the application of the roughness penalty idea to readers' own smoothing problems.

This paper is an edited version of part of one chapter of the recent monograph by Green and Silverman (1994), and is reproduced by permission of the authors. References to sections numbered in the form 5.2 are to that monograph.

2 Nonparametric link functions

Consider an ordinary GLM, but suppose that in addition to the regression parameter $\boldsymbol{\beta}$, the link function G is also unknown. Of course, in truth, this is almost always the case, except when a very definite model for the mechanism generating the data is available, but the usual practice is to fix G nevertheless, and to concentrate on the estimation of $\boldsymbol{\beta}$.

If estimating G is of interest, one possible approach is the parametric link function method of Scallan, Gilchrist and Green (1984). However it is more in the spirit of the roughness penalty approach to assume that G is completely unknown, but for smoothness assumptions, and to estimate it nonparametrically simultaneously with $\boldsymbol{\beta}$; in this section we discuss such an approach.

It turns out to be more straightforward to work not directly with the link function defined by

$$G(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

but with a function related to its inverse. This will avoid the need to impose awkward constraints to guarantee the existence of certain inverse functions to make our model well-defined. We will therefore assume that our responses are drawn independently from an exponential family density or probability function, where the natural parameter θ_i for the i^{th} observation is related to the linear predictor by

$$\theta_i = g(\mathbf{x}_i^T \boldsymbol{\beta}). \tag{1}$$

Both g and $\boldsymbol{\beta}$ are unknown, and in the spirit of our general approach, we could attempt to estimate them both by maximizing the penalized likelihood

$$\ell(\boldsymbol{\theta}, \phi) - \tfrac{1}{2}\lambda \int g''(\eta)^2 d\eta. \tag{2}$$

Notice, however, that the integration is now in the space of the linear predictor η, not an explanatory variable t. An appealing property of this criterion combined with this parameterization of the model is that perfectly smooth g (those for which $\int g''^2 = 0$) correspond to choice of the canonical link function.

Inspection of (1) reveals a certain lack of identifiability. For example if $\boldsymbol{\beta}$ is replaced by $\tilde{\boldsymbol{\beta}}$ for some scalar c, and g is replaced by $\tilde{g}$, where $\tilde{g}(\eta_i) = g(c^{-1}\eta_i)$, then $\boldsymbol{\theta}$ is unchanged. Similarly, if the linear predictor includes a constant term, there is a translation equivariance as well. A full analysis of such equivariances is quite complicated, and will not be attempted here. Instead, we will only consider models in which the linear predictor *does* include a constant term, and constrain the parameter vector $\boldsymbol{\beta}$ so that the linear predictor is standardized to have mean 0 and variance 1, for the particular set of explanatory variables $\mathbf{x}_i$ to hand. It is evident that this can always be imposed by an explicit scaling of $\boldsymbol{\beta}$, and a shift of the constant term. The rationale for this pair of constraints is that these are precisely the degrees of freedom absorbed by the linear regressions that form the perfectly smooth functions g.

[1]Department of Mathematics, University of Bristol, Bristol, BS8 1TW, UK. Email: `P.J.Green@bristol.ac.uk`

The estimating equations that arise from maximizing (2) subject to the constraints

$$\sum \eta_i = 0 \quad \text{and} \quad \sum \eta_i^2 = (n-1) \tag{3}$$

have awkward Lagrange multiplier terms. This route will not be pursued here. Instead we will simply alternate between maximizing (2) over $\boldsymbol{\beta}$ for fixed g, and over g for fixed $\boldsymbol{\beta}$. After each update of $\boldsymbol{\beta}$, $\boldsymbol{\eta}$ and g will be rescaled as described above to impose (3), while leaving $\boldsymbol{\theta}$ unchanged.

For fixed g, the model is an ordinary parametric GLM, so maximization of (2) is an application of the methods of Section 5.2. An updated estimate of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}^{new} = (X^T W^{(1)} X)^{-1} X^T W^{(1)} \mathbf{z}^{(1)},$$

where

$$z_i^{(1)} = \frac{Y_i - \mu_i}{b''(\theta_i) g'(\eta_i)} + \mathbf{x}_i^T \boldsymbol{\beta},$$

and $W^{(1)}$ is the diagonal matrix with

$$W_{ii}^{(1)} = b''(\theta_i)\{g'(\eta_i)\}^2.$$

Conversely for $\boldsymbol{\beta}$ and hence $\boldsymbol{\eta}$ fixed, our model is a purely nonparametric GLM, using the canonical link function, with η_i as the values of a one-dimensional working explanatory variable, so the methods of Section 5.3 are relevant. The update is given by

$$\mathbf{g}^{new} = (W^{(2)} + \alpha K)^{-1} W^{(2)} \mathbf{z}^{(2)},$$

with

$$z_i^{(2)} = \theta_i + \frac{Y_i - b'(\theta_i)}{b''(\theta_i)}$$

and

$$W_{ii}^{(2)} = b''(\theta_i).$$

Thus no really new ideas are needed to handle the nonparametric link function model: this is another example of the flexible, modular way in which computational tools arising from roughness penalty methods can be combined to produce new methodology.

Before giving an example, we close this section with a word of warning. As with many iterative algorithms for nonlinear models, there is no guarantee of convergence from all initial estimates. In most cases, the algorithms are well-behaved for most data sets, but it does seem that the algorithm above is prone to limit-cycle behaviour in some cases, typically when the smoothing parameter is such as to allow more than about 5 degrees of freedom for the nonparametric curve. In our experience, this problem can be circumvented by adjusting the step size in both of the updating steps, that is, by updating to a convex combination of the previous value and that suggested by the algorithm as described. Since both $\boldsymbol{\beta}$ and g are updated twice in each cycle, because of the rescaling, it is of course the previous value *at that point in the cycle* that should be used here.

2.1 Application to tumour prevalence data

As a demonstration of the methodology introduced above, we show in Figure 1 the results of applying it to the data on male rats from the tumour prevalence study described in Section 5.5. (The data is that of Dinse and Lagakos (1983, 1984).) The number of degrees of freedom for the smooth link function was fixed at 5, and the linear predictor included terms corresponding to all four of the explanatory variables (x_1, x_2, x_3 and t), all featuring linearly. Figure 1 displays the observed responses Y_i, plotted against the fitted values of the linear predictor $\hat{\eta}_i$, and the smooth fitted probability curve $\exp\{\hat{g}(\eta)\}/[1+\exp\{\hat{g}(\eta)\}]$.

In order to provide a comparison with a standard analysis, the fitted probabilities from a fully parametric logistic regression on x_1, x_2, x_3 and t are also displayed, plotted against the *non*parametric fitted $\hat{\eta}_i$.

3 Composite likelihood function regression models

We now turn to extensions of the generalized linear model structure. In this section, we relax further some of the specific assumptions used by Nelder and Wedderburn (1972) to define generalized linear models. In particular, we will no longer insist on the exponential family of distributions, on the independence of the

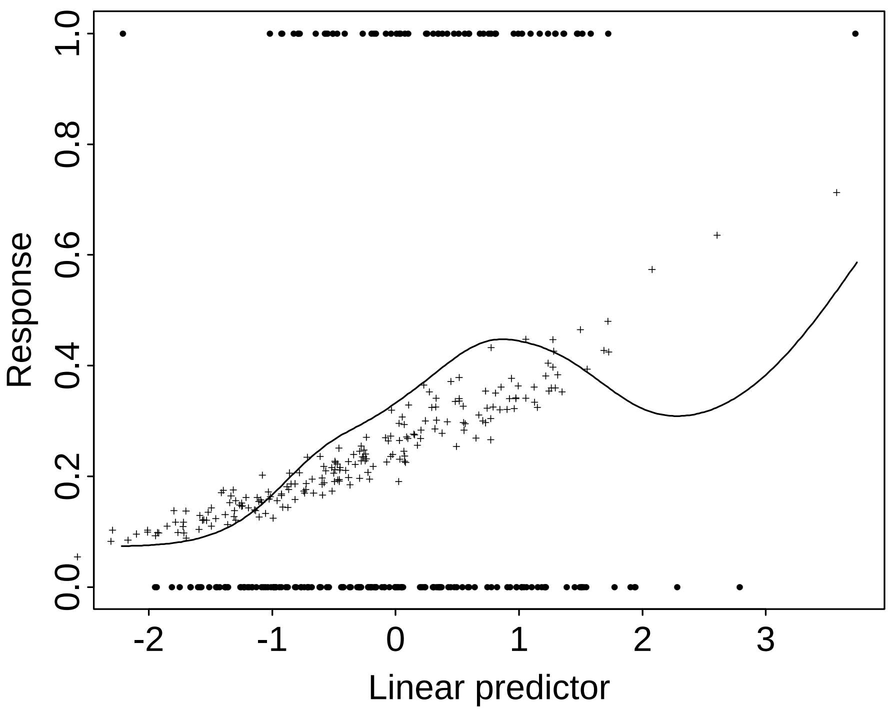

Figure 1: Nonparametric link function analysis, allowing 5 d.f., for the male rats in the Dinse and Lagakos study. Smooth curve shows nonparametric fitted probabilities, + signs those obtained from a standard logistic regression, and points the raw data.

responses, or on the linearity of the regression function. What will be kept is the pivotal role of the *predictors* $\{\eta_i\}$ in decomposing the regression model into systematic and random components. Among examples that can then be covered that lie outside the scope of generalized linear models are

- multiple linear regression with an arbitrary error distribution,
- Gaussian nonlinear regression,
- models for censored or grouped data,
- generalizations of probit analysis to allow synergistic or antagonistic interaction between drugs, and
- multinomial models for regression analysis of ordered categorical data.

This framework for regression has been further developed by Green (1984) and Green (1987), particularly from an algorithmic perspective.

We suppose that we are given a regression function that determines the predictors $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^T$ in terms of an unknown p-vector of parameters $\boldsymbol{\beta}$, an unknown function g of a single variable t, and observed explanatory variables. Suppose that $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\beta}, g)$ depends on g only through the values of $g(t_1), g(t_2), \ldots,$ $g(t_m)$, where $t_1, t_2, \ldots, t_m$ are known. Often m will be the same as n, although this is not necessary, and $t_1, t_2, \ldots, t_n$ may then be observed values of an explanatory variable for each of n cases.

This regression function provides the systematic component of the model, and the random component is provided by the log-likelihood

$$\ell = \ell(\mathbf{Y}; \boldsymbol{\eta}, \phi) = \ell(\boldsymbol{\eta}, \phi), \text{say},$$

of the data $\mathbf{Y}$ in terms of the vector of predictors $\boldsymbol{\eta}$. We are interested in estimating $\boldsymbol{\beta}$ and g, in order to quantify the dependence of the responses $\mathbf{Y}$ on the explanatory variables $\mathbf{x}$ and t. We do so by maximizing the penalized log-likelihood

$$\Pi = \ell\{\boldsymbol{\eta}(\boldsymbol{\beta}, g), \phi\} - \tfrac{1}{2}\lambda \int g''(t)^2 dt. \tag{4}$$

The extra generality actually causes very few new difficulties. Suppose for notational simplicity that there is no nuisance parameter ϕ parameterizing the random component of the model or, which amounts to the same thing, that $\phi = 1$. The restrictions placed on the dependence of $\boldsymbol{\eta}$ on g ensure that once again the estimate of g is guaranteed to be a natural cubic spline and we have only to solve a finite-dimensional optimization problem. Once again we attempt an iterative solution by Fisher scoring, and evaluate the requisite derivatives of Π by the chain rule. Let $\mathbf{g} = (g(s_1), \ldots, g(s_q))^T$ where $s_1, s_2, \ldots, s_q$ denote the ordered distinct values among $t_1, t_2, \ldots, t_m$. Then define

$$X = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}, \quad N = \frac{\partial \boldsymbol{\eta}}{\partial \mathbf{g}} \tag{5}$$

and

$$W = E\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}\right); \tag{6}$$

let

$$\mathbf{z} = W^{-1}\frac{\partial \ell}{\partial \boldsymbol{\eta}} + X\boldsymbol{\beta} + N\mathbf{g}. \tag{7}$$

With these changes of definition, all of the arguments of Section 5.4 still apply and we again obtain the updating equations

$$\begin{bmatrix} X^T W X & X^T W N \\ N^T W X & N^T W N + \alpha K \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}^{new} \\ \mathbf{g}^{new} \end{pmatrix} = \begin{pmatrix} X^T W \mathbf{z} \\ N^T W \mathbf{z} \end{pmatrix}. \tag{8}$$

Various special cases must be distinguished when it comes to the practical implementation of these updating equations. The standard, easy, case arises when W is diagonal and N is an incidence matrix.

4 A varying coefficient model with censoring

To give some of the flavour of the models of the preceding section, we consider here a rather contrived example in which appear the ingredients of semiparametric modelling, censoring, an error distribution outside the exponential family, and the varying coefficient models introduced and studied by Hastie and Tibshirani (1993). This combination by no means exhausts the scope of composite likelihood regression models!

Suppose that potential responses $\{\tilde{Y}_i, i = 1, 2, \ldots, n\}$ are independently drawn from a location family

$$P\{\tilde{Y}_i \leq y\} = \Psi(y - \eta_i) \tag{9}$$

where Ψ is a known distribution function with density ψ, and the predictor η_i is given by

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + s_i g(t_i). \tag{10}$$

Here, $\mathbf{x}_i, s_i$ and t_i are observed explanatory variables (a p-vector and two scalars, respectively) and the parameter $\boldsymbol{\beta}$ and function g are, as usual, to be estimated. Note that the nonparametric function g appears here as the *slope* of the regression on the variable s; its argument t is called an *effect modifying variable* by Hastie and Tibshirani.

The potential responses are subject to censoring, and we actually observe only

$$Y_i = \min(\tilde{Y}_i, c_i)$$

and

$$\delta_i = \begin{cases} 1 & \text{if} \quad Y_i < c_i \\ 0 & \text{if} \quad Y_i = c_i, \end{cases}$$

where c_i is the non-random censoring time for the i^{th} observation.

The log-likelihood thus takes the form

$$\ell(\boldsymbol{\eta}) = \sum_{i=1}^{n} [\delta_i \log \psi(Y_i - \eta_i) + (1 - \delta_i) \log\{1 - \Psi(Y_i - \eta_i)\}] \tag{11}$$

and the maximum penalized likelihood estimates of $\boldsymbol{\beta}$ and g are those maximizing

$$\ell(\boldsymbol{\eta}) - \tfrac{1}{2}\lambda \int g''(t)^2 dt$$

as usual. This is an example of (4). Construction of an algorithm to carry out the estimation proceeds as in Section 3, with one exception. Since it is not usually reasonable to assume that censoring times $\{c_i\}$ are known for *uncensored* observations, the expectation in (6) cannot be computed. We therefore take instead

$$W = -\frac{\partial^2 \ell}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T}$$

that is, use observed rather than expected information in the scoring step. This substitution does not usually materially affect the convergence of the scoring algorithm; for further discussion of this point, see Jørgensen (1983). The matrix W is diagonal, consisting of weights that arise by straightforward but tedious differentiation of (11).

The derivative matrices X and N are easily obtained for this model. The rows of X are just the vectors of explanatory variables $\mathbf{x}_i^T$, while it is clear from (10) that

$$N = UN^*,$$

where N^* is the incidence matrix describing the ordering and ties among $\{t_1, t_2, \ldots, t_n\}$, and

$$U = \mathrm{diag}(s_1, s_2, \ldots, s_n).$$

Fitting the model can proceed by the backfitting method. We alternate between use of the equations

$$\boldsymbol{\beta} = (X^T W X)^{-1} X^T W (\mathbf{z} - N\mathbf{g}),$$

and

$$N\mathbf{g} = S(\mathbf{z} - X\boldsymbol{\beta}),$$

where

$$S = N(N^T W N + \alpha K)^{-1} N^T W.$$

Since $N = UN^*$, we can write

$$S = U\{N^*({N^*}^T UWUN^* + \alpha K)^{-1} {N^*}^T UWU\} U^{-1}$$

so that the smoothing operation consists of weighted cubic spline smoothing allowing for ties and using weight matrix UWU, preceded and followed, respectively, by division and multiplication by U.

As an alternative to backfitting, there is a direct solution, not requiring nested iteration, by eliminating $\mathbf{g}$ from the updating equations to get the $p \times p$ system

$$X^T W(I - S) X\boldsymbol{\beta} = X^T W(I - S)\mathbf{z},$$

and solving this for $\boldsymbol{\beta}$. Note that this only requires application of the smoother S to the columns of X and to $\mathbf{z}$, not construction of the full smoothing matrix. Given the solution for $\boldsymbol{\beta}$, $N\mathbf{g}$ is obtained as above.

Various adaptations of this model could be contemplated. For example, we might wish to replace the additive predictor (10) by a function nonlinear in $\boldsymbol{\beta}$ and g, such as

$$\eta_i = \beta_1 x_i^{\beta_2} \beta_3^{(x_i^{\beta_4})} [1 - \exp\{-s_i g(t_i)\}]^{\beta_5} x_i^{\beta_6},$$

(a modification of a model that has been used to describe population growth). This would not change the algorithm, but merely mean that the matrices X and U would have more complicated forms, and values that changed with each iteration. Similarly, replacing the location family (9) by a more general dependence

$$P\{\tilde{Y}_i \leq y\} = \Psi(y; \eta_i)$$

would simply mean a different calculation for W and $\mathbf{z}$.

In all of these variations, the fixed point justification of the backfitting algorithm for maximum penalized likelihood estimation of $\boldsymbol{\beta}$ and g is retained. More care may be needed, however, in the more exotically nonlinear models, and it may also be necessary for convergence that the step size is adjusted, for example by taking a fixed convex combination of the old and proposed new values of each parameter at each update.

5 Nonparametric quantile regression

In medical practice, a useful device for screening patients is a table or chart recording the distribution of a biometrical measurement, such as height, weight or middle-upper-arm-circumference, for various values of an appropriate covariate, often age. A physician can then refer a measurement on a current patient to this chart, in order to assess, for example, whether a child is unusually short or tall for his age. Naturally, the chart has to be constructed from data from a relevant population.

In most cases of interest, the covariate concerned takes continuous values, and the distribution of the measurement can be assumed to vary smoothly with the covariate. Reference charts should therefore be constructed so as to respect this continuity, and without making use of arbitrary discretization or grouping. Further, although it is sometimes argued on general grounds that biometrical measurements are approximately normally distributed, it seems more appropriate not to make strong distributional assumptions, or to assume a particular parametric form for the dependence on the covariate.

5.1 The LMS method

A variety of methods is available for constructing reference curves for populations, not all of them meeting all the criteria set out above. A method that does meet them was proposed by Cole (1988). In what he terms the LMS method, it is assumed that for each value of the covariate t, the measurement Y, *after a Box-Cox power transformation*, is normally distributed. It is supposed that the expectation, variance, and the power of the transformation, all vary smoothly with t. Thus there is flexibility over the way that the location, spread and shape of the distribution of Y given t depend on t; it is natural, and in the spirit of this monograph, to allow this dependence to be nonparametric. Because of the flexibility afforded by these three varying characteristics, the normal assumption is not a particularly strong one.

The methodology described here was originally developed with the application to reference growth curve data very much in mind, but of course it may have applications in other areas where it is required to produce, as a function of a covariate, an estimate not just of the expected response, but of the whole distribution.

To define the LMS method, suppose that $\lambda(t)$, $\mu(t)$ and $\sigma(t)$ are functions of t, with $\mu(t)$ and $\sigma(t)$ supposed strictly positive. It is the initial letters of λ, μ and σ that provide the acronym LMS. If a measurement Y is made at t, denote by Z the transformed variate

$$Z = \begin{cases} \sigma(t)^{-1}\lambda(t)^{-1}[\{Y/\mu(t)\}^{\lambda(t)} - 1] & \text{if} \quad \lambda(t) \neq 0 \\ \sigma(t)^{-1}\log\{Y/\mu(t)\} & \text{if} \quad \lambda(t) = 0 \end{cases}$$

Then the conditional distribution of Z given t is assumed to be standard normal. If Φ denotes the normal distribution function, then the $100\alpha\%$ quantile of Y at t will be

$$\mu(t)\{1 + \lambda(t)\sigma(t)\Phi^{-1}(\alpha)\}^{1/\lambda(t)} \qquad \text{for} \qquad \lambda(t) \neq 0$$

or

$$\mu(t)\exp\{\sigma(t)\Phi^{-1}(\alpha)\} \qquad \text{for} \qquad \lambda(t) = 0.$$

Thus if the λ, μ and σ curves are smooth, so are all quantile curves. Cole's proposal is to use a plot of selected quantile curves as a summary of reference growth data.

5.2 Estimating the curves

There are various ways in which the λ, μ and σ curves could be estimated from data. Here, we describe the approach suggested by Cole and Green (1992), which is appropriate in the absence of assumptions about the forms of the curves other than that they are smooth. As described here, the method applies to data consisting of independent observations $(t_i, y_i), i = 1, 2, \ldots, n$, such as would arise in cross-sectional studies, and a modification to the likelihood would be needed to deal with the (more common) case of longitudinal data.

The penalized log-likelihood is

$$\begin{aligned} \Pi \quad &= \quad \sum_{i=1}^{n}\left(\lambda(t_i)\log\frac{Y_i}{\mu(t_i)} - \log\sigma(t_i) - \tfrac{1}{2}Z_i^2\right) \\ &\qquad - \tfrac{1}{2}\alpha_\lambda\int\lambda''(t)^2dt - \tfrac{1}{2}\alpha_\mu\int\mu''(t)^2dt - \tfrac{1}{2}\alpha_\sigma\int\sigma''(t)^2dt, \end{aligned}$$

where α_λ, α_μ and α_σ are smoothing parameters, and Z_i denotes Y_i transformed as above. A minor difficulty now is that there is a singularity in the likelihood as $\sigma(t) \to 0$ for all t. This gives an infinite value for Π. However, there is in our experience with the model always a local maximum of Π which is a stable fixed point of the algorithm about to be described, and to which the algorithm converges from 'sensible' initial estimates.

To locate a local maximum of Π, we now follow the familiar paradigm of using Fisher scoring. Let $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ denote the vectors of values of the corresponding curves at the distinct ordered values of $\{t_i\}$. Given $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ at the current iteration, updated estimates are given by the solutions to the equations

$$\begin{bmatrix} W_\lambda + \alpha_\lambda K & W_{\lambda\mu} & W_{\lambda\sigma} \\ W_{\mu\lambda} & W_\mu + \alpha_\mu K & W_{\mu\sigma} \\ W_{\sigma\lambda} & W_{\sigma\mu} & W_\sigma + \alpha_\sigma K \end{bmatrix} \begin{pmatrix} \boldsymbol{\lambda}^{new} - \boldsymbol{\lambda} \\ \boldsymbol{\mu}^{new} - \boldsymbol{\mu} \\ \boldsymbol{\sigma}^{new} - \boldsymbol{\sigma} \end{pmatrix} = \begin{pmatrix} \mathbf{u}_\lambda - \alpha_\lambda K\boldsymbol{\lambda} \\ \mathbf{u}_\mu - \alpha_\mu K\boldsymbol{\mu} \\ \mathbf{u}_\sigma - \alpha_\sigma K\boldsymbol{\sigma} \end{pmatrix}.$$

The $\mathbf{u}$'s and W's are first and expected second derivatives of the log likelihood with respect to the variables specified in the subscripts, and are sums of terms over observations at each distinct value of t. (In the case of W_λ, the expectation of the second derivative is replaced by that of the first few terms of a Taylor approximation, as derivatives of the log-likelihood with respect to $\boldsymbol{\lambda}$ do not have finite expectations.) This system of equations, while awkward to handle directly, is very amenable to backfitting: each of the three block rows of the system can be rearranged in the form of an updating equation for $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ respectively, which is an application of a weighted cubic spline smoother. Cole and Green find that this algorithm converges in typically four to eight iterations of the outer scoring loop. The iteration can be initialized by setting $\boldsymbol{\lambda}$ identically to 1, obtaining $\boldsymbol{\mu}$ by a simple smooth of $\mathbf{Y}$, and $\boldsymbol{\sigma}$ by smoothing the vector of values $\{Y_i/\mu(t_i) - 1\}^2$, and then taking square roots.

Further details of the methodology can be found in Cole and Green (1992), which includes a discussion of the relationship between the smoothing parameters and the corresponding equivalent degrees of freedom for each curve, and illustrations of the LMS method applied to data on triceps skinfold in Gambian females, and on body weight in U.S. girls.

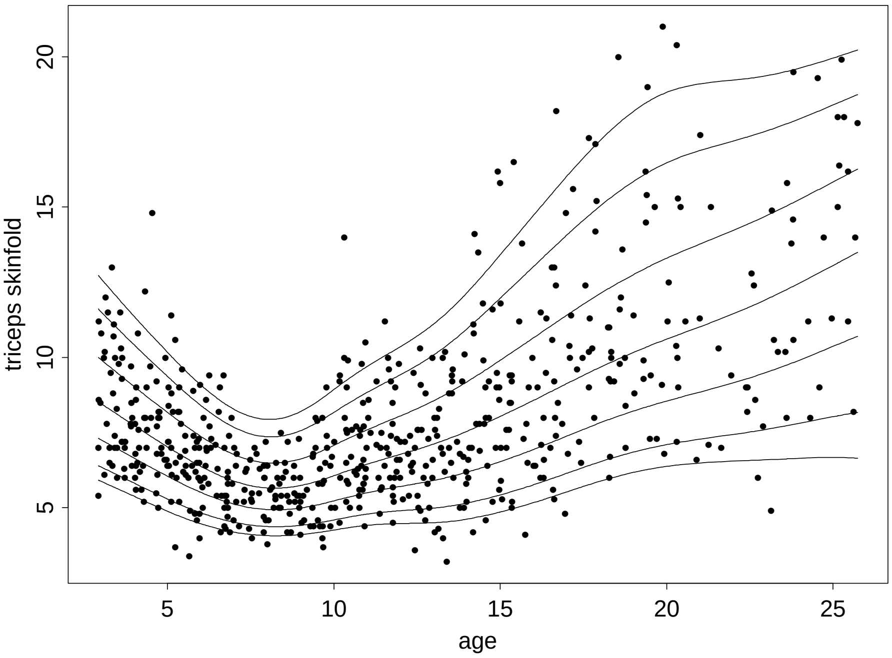

Figure 2: Smoothed reference centile curves for triceps skinfold among Gambian females: 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th} and 95^{th} percentiles.

Here, we provide some illustrations from the Gambian study. Figure 2 shows a scatter plot of triceps skinfold on age for Gambian females aged from about 3 to 26 years. The percentile curves shown are estimated by the method we have described, using smoothing parameters that allow 6 equivalent degrees of freedom for each curve. Note the 'notch' in the dependence at around 9 years. Figure 3 displays the fitted λ curve for these data, and superimposed the fitted σ curve. The distribution of triceps skinfold is apparently positively skewed for ages less than 20 years, with the skew decreasing abruptly between 20 and 25, although as can be seen in Figure 2, the data are somewhat sparse in this range. The relative spread, as measured by $\sigma(t)$, is fairly stable, but increases over the range 10 to 16 years.

References

Cole, T. J. (1988) Fitting smoothed centile curves to reference data (with discussion). *J. Roy. Statist. Soc.* A, **151**, 385–418.

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

Dinse, G. E. and Lagakos, S. W. (1983) Regression analysis of tumour prevalence data. *Applied Statistics*, **32**, 236–248.

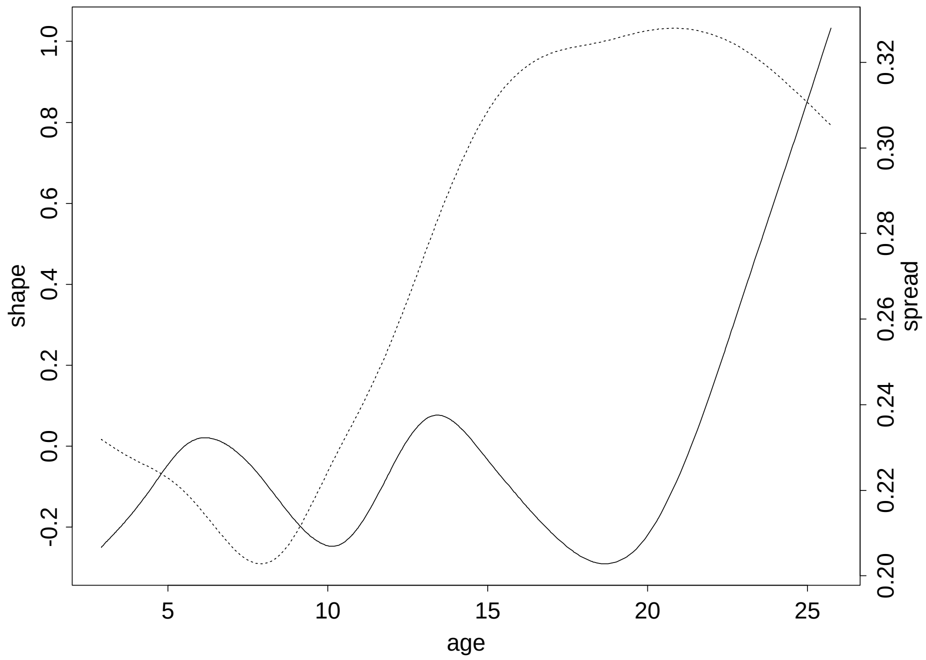

Figure 3: λ (shape, solid line) and σ (spread, broken line) curves for triceps skinfold among Gambian females.

Dinse, G. E. and Lagakos, S. W. (1984) Correction to Dinse and Lagakos (1983). *Applied Statistics*, **33**, 79–80.

Green, P. J. (1984) Iterated reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *J. Roy. Statist. Soc.* B, **46**, 149–192.

Green, P. J. (1987) Penalized likelihood for generalized semi-parametric regression models. *Int. Statist. Rev.*, **55**, 245–259.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric regression and generalized linear models: a roughness penalty approach.* London: Chapman and Hall.

Hastie, T. J. and Tibshirani, R. J. (1993) Varying-coefficient models (with discussion). *J. Roy. Statist. Soc.* B, **55**, 757–796.

Jørgensen, B. (1983) Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika*, **70**, 19–28.

Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. Roy. Statist. Soc.* A, **135**, 370–384.

Scallan, A., Gilchrist, R. and Green, M. (1984) Fitting parametric link functions in generalised linear models. *Comp. Statist. and Data Anal.*, **2**, 37–49.