Smoothing and mixed models

M. P. Wand

Department of Biostatistics, School of Public Health, Harvard University, 665 Huntington Avenue, Boston, MA 02115, USA

Summary

Smoothing methods that use basis functions with penalisation can be formulated as maximum likelihood estimators and best predictors in a mixed model framework. Such connections are at least a quarter of a century old but, perhaps with the advent of mixed model software, have led to a paradigm shift in the field of smoothing. The reason is that most, perhaps all, models involving smoothing can be expressed as a mixed model and hence enjoy the benefit of the growing body of methodology and software for general mixed model analysis. The handling of other complications such as clustering, missing data and measurement error is generally quite straightforward with mixed model representations of smoothing.

Keywords: Best prediction, Generalised linear mixed models, Nonparametric regression, Kriging, Maximum likelihood, Variance components, Restricted maximum likelihood.

1 Introduction

If you ask a randomly chosen statistician the first thing that comes to mind when he or she hears the phrase 'mixed models', then a likely response would be that they are a vehicle for analysis of data such as those depicted in Figure 1. These data are weight measurements on 48 pigs, for 9 successive weeks and are an example of *longitudinal* data (source: Diggle, Liang and Zeger, 1995). Mixed models have been prominent in longitudinal data analysis going back

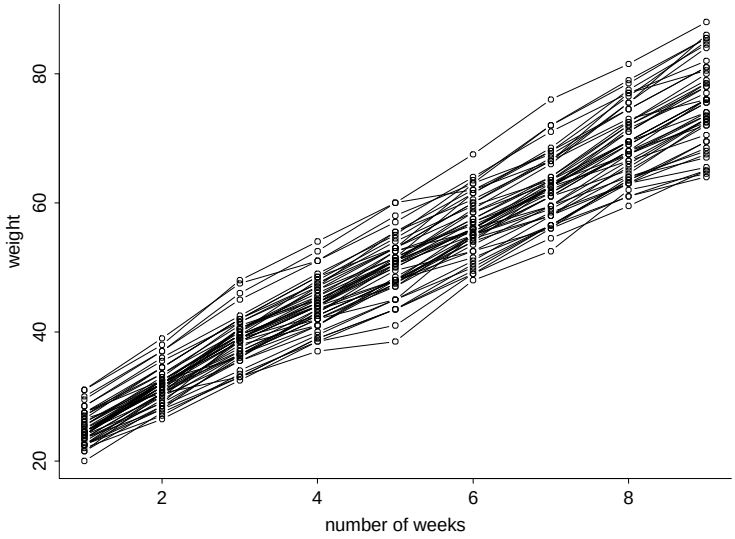

Figure 1: Pig weight data. Lines join repeated measurements on the same pig.

to the seminal paper of Laird and Ware (1982). A reasonable model for these data is a straight line with random intercept model:

$$\texttt{weight}_{ij} = \beta_0 + U_i + \beta_1 \texttt{week}_j + \varepsilon_{ij}, \quad 1 \leq i \leq 48, \quad 1 \leq j \leq 9, \qquad (1.1)$$

where $U_1, \ldots, U_{48} \overset{\text{ind.}}{\sim} N(0, \sigma_U^2)$ are the random intercepts and $\varepsilon_{ij} \overset{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2)$. Note that the U_i induce correlation among the measurements within each pig. Since the model contains both fixed effects (β_0 and β_1) and random effects (the U_i) the label *mixed model* applies. Similar models are used in other situations where there is grouping structure in the data. For example, 'small area' disease maps can be smoothed using Poisson mixed models (e.g. Breslow and Clayton, 1993).

However, mixed models have much wider generality than those used for handling grouping structure. Figure 2 shows a scatterplot corresponding to some fossil data, as described in Chaudhuri and Marron (1999). The smooth of the scatterplot corresponds to a mixed model fit. The amount of smoothing was also chosen automatically from the data using standard mixed model methodology.

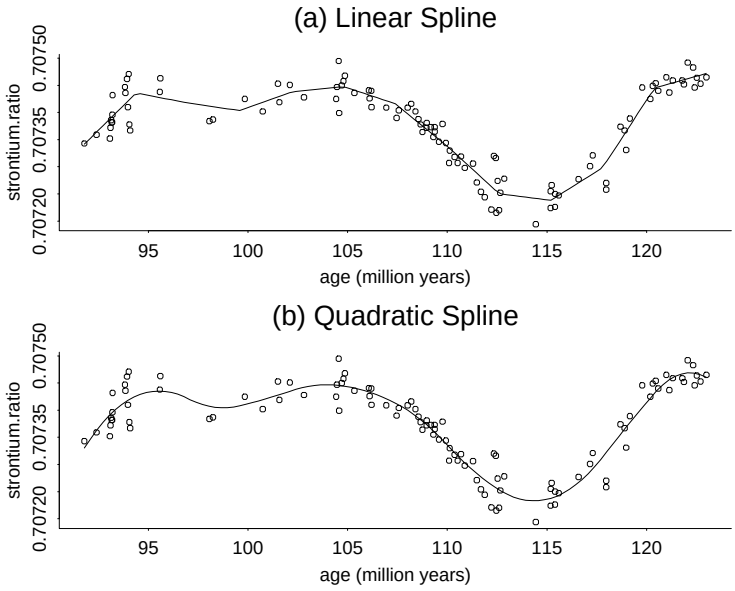

Figure 2: Scatterplot corresponding to fossil data, with automatic scatterplot smooth added.

Figure 2 is an example of nonparametric regression, although the terms *scatterplot smoothing*, or simply *smoothing* are often used instead. In this paper we use the term 'smoothing' as follows: use of regression models that contain at least one function being modelled nonparametrically. Such models might also be called *semiparametric* regression models.

Speed (1991) is the earliest reference known to me that explicitly makes the connection between nonparametric regression and mixed models. Earlier work by Wahba (1978), Wecker and Ansley (1983) and Green (1985) make connections of this type, but do not mention mixed models (although 'random effects model' is mentioned on page 17 of Wahba (1990) when discussing Wahba (1978)). The late 1990s saw a surge of research in mixed model-based smoothing, partly driven by the availability of mixed model software in packages such as `SAS` (SAS Institute Inc., 2002) and `S-PLUS` (MathSoft Inc., 2002). References include Verbyla (1994), O'Connell and Wolfinger (1997), Wang (1998a,1998b), Brumback and Rice (1998), Verbyla, Cullis, Kenward and Welham (1999), Lin and Zhang (1999) and Brumback, Ruppert and Wand (1999). Another noteworthy reference is Diggle (1997) in which the commonalities between longitudinal data analysis and spatial statistics are observed. The current paper argues that smoothing can be added to this list.

Section 2 gives an overview of general design mixed models and Section 3 illustrates their richness from a smoothing standpoint. Generalized responses

are considered in Section 4. Some departures from the 'clean data' situation are discussed in Section 5, where it is pointed out that the mixed model approach to smoothing has several advantages. Inference, model selection and diagnostics are briefly mentioned in Section 6; as is Bayesian methodology in Section 7. Pointers to software are given in Section 8. I conclude with some remarks about the minimalisation of methodology that comes out of the mixed model approach to smoothing in Section 9.

2 General Design Mixed Models

The majority of mixed models in common use have the generic form

$$\mathsf{E}(\mathbf{y}|\mathbf{u}) = g(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}), \quad \mathbf{u} \sim (\mathbf{0}, \mathbf{G})$$

for response vector $\mathbf{y}$, design matrices $\mathbf{X}$ and $\mathbf{Z}$, fixed effects vector $\boldsymbol{\beta}$ and random effects vector $\mathbf{u}$. Here g is a scalar 'link' function, and evaluated element-wise for vector arguments. For a general random vector $\mathbf{v}$, $\mathbf{v} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is shorthand for $\mathsf{E}(\mathbf{v}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{v}) = \boldsymbol{\Sigma}$, the covariance matrix of $\mathbf{v}$. Since $\mathbf{X}$ and $\mathbf{Z}$ are general matrices (with conforming dimension) I use the phrase *general design*.

An important special case is the linear mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right). \tag{2.1}$$

Robinson (1991) and the ensuing discussion provides an excellent overview of general design linear mixed models. A special case of (2.1) is the Gaussian mixed model, for which

$$\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right).$$

While estimation of $\boldsymbol{\beta}$ can be done without the Gaussian assumption using best linear unbiased prediction, I will confine discussion in this section and the next to the Gaussian case. One payoff from the Gaussian assumption is the likelihood-based estimation of the covariance matrices $\mathbf{G}$ and $\mathbf{R}$. The log-likelihood of $(\boldsymbol{\beta}, \mathbf{G}, \mathbf{R})$ under the Gaussian model is

$$\ell(\boldsymbol{\beta}, \mathbf{G}, \mathbf{R}) = -\tfrac{1}{2}\left\{ n\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \tag{2.2}$$

where

$$\mathbf{V} \equiv \mathrm{Cov}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathsf{T}} + \mathbf{R}.$$

Therefore, the maximum likelihood (ML) estimate of $(\boldsymbol{\beta}, \mathbf{G}, \mathbf{R})$ is the one that maximizes the right-hand side of this expression. If one first optimizes over $\boldsymbol{\beta}$, which appears only in the last term, one obtains

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{y}$$

for all $\mathbf{V}$. On substitution into (2.2) one obtains the *profile log-likelihood* for $\mathbf{V}$:

$$\begin{aligned} \ell_P(\mathbf{V}) &= -\tfrac{1}{2}\left\{\log|\mathbf{V}| + (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^{\mathsf{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) + n\log(2\pi)\right\} \\ &= -\tfrac{1}{2}\left[\log|\mathbf{V}| + \mathbf{y}^{\mathsf{T}}\mathbf{V}^{-1}\{\mathbf{I}-\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\}\mathbf{y}\right] \\ &\quad -\tfrac{n}{2}\log(2\pi). \end{aligned} \tag{2.3}$$

Maximum likelihood estimates of $\mathbf{G}$ and $\mathbf{R}$ can be found by maximizing (2.3).

An alternative to ML for estimation of the covariance matrices $\mathbf{G}$ and $\mathbf{R}$ is Restricted ML (REML) (Patterson and Thompson, 1973). Its derivation is more complicated and involves maximizing the likelihood of linear combinations of the elements of $\mathbf{y}$ that do not depend on $\boldsymbol{\beta}$. Details can be found in, for example, Chapter 6 of Searle, Casella and McCulloch (1992). The resulting criterion function is the *restricted log-likelihood*

$$\ell_R(\mathbf{V}) = \ell_P(\mathbf{V}) - \tfrac{1}{2}\log|\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X}|. \tag{2.4}$$

For known $\boldsymbol{\beta}$, $\mathbf{G}$ and $\mathbf{R}$ the random effects vector can be predicted using *best prediction* (BP). This entails determination of the $\widehat{\mathbf{u}}$ that minimises

$$\mathsf{E}\{(\widehat{\mathbf{u}}-\mathbf{u})^{\mathsf{T}}(\widehat{\mathbf{u}}-\mathbf{u})\}$$

for which the solution is the best predictor

$$\widehat{\mathbf{u}} = \mathsf{E}(\mathbf{u}|\mathbf{y}).$$

For the Gaussian mixed model

$$\widehat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^{\mathsf{T}}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

In practice, $\widehat{\mathbf{u}}$ will usually depend on estimates of $\boldsymbol{\beta}$, $\mathbf{G}$ and $\mathbf{R}$ in which case the term *estimated* best predictor is sometimes used. For the sake of simplicity, I will not discriminate between the two in this article.

2.1 Two-point summary of fitting for mixed models

Typical mixed model literature can be somewhat daunting. Reasons include complex subscript notation for handling special cases and focus on computational issues. However, when the mixed model is considered in full generality and computational issues are left aside then fitting boils down to just two fundamental principles: maximum likelihood and best prediction. Specifically, the fitting of mixed models involves:

- Estimation of fixed effects $\boldsymbol{\beta}$ and covariance matrices $\mathbf{G}$ and $\mathbf{R}$ using (restricted) maximum likelihood ((RE)ML).
- Prediction of random effects $\mathbf{u}$ using best prediction (BP); $\widehat{\mathbf{u}} = \mathsf{E}(\mathbf{u}|\mathbf{y})$.

Figure 3 displays this two-point summary in a single graphic.

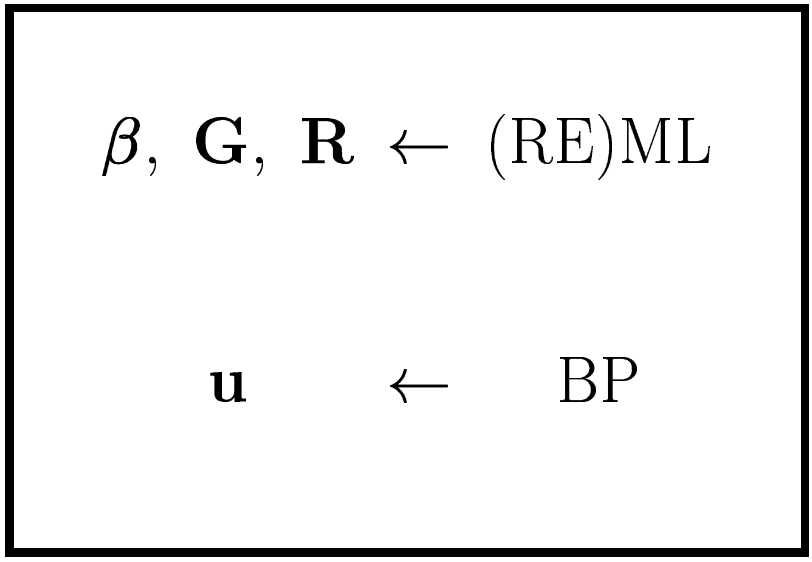

Figure 3: Two-Point summary of fitting mixed models.

3 The Many Uses of Mixed Models

3.1 Random intercept models

As I mentioned at the start of this paper, probably the most common use of mixed models is to handle within subject correlation in linear regression models through the employment of a random intercept. Model (1.1) is then a special case of (2.1) with

$$\mathbf{y} = \begin{bmatrix} \texttt{weight}_{1,1} \\ \vdots \\ \texttt{weight}_{1,9} \\ \vdots \\ \texttt{weight}_{48,1} \\ \vdots \\ \texttt{weight}_{48,9} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & \texttt{week}_1 \\ \vdots & \vdots \\ 1 & \texttt{week}_9 \\ \vdots & \vdots \\ 1 & \texttt{week}_1 \\ \vdots & \vdots \\ 1 & \texttt{week}_9 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}, \quad (3.1)$$

$\boldsymbol{\beta} = [\beta_0 \ \beta_1]^\mathsf{T}$, $\mathbf{u} = [U_1, \ldots, U_{48}]^\mathsf{T}$, $\mathbf{G} = \sigma_U^2\mathbf{I}$ and $\mathbf{R} = \sigma_\varepsilon^2\mathbf{I}$.

Unfortunately, the majority of the literature and some software packages on mixed models are geared towards to the analysis of repeated measures data sets such as the one in this example. In particular, they cater only for the

random effects design matrix $\mathbf{Z}$ having special forms like the one appearing in (3.1). As I will show in the remainder of this section, there are several important mixed models that fall outside of this format.

3.2 Scatterplot smoothing

The means by which mixed models can be tricked into doing smoothing can be explained in just a few lines of mathematics. Figure 4 may also be helpful for understanding. The data in each panel of this figure is the same, and was generated as

$$y_i = f(x_i) + 0.4\varepsilon_i$$

where the x_i and ε_i are random samples from the uniform distribution on $(0,1)$ and the standard normal distribution respectively. Although the x_i's are generated by a random process I will consider them fixed in this discussion. The mean function f is $f(x) = \sin(3\pi x)$. In both panels linear models of the form

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+ + \varepsilon_i \tag{3.2}$$

have been fit to the data. The function

$$(x - \kappa_k)_+ = \begin{cases} 0, & x \leq \kappa_k \\ x - \kappa_k, & x > \kappa_k \end{cases}$$

represents a 'broken' line with a join-point, or *knot*, at κ_k. The choice of the κ_k's is discussed in Section 3.2.3.

The bar at the base of each panel shows the truncated line basis functions $(x - \kappa_k)_+$, $1 \leq k \leq K$. Panel (a) is just an ordinary least squares fit to the scatterplot; but is quite rough due to the large number of truncated line functions being fit. Panel (b) remedies this through one simple modification:

$$u_k \overset{\text{ind.}}{\sim} N(0, \sigma_u^2). \tag{3.3}$$

For $\sigma_u^2 < \infty$ (3.3) imposes a restriction on the restriction u_k: they are no longer allowed to range between $-\infty$ and ∞ as they please. Instead they must obey the laws of a normal probability distribution with zero mean and finite variance. This tends to shrink the u_k and leads to the smooth fit shown in Figure 4(b).

If one defines the design matrices

$$\mathbf{X} = [1 \; x_i]_{1 \leq i \leq n}, \qquad \mathbf{Z} = \underset{1 \leq k \leq K}{[(x_i - \kappa_k)_+]_{1 \leq i \leq n}}$$

and set $\boldsymbol{\beta} = [\beta_0, \beta_1]^\mathsf{T}$, $\mathbf{u} = [u_1, \ldots, u_K]^\mathsf{T}$ then one can rewrite (3.2) and (3.3) as the linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \mathbf{I} \end{bmatrix} \right).$$

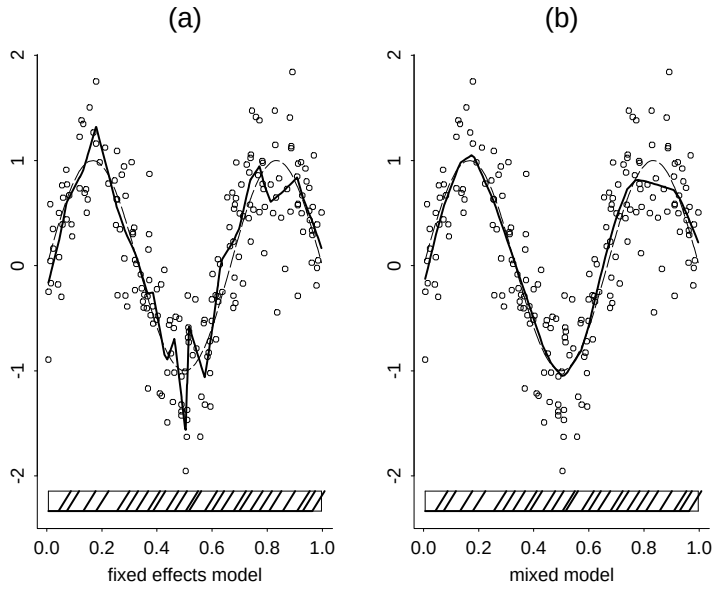

Figure 4: How mixed models do smoothing. In (a) all coefficients are fixed effects, while in (b) the coefficients of the knots are random effects. The solid curve is the estimated curve, while the dashed curve is the function from which the data were generated.

The scatterplot smooth can then be obtained by applying (restricted) maximum likelihood to $\boldsymbol{\beta}$, σ_u^2 and σ_ε^2 and best prediction to $\mathbf{u}$.

3.2.1 Connection with penalized regression

For given values of σ_u^2 and σ_ε^2 application of ML and BP to obtain $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ is equivalent to solving the penalized least squares problem

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{bmatrix} = \underset{\boldsymbol{\beta},\mathbf{u}}{\operatorname{argmin}}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 + \alpha\|\mathbf{u}\|^2) \tag{3.4}$$

where $\alpha \equiv \sigma_\varepsilon^2/\sigma_u^2$ and, for a general vector $\mathbf{v}$, $\|\mathbf{v}\| \equiv \sqrt{\mathbf{v}^\mathsf{T}\mathbf{v}}$ (e.g. Robinson, 1991). This is an example of *penalized least squares* (e.g. Green, 1987) since minimisation of the least squares $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2$ is subject to the penalty $\alpha\|\mathbf{u}\|^2$ being imposed on the coefficients in $\mathbf{u}$. The solution is easily shown to be

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{bmatrix} = (\mathbf{C}^\mathsf{T}\mathbf{C} + \alpha\mathbf{D})^{-1}\mathbf{C}^\mathsf{T}\mathbf{y}$$

where $\mathbf{C} = [\mathbf{X}\ \mathbf{Z}]$ and $\mathbf{D} = \operatorname{diag}(0, 0, 1, \ldots, 1)$.

The penalty in (3.4) is quadratic in $\mathbf{u}$:

$$\alpha\|\mathbf{u}\|^2 = \alpha \sum_{k=1}^{K} u_k^2$$

and arises from the normality condition (3.3). Different penalties can be obtained by modifying the distribution of the u_k's. For example, the condition

$$u_k \overset{\text{ind.}}{\sim} \text{Laplace}(0, \sigma_u^2),$$

where $X \sim \text{Laplace}(\mu, \sigma^2)$ if and only if its density function is $f_X(x) = \frac{1}{2}\sigma^{-1}\exp\{-|x-\mu|/\sigma\}$, gives rise to the sum of absolute values penalty

$$\alpha \sum_{k=1}^{K} |u_k|, \quad \alpha \equiv \sigma_\varepsilon^2/\sigma_u.$$

This penalty is called the *lasso* (*least absolute shrinkage and selection operator*) in Tibshirani (1996).

As shown in Antoniadis and Fan (2001) there are also equivalences between penalized regression and *thresholding* as commonly used in wavelet regression.

Scatterplot smoothers where the number of basis functions is less than the sample size, presented in this and the previous section, go back at least to Parker and Rice (1985), O'Sullivan (1986,1988), Gray (1992) and Kelly and Rice (1990). More recent references are Eilers and Marx (1996), Hastie (1996) and Ruppert and Carroll (2000) where the following names:

- P-splines,
- penalised splines,
- pseudosplines, and
- low-rank smoothers

have been coined. Since each of these are virtually synonymous, I will use penalised splines in the rest of this paper.

Penalised splines typically use $K << n$ knots with little degradation in the quality of the fit. This leads to a number of computational payoffs: faster scatterplot smoothing for very large n, direct fitting of additive models and exact computation of auxiliary quantities such as degrees of freedom measures and standard errors (e.g. Hastie, 1996). As pointed out in French, Kammann and Wand (2001), even the `S-PLUS` smoothing spline function `smooth.spline()` uses a rule similar to (3.6) for $n > 50$.

3.2.2 Other spline bases

I like the truncated line basis because it is very easy to see how the curve fitting is being done. Through pictures like Figure 4 one can explain smoothing to someone with only a cursory knowledge of regression. The simple mathematical form of truncated lines is also useful when formulating more complicated models. Smoother fits can be achieved using higher degree truncated polynomial bases. For example, the truncated cubic basis for the knots $\kappa_1, \ldots, \kappa_K$ is

$$1, x, x^2, x^3, \{(x-\kappa_1)_+\}^3, \ldots, \{(x-\kappa_K)_+\}^3.$$

Many people criticise the truncated polynomial bases because of their suboptimal numerical properties, but I prefer to leave numerical details out of model formulation. This is in keeping with elementary regression textbooks (e.g. Draper and Smith, 1998) which present expressions like

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} \tag{3.5}$$

even though direct use of (3.5) not recommended for actual implementation; QR decomposition is recommended instead. Bases with better numerical properties, such as B-spline and Demmler-Reinsch bases, are equivalent from a mathematical standpoint. However, in my view, their use in model formulation causes unnecessary complication (especially for the uninitiated) and numerical issues are best left aside as an implementational detail. I do exactly this in the current paper by including an appendix that explains how to fit penalised splines in a stable and efficient manner. Truncated polynomial penalised spline basis functions could be used as inputs, but the algorithm transforms them to a Demmler-Reinsch basis for actual fitting.

An alternative to the B-spline-type bases is radial bases; which are useful for higher dimensional extension. Details are given in Section 3.6. Penalised splines with cubic radial basis functions are a low-rank generalisation of smoothing splines (see French, Kammann and Wand, 2001).

3.2.3 Knot specification

Based on about five years of practical experience, and some unpublished simulation studies, I have found knot specification to be very much a minor detail for penalised splines. French, Kammann and Wand (2001) provide a small amount of evidence to support this claim, as does Ruppert (2002). Indeed, I have found that simple rules such as

$$\kappa_k = \left(\tfrac{k+1}{K+2}\right) \text{th sample quantile of unique } x_i\text{'s}, \quad 1 \le k \le K, \tag{3.6}$$

with $K = \min(n/4, 35)$ work quite well. Occasionally, some manual adjustment is in order. See Ruppert (2002) for empirical justification. Some writers

(e.g. Ke and Wang, 2001) have criticised having to specify the knots, but I think that it is small price to pay for the computational advantages that penalised splines offer (Section 3.2.1).

3.3 Additive models

The simplest example of additive model is

$$y_i = \beta_0 + f(s_i) + g(t_i) + \varepsilon_i \tag{3.7}$$

where (s_i, t_i) represent measurements on two continuous predictors s_i and t_i. Hastie and Tibshirani (1990) provide a thorough coverage of such models. Software for their fitting is provided through the function `gam()` in `S-PLUS` and `PROC GAM` in `SAS`. However, (3.7) may also be fit using mixed model software by expressing it as the mixed model

$$y_i = \beta_0 + \beta_s s_i + \sum_{k=1}^{K_s} u_k^s (s_i - \kappa_k^s)_+ + \beta_t t_i + \sum_{k=1}^{K_t} u_k^t (t_i - \kappa_k^t)_+ + \varepsilon_i$$

with

$$u_k^s \overset{\text{ind.}}{\sim} N(0, \sigma_s^2) \quad \text{and} \quad u_k^t \overset{\text{ind.}}{\sim} N(0, \sigma_t^2),$$

independently of one another. Setting up

$$\boldsymbol{\beta} = [\beta_0, \beta_s, \beta_t]^\mathsf{T}, \quad \mathbf{u} = [u_1^s, \ldots, u_{K_s}^s, u_1^t, \ldots, u_{K_t}^t]^\mathsf{T},$$

$$\mathbf{X} = [1 \ s_i \ t_i]_{1 \le i \le n}, \quad \text{and} \quad \mathbf{Z} = [\underset{1 \le k \le K^s}{(s_i - \kappa_k^s)_+}, \underset{1 \le k \le K^t}{(t_i - \kappa_k^t)_+}]_{1 \le i \le n},$$

the model can be expressed in the standard form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \mathsf{E}\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \mathbf{0}, \quad \text{Cov}\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_s^2 \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_t^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_\varepsilon^2 \mathbf{I} \end{bmatrix}.$$

3.4 Additive models with interactions

An extension of (3.7) that has been required in some of our applications is

$$y_i = \beta_0 + f_{z_i}(s_i) + g(t_i) + \varepsilon_i \tag{3.8}$$

where z_i is a categorical variable. This is no longer an additive model since the term $f_{z_i}(s_i)$ corresponds to an interaction between z and s. Such models have been studied by, for example, Wahba (1986) and Chen (1993). Coull, Ruppert and Wand (2001) describe how to treat (3.8) as a mixed model.

3.5 Varying coefficient models

Let (y_i, x_i, s_i) represent measurements on three variables, y, x and s. Then a *varying-coefficient model* for these data is

$$y_i = \alpha(s_i) + \beta(s_i)x_i + \varepsilon_i \tag{3.9}$$

(Hastie and Tibshirani, 1993). The interpretation is that, for any fixed value of the variable s, there is a linear relationship between y and x, but the coefficients are smooth functions of the s variable. The penalized spline version of this model is

$$y_i = \alpha_0 + \alpha_1 s_i + \sum_{k=1}^{K} u_k^{\alpha}(s_i - \kappa_k)_+ + \left\{\beta_0 + \beta_1 s_i + \sum_{k=1}^{K} u_k^{\beta}(s_i - \kappa_k)_+\right\} x_i + \varepsilon_i.$$

where $\kappa_1, \ldots, \kappa_K$ are knots over the range of the s_i values. A mixed model representation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$ is obtained by setting

$$\mathbf{X} = [1 \quad s_i \quad x_i \quad s_i x_i]_{1\le i\le n}, \qquad \mathbf{Z} = [\underset{1\le k\le K}{(s_i - \kappa_k)_+} \quad \underset{1\le k\le K}{x_i(s_i - \kappa_k)_+}]_{1\le i\le n,},$$

$\boldsymbol{\beta} = [\alpha_0 \ \alpha_1 \ \beta_0 \ \beta_1]^{\mathsf{T}}$, $\mathbf{u} = [u_1^{\alpha}, \ldots, u_K^{\alpha}, u_1^{\beta}, \ldots, u_K^{\beta}]^{\mathsf{T}}$ and

$$\text{Cov}(\mathbf{u}) = \begin{bmatrix} \sigma_{\alpha}^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_{\beta}^2 \mathbf{I} \end{bmatrix}.$$

3.6 Multivariate smoothing

Penalised spline regression can be extended to the bivariate situation in at least two ways: taking products of one-dimensional splines or working with radial basis functions which depend only on the distances between the data and the knots. I have a preference for the latter; and will start with univariate radial smoothing. The multivariate extension follows straightforwardly.

3.6.1 Univariate radial smoothers

Penalised spline smoothers with radial bases, or *radial smoothers*, and their relationship to smoothing/thin plate splines and kriging are summarised in French, Kammann and Wand (2001). For $x_i \in \mathbb{R}$ a useful class of low-rank radial smoothers is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{\mathrm{K}}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \text{Cov}(\mathbf{u}) = \sigma_u^2(\boldsymbol{\Omega}_{\mathrm{K}}^{-1/2})(\boldsymbol{\Omega}_{\mathrm{K}}^{-1/2})^{\mathsf{T}}$$

where $\mathbf{X} = [1 \ x_i \ \ldots \ x_i^{m-1}]_{1\le i\le n}$,

$$\mathbf{Z}_{\mathrm{K}} = [\underset{1\le k\le K}{|x_i - \kappa_k|^{2m-1}}]_{1\le i\le n} \quad \text{and} \quad \boldsymbol{\Omega}_{\mathrm{K}} = [\underset{1\le k,k'\le K}{|\kappa_k - \kappa_{k'}|^{2m-1}}].$$

Using the transformation $\mathbf{Z} = \mathbf{Z}_K \mathbf{\Omega}_K^{-1/2}$ the model can be rewritten as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \text{Cov} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_u^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \mathbf{I} \end{bmatrix}. \tag{3.10}$$

This form allows fitting through standard mixed model software (see Section 3.2.3).

Note that

$$C(r) = (-1)^m |r|^{2m-1}$$

is a so-called *generalized covariance function* and could be replaced by any of the proper covariance functions used in kriging (e.g. Cressie 1993; O'Connell & Wolfinger 1997; Stein 1999).

3.6.2 Extension to higher dimensions

For $\mathbf{x}_i \in \mathbb{R}^d$, $1 \le i \le n$, and $\boldsymbol{\kappa}_k \in \mathbb{R}^d$, $1 \le k \le K$, then higher dimension approximate smoothing splines (also called *thin plate* splines) with smoothness parameter m can be obtained by taking $\mathbf{X}$ to have columns spanning the space of all d-dimensional polynomials in the components of $\mathbf{x}_i$ with degree less than m and

$$\mathbf{Z} = [C(\|\mathbf{x}_i - \boldsymbol{\kappa}_k\|)]_{\substack{1 \le i \le n \\ 1 \le k \le K}} [C(\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}\|)]^{-1/2}_{1 \le k, k' \le K}$$

where

$$C(\mathbf{r}) = \begin{cases} \|\mathbf{r}\|^{2m-d} & d \text{ odd} \\ \|\mathbf{r}\|^{2m-d} \log \|\mathbf{r}\| & d \text{ even} \end{cases}$$

(e.g. Nychka, 2000).

Once again, $C(\cdot)$ could also be a covariance function such as those used in kriging (e.g. Cressie 1993; O'Connell & Wolfinger 1997; Stein 1999).

The choice of the bivariate knots $\boldsymbol{\kappa}_k$, $1 \le k \le K$, is somewhat more challenging. I have had good experience with knots chosen via an efficient space filling algorithm (e.g. Johnson, Moore and Ylvisaker, 1990; Nychka and Saltzman, 1998). The `S-PLUS` module `FUNFITS` (Nychka, Haaland, O'Connell and Ellner, 1998) supports space filling algorithms. Figure 5 shows the result of applying such an algorithm to the (jittered) locations in the example used by Kammann and Wand (2003) for $d = 2$.

3.7 Combinations

One of the advantages of handling smoothing through mixed models is the seamlessness with which other complications can be handled. Figure 6 provides an illustration. It shows a data set similar in nature to the pig weight data, corresponding to spinal bone mineral density (SBMD) measurements of 230 female subjects aged between 8 and 27. Each subject is measured

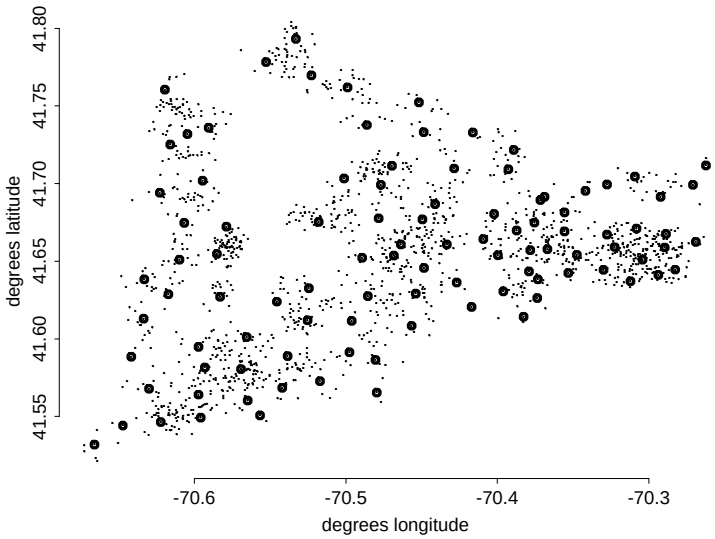

Figure 5: The smaller dots correspond to the geographical locations in the Upper Cape Cod reproductive data, with jittering to protect identity. The larger dots correspond to a representative subset of 100 locations for performing radial penalized spline smoothing.

1,2,3 or 4 times. Hastie and Tibshirani (2000) provides a thorough semiparametric analysis of these data for the subjects with 2 or more measurements. Other analyses are given in James, Hastie and Sugar (2000) and James and Hastie (2001). The main difference between the data presented in Figure 6 and those in Figure 1 is that the latter exhibit a high degree of nonlinearity.

An appropriate model is

$$\texttt{SBMD}_{ij} = U_i + f(\texttt{age}_{ij}) + \varepsilon_{ij}, \quad 1 \le j \le n_i, \quad 1 \le i \le 230. \tag{3.11}$$

Here $\texttt{age}_{ij}$ and $\texttt{SBMD}_{ij}$ are the jth measurements, $1 \le j \le n_i$, of age and spinal bone mineral density respectively for on subject i, f is some smooth function, U_i is a random intercept and ε_{ij} are random errors. One can do smoothing and have a random intercept through the mixed model

$$\texttt{SBMD}_{ij} = U_i + \beta_0 + \beta_1 \texttt{age}_{ij} + \sum_{k=1}^{K} u_k (\texttt{age}_{ij} - \kappa_k)_+ + \varepsilon_{ij},$$

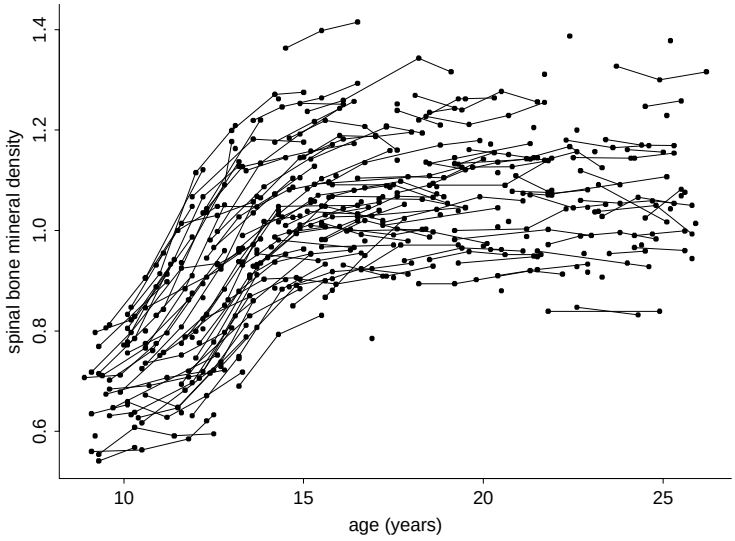

Figure 6: Spinal bone mineral density data.

where $U_i \overset{\text{ind.}}{\sim} N(0, \sigma_U^2)$ and $u_k \overset{\text{ind.}}{\sim} N(0, \sigma_u^2)$. If one defines

$$\mathbf{X} = \begin{bmatrix} 1 & \mathtt{age}_{11} \\ \vdots & \vdots \\ 1 & \mathtt{age}_{1n_1} \\ \vdots & \vdots \\ 1 & \mathtt{age}_{m1} \\ \vdots & \vdots \\ 1 & \mathtt{age}_{mn_m} \end{bmatrix},$$

$$\mathbf{Z} = \begin{bmatrix} 1 & \cdots & 0 & (\mathtt{age}_{11} - \kappa_1)_+ & \cdots & (\mathtt{age}_{11} - \kappa_K)_+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 0 & (\mathtt{age}_{1n_1} - \kappa_1)_+ & \cdots & (\mathtt{age}_{1n_1} - \kappa_K)_+ \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & (\mathtt{age}_{m1} - \kappa_1)_+ & \cdots & (\mathtt{age}_{m1} - \kappa_K)_+ \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & (\mathtt{age}_{mn_m} - \kappa_1)_+ & \cdots & (\mathtt{age}_{mn_m} - \kappa_K)_+ \end{bmatrix}$$

$\boldsymbol{\beta} = [\beta_0, \beta_1]^\mathsf{T}$ and $\mathbf{u} = [U_1, \ldots, U_m, u_1, \ldots, u_K]^\mathsf{T}$ then one can simultaneously estimate variance components for the random intercept and the amount of

smoothing for f through the mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \text{Cov}\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_U^2\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_\varepsilon^2\mathbf{I} \end{bmatrix}. \tag{3.12}$$

Here σ_U^2 measures the between subject variation, σ_ε^2 measures within subject variation and σ_u^2 controls the amount of smoothing done to estimate f.

A similar such combination is performed by Kammann and Wand (2003). There geostatistical analyses, similar to kriging, is combined with additive models to account for non-linear effects. They name the resulting models *geoadditive* models.

4 Generalized Responses

The extension to generalized responses, such as binary and count variables, entails *generalized* mixed models. The most common is the *generalized linear mixed model*, corresponding to the one-parameter exponential family and Gaussian random effects, for which

$$f(\mathbf{y}|\mathbf{u}) = \exp\{\mathbf{y}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^\mathsf{T}b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) + \mathbf{1}^\mathsf{T}c(\mathbf{y})\}$$

is the density of $\mathbf{y}$ given $\mathbf{u}$ and

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$$

is a $q \times 1$ random effects vector. The logistic-normal mixed model corresponds to

$$b(x) = \log(1 + e^x), \quad c(x) = 0$$

while the Poisson-normal mixed model corresponds to

$$b(x) = e^x, \quad c(x) = -\log(x!).$$

McCulloch and Searle (2000) provides an excellent overview of generalized linear mixed models.

In theory, generalized linear mixed models can be fit using Figure 3:

$$\begin{aligned} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{G}} \end{bmatrix} &= \text{maximum of likelihood function for } (\boldsymbol{\beta}, \mathbf{G}) \\ &= \underset{\boldsymbol{\beta},\mathbf{G}}{\text{argmax}}\, f(\mathbf{y}; \boldsymbol{\beta}, \mathbf{G}) \\ &= \underset{\boldsymbol{\beta},\mathbf{G}}{\text{argmax}} (2\pi)^{-q/2}|\mathbf{G}|^{-1/2} \exp\{\mathbf{1}^\mathsf{T}c(\mathbf{y})\} \\ &\quad \times \int_{\mathbb{R}^q} \exp\{\mathbf{y}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^\mathsf{T}b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \tfrac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{G}^{-1}\mathbf{u}\}\, d\mathbf{u} \end{aligned}$$

and

$$\begin{aligned}\widehat{\mathbf{u}} &= \text{best predictor of } \mathbf{u} \text{ given } \mathbf{y} \\ &= \mathsf{E}(\mathbf{u}|\mathbf{y}) \\ &= \frac{\int_{\mathbb{R}^q} \mathbf{u} \exp\{\mathbf{y}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \mathbf{1}^\mathsf{T} b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{G}^{-1}\mathbf{u}\}\, d\mathbf{u}}{\int_{\mathbb{R}^q} \exp\{\mathbf{y}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \mathbf{1}^\mathsf{T} b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}^\mathsf{T}\mathbf{G}^{-1}\mathbf{u}\}\, d\mathbf{u}.}\end{aligned}$$

In practice, computation of $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{G}}$ and $\widehat{\mathbf{u}}$ is hindered by the integrals over $\mathbb{R}^q$. McCulloch and Searle (2000, Chapter 9) summarise strategies for dealing with this computation. The simplest way to handle the integrals is to invoke Laplace's approximation. This coincides with penalized likelihood (e.g. Green, 1987).

4.1 Hazard estimation

An interesting connection between Poisson-normal mixed models and hazard function estimation is made by Cai, Hyndman and Wand (2002). The upshot is that automatic hazard estimation can be done using Poisson-normal software with an offset.

5 Twists

In many applications, the data do not arrive in a 'clean' form and adjustments are necessary. The mixed model approach to smoothing seems to lend itself to the modular handling of such twists.

5.1 Measurement error

In regression models in general, it is well-established that measurement error in a predictor distorts the relationship to the response; see Fuller (1987) and Carroll, Ruppert and Stefanski (1995). However, some measurement error models for parametric regression are simply based on maximum likelihood ideas (e.g. Section 12.3.3, Casella and Berger, 1990). Since the approach to smoothing described in the previous sections are also maximum likelihood, it is possible to fit semiparametric regression models through maximum likelihood and best prediction (Figure 3). Details are given in Ganguli, Staudenmayer and Wand (2002).

5.2 Missing data

Missing data is a common problem in many application areas, and careful adjustment is often necessary for trustworthy and efficient analyses. There are numerous proposed strategies for dealing with missing data; surveyed in the books by Little and Rubin (1987) and Schafer (1997). Particularly

noteworthy in the context of this article are likelihood-based models that account for missing data; for example, Little and Rubin, (1987), Ibrahim, (1990) and Ibrahim, Chen and Lipsitz (2001). The latter reference deals with generalized linear mixed models. French and Wand (2002) explored the extension of these ideas to semiparametric regression models with mixed model representation.

5.3 Outliers

Regression methodology that is resistant to the adverse effects of outlying response values has been the subject of an enormous amount of literature over the past few decades (e.g. Rousseeuw and Leroy, 1987). Some of the main approaches to robust regression involve M-estimation (e.g Huber, 1983) and the t-distribution likelihood (e.g. Lange, Little and Taylor, 1989). Welsh and Richardson (1997) provide a detailed survey of robustness in the linear mixed model, and the approaches described there could be used to 'robustify' the semiparametric regression estimators in this article. Kammann, Staudenmayer and Wand (2002) explored the t-distribution approach for penalized splines in the mixed model framework, which again just follows the paradigm of Figure 3.

6 Inference, Model Selection and Diagnostics

This paper concentrates on model fitting and vitally important follow-up topics: inference, model selection and diagnostics are neglected. However, since every model presented here is a mixed model, general mixed model methodology for each of these topics should be applicable. Pinheiro and Bates (2000), for example, describe inference and diagnostics for mixed models in longitudinal contexts. Shively, Kohn and Wood (1999) developed high-performance model selection strategies that could be useful for mixed model-based smoothing. Fung, Zhu, Wei and He (2002) develop influence diagnostics for a general class of mixed models.

7 Bayesian Analogue

This article follows a frequentist approach throughout. However, everything presented here has a Bayesian analogue, where $\boldsymbol{\beta}$, $\mathbf{G}$ and $\mathbf{R}$ are treated as random, with prior distributions imposed. Rather than the ML/BP approach used in the frequentist approach, estimation and inference for $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{G}$ and $\mathbf{R}$ is based on their respective posterior distributions (e.g. Gelman, Carlin, Stern and Rubin, 1995). Often the posterior distribution is computationally intractable and estimation and inference can be based on samples drawn from this distribution using Markov Chain Monte Carlo algorithms (e.g. Gilks, Richardson and Spiegelhalter, 1996).

8 Software

The SAS procedure PROC MIXED and the lme() function in both S-PLUS and R fit general design linear mixed models and therefore can aid the computation of many of the estimators described in this article. General design generalized linear mixed models can be fit using the glimmix macro in SAS. Ngo and Wand (2002) give several examples illustrating the fitting of the models of Sections 3 and 4.

9 Minimalist Statistics

One of the major themes of this paper is the use of the mixed model framework to fit and make inference for a wide variety of semiparametric regression models. This approach has the advantage of requiring little more than familiarity with mixed model methodology as outlined in Section 3 and Section 4. In particular, fitting is achieved through just two fundamental and well-established principles:

Estimation of parameters via (Restricted) Maximum Likelihood.

Prediction of random effects via Best Prediction.

When there is a twist, such as a predictor being subject to measurement error, then these principles can still be used for fitting. However, as seen in Section 4, maximum likelihood and best prediction are sometimes hindered by the presence of intractable integrals. Computational schemes such as Laplace approximation, Monte Carlo Expectation Maximisation and Markov Chain Monte Carlo algorithms are then important for implementation. If a Bayesian approach is used then similar comments apply; with fitting corresponding to finding the mode or mean of a posterior density. Inference, model selection and diagnostics can be made within the mixed model framework.

I like to call this streamlining of statistical methodology *minimalist statistics*. As the field of Statistics finds itself increasingly intertwined with other disciplines, and as required models become more complicated, I believe that such minimalism is very important. There is only so much time available to educate interdisciplinary researchers and practitioners in statistical theory and methodology. A forthcoming book: Ruppert, Wand and Carroll (2003) will expand on this theme.

Appendix: Demmler-Reinsch orthogonalisation

If $\mathbf{X}$ and $\mathbf{Z}$ contain the fixed and random effect basis functions for a scatterplot smooth (e.g. as in Section 3.2 or Section 3.6.1) and, as shown in Section

3.2.1, penalised spline regression corresponds to the ridge regression

$$\widehat{\mathbf{f}}_\alpha = \mathbf{C}(\mathbf{C}^\mathsf{T}\mathbf{C} + \alpha\mathbf{D})^{-1}\mathbf{C}^\mathsf{T}\mathbf{y} \tag{9.1}$$

for some diagonal matrix $\mathbf{D}$ and with $\mathbf{C} = [\mathbf{X}\ \mathbf{Z}]$. Here α controls the amount of smoothing and in the mixed model formulation of penalised splines $\alpha = \sigma_\varepsilon^2/\sigma_u^2$. Algorithm 1 allows for fast and stable calculation of (9.1).

Algorithm 1

Inputs: $\mathbf{y}$, $\mathbf{C}$, $\mathbf{D}$, α.

1. Obtain the Cholesky decomposition of $\mathbf{C}^\mathsf{T}\mathbf{C}$:

$$\mathbf{C}^\mathsf{T}\mathbf{C} = \mathbf{R}^\mathsf{T}\mathbf{R}.$$

2. Form the symmetric matrix $\mathbf{R}^{-\mathsf{T}}\mathbf{D}\mathbf{R}^{-1}$ and obtain its singular value decomposition:

$$\mathbf{R}^{-\mathsf{T}}\mathbf{D}\mathbf{R}^{-1} = \mathbf{U}\mathrm{diag}(\mathbf{s})\mathbf{U}^\mathsf{T}.$$

3. Compute the matrix and vector

$$\mathbf{A} \equiv \mathbf{C}\mathbf{R}^{-1}\mathbf{U} \quad \text{and} \quad \mathbf{b} \equiv \mathbf{A}^\mathsf{T}\mathbf{y}.$$

4. The fitted values are then (with element-wise division)

$$\widehat{\mathbf{f}}_\alpha = \mathbf{A}\left(\frac{\mathbf{b}}{\mathbf{1} + \alpha\,\mathbf{s}}\right).$$

The Cholesky decomposition applies only to nonsingular matrices. If $\mathbf{C}$ is ill-conditioned, it is advisable to add a small multiple of $\mathbf{D}$ to $\mathbf{C}^\mathsf{T}\mathbf{C}$ before applying the Cholesky decomposition, so that

$$\mathbf{C}^\mathsf{T}\mathbf{C} + \delta\mathbf{D} = \mathbf{R}^\mathsf{T}\mathbf{R},$$

where δ is small, e.g., $\delta = 10^{-10}$.

Once the matrix $\mathbf{A}$ and vectors $\mathbf{b}$ and $\mathbf{s}$ have been computed, the vector of fitted for different values of α reduces to a matrix multiplication. Therefore, $\widehat{\mathbf{f}}_\alpha$ can be computed cheaply for several α values. This is particularly useful for automatic smoothing parameter selection.

Justification of Algorithm 1

Now

$$\mathbf{R}^{-\mathsf{T}}\mathbf{D}\mathbf{R}^{-1} = \mathbf{U}\mathrm{diag}(\mathbf{s})\mathbf{U}^{\mathsf{T}} \quad \text{with} \quad \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}.$$

Since $\mathbf{U}$ is a square matrix, $\mathbf{U}^{\mathsf{T}} = \mathbf{U}^{-1}$ and so

$$\mathbf{D} = \mathbf{R}^{\mathsf{T}}\mathbf{U}\mathrm{diag}(\mathbf{s})\mathbf{U}^{-1}\mathbf{R}.$$

Also,

$$\mathbf{C}^{\mathsf{T}}\mathbf{C} = \mathbf{R}^{\mathsf{T}}\mathbf{R} = \mathbf{R}^{\mathsf{T}}\mathbf{U}\mathbf{U}^{-1}\mathbf{R}$$

and consequently

$$\mathbf{C}^{\mathsf{T}}\mathbf{C} + \alpha\mathbf{D} = \mathbf{R}^{\mathsf{T}}\mathbf{U}\{\mathbf{I} + \alpha\mathrm{diag}(\mathbf{s})\}\mathbf{U}^{-1}\mathbf{R}.$$

Hence

$$\begin{aligned}\widehat{\mathbf{f}}_\alpha &= \mathbf{C}[\mathbf{R}^{\mathsf{T}}\mathbf{U}\{\mathbf{I} + \alpha\mathrm{diag}(\mathbf{s})\}\mathbf{U}^{-1}\mathbf{R}]^{-1}\mathbf{C}^{\mathsf{T}}\mathbf{y} \\ &= (\mathbf{C}\mathbf{R}^{-1}\mathbf{U})\{\mathrm{diag}(\mathbf{1} + \alpha\mathbf{s})\}^{-1}(\mathbf{C}\mathbf{R}^{-1}\mathbf{U})^{\mathsf{T}}\mathbf{y} = \mathbf{A}\left(\frac{\mathbf{b}}{\mathbf{1} + \alpha\,\mathbf{s}}\right)\end{aligned}$$

where $\mathbf{A} \equiv \mathbf{C}\mathbf{R}^{-1}\mathbf{U}$ and $\mathbf{b} \equiv \mathbf{A}^{\mathsf{T}}\mathbf{y}$.

The new expression for $\widehat{\mathbf{f}}_\alpha$ is thus of the form

$$\widehat{\mathbf{f}}_\alpha = \mathbf{A}\{\mathbf{A}^{\mathsf{T}}\mathbf{A} + \alpha\mathrm{diag}(\mathbf{s})\}^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{y}.$$

Comparison with (9.1) shows that we have effectively replaced the basis functions in $\mathbf{C}$ with those in $\mathbf{A}$ where this design matrix has the *orthogonality* property $\mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{I}$. The columns of $\mathbf{A}$ correspond to the *Demmler-Reinsch* basis for the vector space spanned by $\mathbf{C}$. The orthogonality property is crucial for the fast computation over several smoothing parameters.

Acknowledgements

The ideas summarised in this article are the result of interaction with several of my colleagues at Harvard School of Public Health in the period 1997–2002: Babette Brumback, Tianxi Cai, Brent Coull, Jonathan French, Bhaswati Ganguli, Erin Kammann, Long Ngo, Nan Laird, Helen Parise, Louise Ryan, Misha Salganik, Joel Schwartz, John Staudenmayer, Sally Thurston, Jim Ware and Yihua Zhao. The paper has also benefited greatly from conversations with Marc Aerts, Ray Carroll, Gerda Claeskens, Ciprian Crainiceanu, Maria Durban, Jim Hobert, Robert Kohn, Xihong Lin, Mary Lindstrom, Michael O'Connell, José Pinheiro and David Ruppert. I am grateful to Professors Trevor Hastie and Gareth James for making the spinal bone mineral density data available. Finally, thank you to participants in the Eurowork-shop on Nonparametric Models (HPCFCT-2000-00041) held in Bernried, Germany in November, 2001 and for its co-organiser, Göran Kauermann, for encouraging me to write this paper. This paper was supported by U.S. National Institute of Environmental Health Sciences grant R01-ES10844-01.

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations (with discussion). *Journal of the American Statistical Association*, **96**, 939–967.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brumback, B.A. and Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, **93**, 961–994.

Brumback, B.A., Ruppert, D. & Wand, M.P. (1999). Comment on Shively, Kohn and Wood. *Journal of the American Statistical Association*, **94**, 794–797.

Cai, T., Hyndman, R.J. and Wand, M.P. (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics*, **11**, in press.

Carroll, R. J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Casella, G. and Berger, R. L. (1990). *Statistical Inference (Second Edition)*. Pacific Grove, California: Thomson Learning.

Chaudhuri, P. & Marron, J.S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807–823.

Chen, Z. (1993). Fitting multivariate regression functions by interaction spline models. *Journal of the Royal Statistics Society, Series B* , **55**, 473–491.

Coull, B.A., Ruppert, D. and Wand, M.P. (2001). Simple incorporation of interactions into additive models. *Biometrics*, **57**, 539–545.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Diggle, P., Liang, K.-L. and Zeger, S. (1995). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Diggle, P. (1997). Spatial and longitudinal data analysis: Two histories with a common future? In *Proceedings of the Nantucket conference on Modeling Longitudinal and Spatially Correlated Data : Methods, Applications, and Future Directions. Lecture Notes in Statistics 122,* Gregoire, T., Brillinger, D.R., Diggle, P.J., Rusek-Cohen, E., Warren, W.G., Wolfinger, R.D. (eds), Springer-Verlag, New York, 387–402.

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis (Third Edition).* New York: John Wiley & Sons.

Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

French, J.L., Kammann, E.E. & Wand, M.P. (2001). Comment on Ke and Wang. *Journal of the American Statistical Association*, **96**, 1285–1288.

French, J.F. and Wand, M.P. (2002). Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics*, **3**, 000-000.

Fuller, W.A. (1987). *Measurement Error Models.* New York: John Wiley & Sons.

Fung, W.-K., Zhu, Z.-Y., Wei, B.-C. and He, X. (2002). Influence diagnostics and outlier tests for semiparametric mixed models. *Journal of the Royal Statistical Society, Series B.* **64**, 565–579.

Ganguli, B., Staudenmayer, J. and Wand, M.P. (2002). Additive models with predictors subject to measurement error. Unpublished manuscript.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis.* Boca Raton, Florida: Chapman and Hall.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Gray, R. J. (1992). Spline-based tests in survival analysis. *Biometrics*, **50**, 640–652.

Green, P.J. (1985). Linear models for field trials, smoothing and cross-validation. *Biometrika*, **72**, 523–537.

Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, **55**, 245–259.

Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society, Series B*, **58**, 379–396.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficients models. *Journal of the Royal Statistics Society, Series B* , **55**, 757–796.

Hastie, T. and Tibshirani, R.J. (2000). Bayesian backfitting. *Statistical Science*, **15**, 196–223.

Huber, P. (1983). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 73–101.

Ibrahim, J.G. (1990). Incomplete data *Journal of the American Statistical Association*, **85**, 765–769.

Ibrahim, J.G., Chen, M.H., and Lipsitz, S.R. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, **88**, 551–564.

James, G.M. and Hastie, T.J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society, Series B*, **63**, 533–550.

James, G.M., Hastie, T.J. and Sugar, C.A. (2000). Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.

Johnson, M.E., Moore, L.M. and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.

Kammann, E.E. & Wand, M.P. (2003). Geoadditive models. *Applied Statistics*, **52**, 1–18.

Kammann, E.E., Staudenmayer, J. and Wand, M.P. (2002). Robustness for general design mixed models using the t-distribution. Unpublished manuscript.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association*, **96**, 1272–1281.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, **46**, 1071–1085.

Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lange, K.L., Little, R.J.A. and Taylor, J.M.G. (1989). Robust statistical modeling using the t-distribution. *Journal of the American Statistical Association*, **84**, 881-896.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B*, **61**, 381–400.

Little, R.J. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

MathSoft Inc. (2002).

McCulloch, C.E., and Searle, S.R. (2000). *Generalized, Linear, and Mixed Models.* New York: John Wiley & Sons.

Ngo, L. and Wand, M.P. (2002). Smoothing with mixed model software. Technical report.

Nychka, D.W. (2000). Spatial process estimates as smoothers. In *Smoothing and Regression* (M. Schimek, ed.). Heidelberg: Springer-Verlag.

Nychka, D. & Saltzman, N. (1998). Design of Air Quality Monitoring Networks. In *Case Studies in Environmental Statistics* (D. Nychka, Cox, L., Piegorsch, W. eds.), Lecture Notes in Statistics, Springer-Verlag, 51–76.

Nychka, D., Haaland, P., O'Connell, M., Ellner, S. (1998). FUNFITS, data analysis and statistical tools for estimating functions. In *Case Studies in Environmental Statistics* (D. Nychka, W.W. Piegorsch, L.H. Cox, eds.), New York: Springer-Verlag, 159–179.

O'Connell, M.A. and Wolfinger, R.D. (1997). Spatial regression models, response surfaces, and process optimization. *Journal of Computational and Graphical Statistics*, **6**, 224–241.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems

(with discussion). *Statistical Science*, **1**, 505–527.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, **9**, 363–379.

Parker, R.L. & Rice, J.A. (1985). Discussion of "Some aspects of the spline smoothing approach to nonparametric regression curve fitting" by B.W. Silverman. *Journal of the Royal Statistical Society, Series B*, **47**, 40-42.

Patterson, H.D. and Thompson, R. (1973). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–51.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735–757.

Ruppert, D. & Carroll, R.J. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, **42**, 205–224.

Ruppert, D., Wand, M. P. & Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

SAS Institute, Inc. (2002).

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*. New York: John Wiley & Sons.

Shively, T.S., Kohn, R. and Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, **94**, 777–794.

Speed, T. (1991). Comment on paper by Robinson. *Statistical Science*, **6**, 42–44.

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* New York: Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological*, **58**, 267–288.

Verbyla, A.P. (1994). Testing linearity in generalized linear models. *Contributed Pap. 17th Int. Biometric Conf., Hamilton, Aug. 8th-12th*, 177.

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). *Journal of the Royal Statistics Society, Series C* , **48**, 269–312.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B*, **40**, 364-372.

Wahba, G. (1986). Partial interaction spline models for the semiparametric estimation of functions of several variables. *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface*, 75–80.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

Wang, Y. (1998a). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association* , **93**, 341–348.

Wang, Y. (1998b). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, **60**, 159–174.

Wecker, W.E. and Ansley, C.F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, **78**, 81–89.

Welsh, A.H. and Richardson, A.M. (1997). Approaches to the robust estimation of mixed models. In *Handbook of Statistics, Vol. 15* (G. S. Maddala and C.R. Rao eds.), Amsterdam: Elsevier Science.