

Semiparametric Regression During 2003–2007

D. RUPPERT, M.P. WAND and Raymond J. CARROLL

Semiparametric regression is a fusion between parametric regression and nonparametric regression that integrates low-rank penalized splines, mixed model and hierarchical Bayesian methodology – thus allowing more streamlined handling of longitudinal and spatial correlation. We review progress in the field over the five-year period between 2003 and 2007. We find semiparametric regression to be a vibrant field with substantial involvement and activity, continual enhancement and widespread application.

KEY WORDS: Asymptotics; Boosting; BUGS; Functional data analysis; Generalized linear mixed models; Graphical models; Hierarchical Bayesian models; Kernel machines; Longitudinal data analysis; Mixed models; Monte Carlo methods; Penalized splines; R; Spatial statistics.

28th April, 2009

1 Introduction

Semiparametric regression is a fusion between traditional parametric regression analysis (e.g. Cook and Weisberg, 1982; Draper and Smith, 1998) and newer nonparametric regression methods (e.g. Wahba, 1990; Hastie and Tibshirani, 1990; Green and Silverman, 1994). This emerging field synthesizes research across several branches of Statistics: parametric and nonparametric regression, longitudinal and spatial data analysis, mixed and hierarchical Bayesian models, Expectation-Maximization (EM) and Markov Chain Monte Carlo (MCMC) algorithms. Semiparametric regression is a field deeply rooted in applications and its evolution reflects the increasingly large and complex problems that are arising in science and industry.

We do not view semiparametric regression as a competitor to parametric and nonparametric approaches, but rather as a bridge between them. The need for parsimonious statistical models is well-known and parametric models are often a convenient method for achieving parsimony. However, nonparametric models exist because there are many examples where parametric models do not provide a suitable fit to the data. Semiparametric modeling allows a researcher to have the best of both worlds: the parametric and the nonparametric. Those features of the data that are suitable for parametric modeling are modeled that way and nonparametric components are used only where needed. For example, in the study discussed in Section 4.1, the effect of blood lead concentration on a

⁰D. Ruppert is Andrew Schultz, Jr., Professor of Engineering, (E-mail: dr24@cornell.edu) School of Operations Research and Industrial Engineering Cornell University, Ithaca, NY 14853, USA, M.P. Wand is Research Professor in Statistics (E-mail: mwand@uow.edu.au), School of Mathematics and Applied Statistics, University of Wollongong, Wollongong 2522, Australia and Raymond J. Carroll is Distinguished Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX, 77843-3143, USA. Ruppert's research was supported by a grant from the National Cancer Institute (CA57030) and from the National Science Foundation (DMS-0805975). Wand's research was supported by a grant from the Australian Research Council (DP0877055). Carroll's research was supported by grants from the National Cancer Institute (CA57030, CA104620), and also in part by award number KUS-CI-016-04 made by the King Abdullah University of Science and Technology. The authors are grateful to several researchers involved with semiparametric regression for providing copies of relevant papers. We are especially grateful to Ciprian Crainiceanu, Yanan Fan, Yingxing Li, Jean Opsomer and Sue Welham for their assistance in the preparation of this article.

child's intelligence quotient is modeled with a spline in order to detect a nonlinear dose response. In that study, the effects of the numerous confounders are modeled linearly and within-child correlations are modeled by a parametric mixed model to achieve a parsimonious fit. The extreme cases, fully parametric or completely nonparametric models, can be used when they are appropriate.

Two prominent features throughout much of semiparametric regression are:

- keeping the nonparametric regression part relatively simple by using low-rank penalized splines;
- utilizing the mixed model representation of penalized splines.

These bring several benefits. Firstly, longitudinal and spatial effects are easily incorporated. Secondly, fitting and inference can be performed within the established frameworks of maximum likelihood and best prediction. Established mixed model software in R and SAS can aid implementation. If a Bayesian approach is used then the infrastructure of Bayesian inference can be called upon. This includes the BUGS software project (e.g. Lunn *et al.* 2000). The Bayesian/BUGS route is particularly attractive in non-standard situations, such as when the data are overdispersed or incomplete. An overarching benefit is *extendability*: the modularity of the mixed model-based penalized spline approach allows 'twists', such as missing data, to be incorporated in a straightforward manner.

Early contributions on the mixed model representations of curve fitting include Wahba (1978), Green (1985), Thompson (1985), Speed (1991), Verbyla (1994), Donnelly, Laird and Ware (1995), O'Connell and Wolfinger (1997) and Wang (1998). In addition, Parker & Rice (1985), O'Sullivan (1986) and Eilers & Marx (1996) represent early work on low-rank penalized splines in nonparametric regression.

In 2003 we published the *Semiparametric Regression* book (Ruppert, Wand and Carroll, 2003). Since it is the first book to make use of both of these ideas, its publication 6 years ago constitutes some sort of 'line in the sand' for this exciting area of research. Although *Semiparametric Regression* was released in April 2003, the final drafts were written in late 2002. Hence it contains a survey of the literature up until the end of 2002 (roughly).

In this review we revisit the field 5 years later and summarize the state of the field as of the end of 2007. We are pleased to report that semiparametric regression is a thriving area of research with contributions to its theory, methodology and software being continually made by research teams around the world. Especially pleasing is the rate at which semiparametric regression is being used in applications. While surveying the area over 2003–2007 we learned about applications in several fascinating and diverse areas, including on-line auctions, genomics, air pollution, agricultural soil and cosmology. A great deal of penalized splines (especially smoothing splines) research does not make use of their mixed model representation. For example, Wood (2006a) and the accompanying R package `mgcv` (Wood, 2008) mainly uses generalized cross-validation (GCV) and a version of Akaike's Information Criterion (AIC).

There is also an enormous literature on flexible regression analysis that does not involve penalized splines. Examples include regression splines (e.g. Stone *et al.* 1997), local polynomials (e.g. Fan and Gijbels, 1995) and wavelets (e.g. Ogden, 1996). Ruppert *et al.* (2003) discuss each of these choices but promote the mixed model-based penalized spline approach to semiparametric regression. Largely because of time and space limitations, we will stay mainly with this approach throughout the review.

1.1 Summary of mixed model-based penalized spline approach

In this section we provide a summary of the mixed model-based penalized spline approach to semiparametric regression that is adopted by many of the papers in this review.

We begin with some examples of semiparametric regression models:

$$y_i \stackrel{\text{ind.}}{\sim} \text{Poisson}[\exp\{\beta_1 x_{1i} + f_2(x_{2i}) + f_{34}(x_{3i}, x_{4i})\}], \quad 1 \leq i \leq n, \quad (1)$$

$$y_i \stackrel{\text{ind.}}{\sim} \text{Binomial}[n_i, \text{logit}^{-1}\{\beta_0(x_{1i}) + \beta_1(x_{1i})x_{2i}\}], \quad 1 \leq i \leq n, \quad (2)$$

$$y_{ij} | u_{i,\text{sbj}} \stackrel{\text{ind.}}{\sim} N(u_{i,\text{sbj}} + f_1(x_{1i}) + \beta_2^\top \mathbf{x}_{2i}, \sigma_\varepsilon^2), \quad u_{i,\text{sbj}} \stackrel{\text{ind.}}{\sim} N(0, \sigma_{\text{sbj}}^2), \quad 1 \leq j \leq n_i, \quad 1 \leq i \leq m. \quad (3)$$

Here x_{1i}, \dots, x_{4i} are scalar predictors corresponding to the response variable y_i , \mathbf{x}_{2i} is a vector of predictors and $u_{i,\text{sbj}}$ is a random subject intercept with variance σ_{sbj}^2 . The term $f_2(x_{2i})$ means a smooth function of x_{2i} . Other functional notation is defined similarly. Model (1) is an extension of the generalised additive model paradigm that allows non-parametric bivariate components. If (x_{3i}, x_{4i}) correspond to geographic position then (1) is sometimes called a *geoadditive model* (e.g. Kammann & Wand, 2003). In Model (2), β_0 and β_1 are smooth functions of the x_1 variable. This model is known as a *varying coefficient model*. Model (3) is usually called an *additive mixed model* since it represents the fusion of an additive model and a linear mixed model.

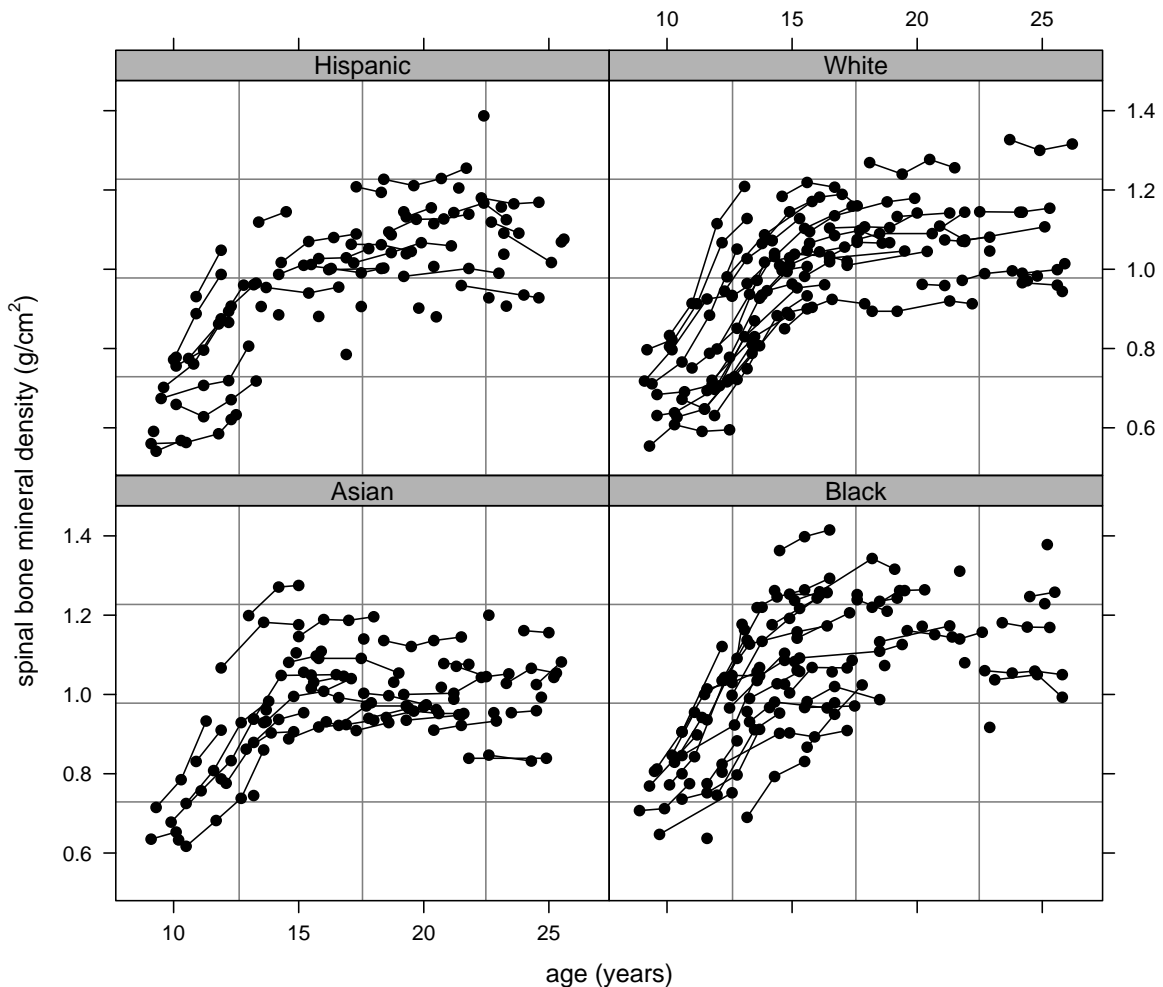


Figure 1: Data on spinal bone mineral density versus age broken down according to ethnicity of the female subjects. Points for the same subject are connected by lines. The source of these data is Bachrach et al. (1999).

An example data set that benefits from (3) is shown in Figure 1. It consists of longitudinal measurements on the spinal bone mineral density of a cohort of young female subjects (source: Bachrach *et al.* 1999). A question of interest is how spinal bone mineral density differs among the four ethnicity groups. However, the analysis is complicated by (a) the non-linear effect of age, and (b) correlation arising from repeated measurements on the same subject. Model (3) with the x_{1i} s corresponding to the age measurements and the x_{2i} s corresponding to ethnicity indicators is appropriate.

In the mixed model approach to semiparametric regression nonparametric functional relationships are handled through modelling mechanisms such as:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^K u_{k,\text{spl}} z_k(x), \quad u_{k,\text{spl}} \text{ i.i.d } N(0, \sigma_{\text{spl}}^2). \quad (4)$$

Here z_1, \dots, z_K are a set of spline basis functions. The simplest example is $z_k(x) = (x - \kappa_k)_+$ for some knot sequence $\kappa_1, \dots, \kappa_K$. Here u_+ equals u for $u \geq 0$ and equals 0 otherwise. However, more sophisticated options now exist and these are reviewed in Section 2.1. Most of the spline bases described there are in accordance with the classical nonparametric regression method known as *smoothing splines* (e.g. Wahba, 1990; Eubank, 1999). This approach is extendable to multivariate functions using either radial basis functions (e.g. Wood, 2003; Ruppert *et al.*, 2003) or tensor products (e.g. Wood, 2006c).

A consequence of (4) is that many frequentist semiparametric regression models are expressible as

$$E(\mathbf{y}|\mathbf{u}) = g(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}), \quad \mathbf{u} \sim (\mathbf{0}, \mathbf{G}), \quad (5)$$

where \mathbf{y} denotes the vector of responses and \mathbf{X} and \mathbf{Z} are design matrices. 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 $E(\mathbf{v}) = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{v}) = \boldsymbol{\Sigma}$. The fixed effects term, $\mathbf{X}\boldsymbol{\beta}$, handles covariates that enter the model linearly. The random effects component $\mathbf{Z}\mathbf{u}$ and corresponding covariance matrix \mathbf{G} handles non-linear effects through spline basis functions, but may also incorporate random subject effects and spatial correlation structure in longitudinal and spatial contexts. There will often be other parameters arising, for example, in the variance structure (e.g. $\mathbf{R} = \text{Cov}(\mathbf{y}|\mathbf{u})$) but we will ignore this in the current discussion.

Most commonly (5) is embedded in a fully specified probabilistic model. This allows fitting and inference to be achieved through the paradigms:

$$\begin{aligned} & \text{Maximum Likelihood (ML) for } \boldsymbol{\beta}; \\ & \text{Restricted Maximum Likelihood (REML) for } \mathbf{G}; \\ & \text{Best Prediction (BP) for } \mathbf{u}. \end{aligned} \quad (6)$$

BP is defined according to minimum mean squared error and has the solution $\hat{\mathbf{u}} = E(\mathbf{u}|\mathbf{y})$ (e.g. McCulloch, Searle & Neuhaus, 2008). Depending on the form of the model (e.g. normal versus Poisson) execution of (6) can range from easy exact calculation using standard mixed model software (e.g. `lme()` in the R language; R Development Core Team, 2008) to difficult approximation via computationally intensive algorithms such as MCMC.

The hierarchical Bayesian version of (5) takes the form

$$\begin{aligned} [\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}] &= f_1(\mathbf{y}; \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}); \quad [\mathbf{u}|\mathbf{G}] = f_2(\mathbf{u}; \mathbf{G}) \\ [\boldsymbol{\beta}] &= f_3(\boldsymbol{\beta}; \mathbf{A}_\boldsymbol{\beta}); \quad [\mathbf{G}] = f_4(\mathbf{G}; \mathbf{A}_\mathbf{G}) \end{aligned} \quad (7)$$

where $\mathbf{A}_\boldsymbol{\beta}$ and $\mathbf{A}_\mathbf{G}$ are hyper-parameters, f_1, \dots, f_4 are fixed conditional density or probability mass functions and $[v|w]$ denotes the conditional density or probability mass function of v given w . Inference is based on posteriors for parameters of interest; in particular

$$[\boldsymbol{\beta}|\mathbf{y}], \quad [\mathbf{u}|\mathbf{y}] \quad \text{and} \quad [\mathbf{G}|\mathbf{y}].$$

In semiparametric regression it is very rare that analytical solutions for these posteriors exist and approximation methods need to be employed. MCMC approximation via the BUGS software (e.g. Lunn *et al.* 2000) often provides satisfactory solutions.

In the interests of conciseness, we will not give specific details or examples of (5) and (7). These can be found in the Ruppert *et al.* (2003), Wand (2003), Ngo & Wand (2004), Gurrin, Scurrah & Hazelton (2006), Crainiceanu, Ruppert & Wand (2005) and Zhao, Staudenmayer, Coull & Wand (2006), for example.

1.2 Layout of Review

The rapidity with which semiparametric regression is growing as a field means that a concise and informative review of the five years since 2002 is quite challenging. For instance, we estimate that more than three hundred papers in 2003–2007 are connected with the area – most of which we read or skimmed in preparing this review. After surveying the literature we decided on the following layout for the remainder of the article:

Section 2: **Advancement of Primitives**

By *primitives* we mean the ‘nuts and bolts’ of semiparametric regression. Examples include spline basis specification, computing and asymptotic theory. Much of Ruppert *et al.* (2003) is concerned with the primitives of semiparametric regression. However, some have undergone noticeable refinement in the past five years. Section 2 summarizes these developments.

Section 3: **Advancement of Models and Methods**

During 2003–2007 semiparametric regression models have continually become more sophisticated in response to the complexities of contemporary data sets and scientific questions. Section 3 reviews these advancements.

Section 4: **Applications**

Semiparametric regression is very much an applications-oriented branch of Statistics. In Section 4 we highlight several case studies which have benefited from the semiparametric regression paradigm.

1.3 Overlooked literature

The production of this review article has involved an immense amount of retrieval and reading over a relatively short time period. While we have tried hard to peruse all relevant contributions it is certain that some have been inadvertently overlooked. We welcome any omissions being drawn to our attention. Also, we point out that the end of 2007 cut-off for inclusion in this review is slightly fuzzy. For example, some relevant papers that we have known about for some time turned out to be 2008 or 2009 papers. These are still included.

2 Advancement of Primitives

In this section we summarize 2003–2007 research on the primitives of semiparametric regression with emphasis on important advancement.

2.1 Univariate spline bases

All commonly used penalized spline models for a smooth real-valued function f spline can be expressed in the form

$$f(x; p, \mathbf{z}) = \beta_0 + \dots + \beta_p x^p + \sum_{k=1}^K u_k z_k(x)$$

where p is the degree of the polynomial component and $\{z_1(\cdot) : k = 1, \dots, K\}$ is a set of spline basis functions for handling departures from p th degree polynomials. The spline coefficients $\mathbf{u} = (u_1, \dots, u_K)$ are subject to penalization. In the mixed model representation \mathbf{u} is usually taken to be random according to $N(\mathbf{0}, \mathbf{G})$ for some \mathbf{G} . Already it is clear that there are a lot of options for spline bases and the penalization. Without loss of generality, we can take $\mathbf{G} = \sigma_u^2 \mathbf{I}$ since this just involves a linear transformation of the z_k s. There is also a lot to be said for taking the polynomial to be linear – for example, if tests for linearity are of interest. So, while $p > 1$ may sometimes be desirable, the $p = 1$ *canonical form*

$$f(x; \mathbf{z}) = \beta_0 + \beta_1 x + \sum_{k=1}^K u_k z_k(x), \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}), \quad (8)$$

is useful for sorting out the various spline basis options. In addition, (8) is convenient for implementation since it corresponds to the standard mixed model structure (with x_1, \dots, x_n being data on the x variable):

$$\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}), \quad \text{where } \mathbf{X} = [1 \ x_i]_{1 \leq i \leq n} \quad \text{and} \quad \mathbf{Z} = [z_k(x_i)]_{\substack{1 \leq i \leq n \\ 1 \leq k \leq K}}.$$

Ruppert *et al.* (2003, Chapter 2) survey options and strategies for K and the z_k s up to about 2002. However, there have been some interesting developments since then. Currie and Durbán (2002) show how the Eilers and Marx (1996) P-splines can be expressed in mixed model forms such as (8). Welham, Cullis, Kenward and Thompson (2007) produce a useful exposition on how the various versions of penalized splines are connected to each other. In a similar vein to (8) they propose ‘A general model for polynomial splines’ that includes several options under one umbrella. A large-scale simulation comparison study shows no clear winner across all settings. Some practical advice about penalty choice and order of differencing is offered. The ‘minimum worst possible change’ approach of Wood (2003), described in Section 2.2 for multivariate smoothing, also yields univariate low-rank spline bases as a special case.

Wand and Ormerod (2008) study the O’Sullivan (1986) low-rank approximation of smoothing splines. The name *O-splines* is suggested for this exact counterpart of P-splines. Results for exact computation of the z_k are derived that allow implementation in R with only a few lines of code. A simulation study shows O-splines and P-splines to be quite close in the interior, but the latter to have better extrapolation behavior, and also very close to smoothing splines even for $K \ll n$. Given the well-established good properties of smoothing splines, such as natural boundary behavior and asymptotic optimality (Nussbaum, 1985), the evidence points towards O-splines as the better option in comparison with P-splines, and as an excellent default for univariate spline bases in semiparametric regression analysis.

2.2 Multivariate smoothing

In principle, all smoothing techniques can be extended to the multivariate case. In practice, though, this extension is a delicate art because of the additional complications that high-dimensional domains bring. Chapter 13 of Ruppert *et al.* (2003) summarised bivariate smoothing approaches based on kriging and splines, including low-rank extensions.

General multivariate extensions were briefly described. We now summarize interesting new work in this direction from the last few years.

Wood (2003) develops an approach to low-rank thin plate spline smoothing that circumvents the knot placement issue. His basis reduction is based on a ‘worst possible change’ criterion. For small data sets, implementation involves standard linear algebra manipulations, while Lanczos iteration is suggested for larger sample sizes. In Wood (2006c) the same author opts for tensor products as a means of extending penalized splines smoothing to the multivariate situation. Scale invariance is the main mechanism for achieving this extension in an attractive way. An advantage is this version of multivariate smoothing is that each direction has its own smoothing parameter. Particular attention is also paid to the cogent incorporation of random effect structure for generalized additive mixed modeling.

Fahrmeir, Kneib & Lang (2004) show that common geostatistical approaches to bivariate smoothing have a representation in terms of stationary Gaussian random fields. They then point out that Gaussian random fields can be approximated by Markov random fields and that the latter has computational advantages. Markov random fields are also a common vehicle for Bayesian smoothing of spatial count data. Hence, the Markov random field approach to bivariate smoothing has the advantage of being in concert with that used for spatial count data. Kneib & Fahrmeir (2006) also use the Markov random field approach to bivariate smoothing and relate it to mixed models.

Currie, Durban & Eilers (2006) and Eilers, Currie & Durbán (2006) treat the special case of smoothing on multidimensional grids. They develop an arithmetic that results higher computation speed and lower storage requirements.

Paciorek (2007a,b) investigates the use of the spectral representation of stationary Gaussian process structure (Wikle, 2002) in semiparametric regression contexts. He identifies advantages for large sample sizes and MCMC mixing in the generalized response situation.

The problem of complicated domains in bivariate smoothing is addressed by Wang & Ranalli (2007). Motivated by a study on mercury concentrations in estuaries, Euclidean ‘as crow flies’ distance is replaced by a geodesic ‘as fish swims’ distance. This distance depends on the intrinsic structure of the domain and needs to be estimated. A procedure based on shortest path theory and Floyd’s algorithm (Floyd, 1962) is described.

2.3 Bayesian semiparametric regression

Bayesian semiparametric regression is progressively becoming more prevalent, and could eventually challenge the frequentist version in terms of popularity. Reasons include (1) the attractiveness of hierarchical Bayesian models for quantifying multiple sources of variability, (2) models becoming more sophisticated (e.g. dealing with complications such as missingness and measurement error) to the point that standard (likelihood-based) mixed model software cannot be used, (3) continual improvement of Monte Carlo methods for Bayesian inference, and (4) continual improvement of the BUGS computing environment (Lunn *et al.* 2000) for MCMC sampling from posterior distributions of interest. We expand on aspect (4) in Section 2.7.

Recent Bayesian modeling research has also impacted upon Bayesian semiparametric regression since 2002. A prominent example is Gelman (2006) which advises on non-informative prior distribution specification for variance parameters and, in particular, argues against the use of inverse Gamma distributions.

With robustness in mind, Jullion & Lambert (2007) study prior specification for Bayesian P-splines models. Advanced Bayesian hierarchical modeling is used, including the use of Dirichlet-based mixture priors.

Several other Bayesian semiparametric regression contributions, involving new models and methodology, are described in Section 3.

Bayesian methodology and software is currently an area of vigorous research activity – in both Statistics and Computer Science (see Section 2.5). This has led to the Bayesian brand of semiparametric regression becoming more prominent in recent years; a trend that we expect to continue.

2.4 Monte Carlo methods

Since the early 1990s Markov Chain Monte Carlo (MCMC) methods have been a mainstay of Bayesian inference. However, in the intervening years, we have noticed the emergence of new Monte Carlo methods. Some of these are more elaborate versions of MCMC, while others fall outside of the Markov chain paradigm.

Staying first within the MCMC family we note that specifically tailored Metropolis-Hastings schemes are developed by Baladandayuthapani *et al.* (2005), Paciorek & Schervish (2006), Gryparis *et al.* (2007) and Baladandayuthapani *et al.* (2008). The `BayesX` software package makes use of elaborate MCMC schemes. In each case, the goal is improved mixing for the complex semiparametric regression model at hand.

The single component adaptive Metropolis algorithm of Haario, Saksman & Tamminen (2005) is a recent modification of the random walk Metropolis-Hastings algorithm that adapts according to what it has learnt from previous sampled iterates. The resulting chain is not Markovian, although Haario *et al.* (2005) prove that it does lead to samples from the correct posterior distributions. The adaptation aspect means that fiddly tuning runs are not required. Nott (2006) successfully applied Haario *et al.*'s algorithm to a semiparametric regression setting.

Quasi-Monte Carlo is a vibrant research area in the general problem of high-dimensional numerical integration via importance sampling. It differs from ordinary Monte Carlo integral approximation in that random samples are replaced by cleverly chosen deterministic ones. While much of quasi-Monte Carlo research is outside of Statistics, Hickernell, Lemieux & Owen (2005) provides a recent survey for a statistical audience. Kuo, Dunsmuir, Sloan, Wand & Womersley (2008) apply state-of-the art quasi-Monte Carlo algorithms to a class of statistical problems that encompass some important semiparametric regression models.

Sequential Monte Carlo samplers are a generalization of importance sampling that produce weighted samples from the target distribution by sampling sequentially from a slowly evolving set of distributions. Del Moral, Doucet & Jasra (2006) is the main reference for this emerging methodology. Fan, Leslie and Wand (2006) represents early work on application of sequential Monte Carlo samplers to Bayesian semiparametric regression.

Other recent developments in Monte Carlo methods that lend themselves to semiparametric regression applications include slice sampling with 'stepping out' (Neal, 2003) and approximate Bayesian computation (e.g. Beaumont *et al.* 2002; Marjoram *et al.* 2003; Sisson *et al.* 2007).

2.5 Computer Science interface

The foreword of a recent special issue of *Statistical Science* proclaimed the "the dissolving of the frontier between Statistics and Computer Science" (Casella and Robert, 2004). In 2006 *Statistica Sinica* had a special issue titled *Challenges in Statistical Machine Learning*. Hastie, Tibshirani and Friedman's cross-disciplinary book *The Elements of Statistical Learning* has had colossal impact since its publication in 2001. In keeping with this *zeitgeist*, strong connections between semiparametric regression and contemporary Computer Science are becoming apparent.

Most of the connections are concerned with methodology for classification (or *supervised learning* in the Computer Science world) and the sub-field of Computer Science

known as Machine Learning. Support vector machines (e.g. Moguerza & Muñoz, 2006) and other kernel machines share many attributes and issues with nonparametric regression (e.g. Hastie & Zhu, 2006). Wahba’s (2006) comment on Moguerza & Muñoz (2006) describes recent convergence between support vector machine and regularization research. Pearce & Wand (2006) show how penalized splines and semiparametric regression structure such as additive models can be embedded within the kernel machine framework.

Boosting (Schapire, 1990), described in Section 3.1, is another innovation from Machine Learning that is now benefiting semiparametric regression. For example, Bühlmann & Yu (2003) use smoothing spline theory and simulations to explain the interplay between the number of boosting iterations and the ‘weakness’ of the smoother. Tutz & Reithinger (2007) apply their lessons to semiparametric mixed models and derive an alternative fitting algorithm called *BoostMixed*.

Another area on the Computer Science interface where we see great potential for benefits to semiparametric regression is *graphical models* (e.g. Jordan, 2004). Wand (2009) provides detailed discussion on this topic. Directed acyclic graphs have become a common way of representing hierarchical Bayesian models and, indeed, comprise the architecture on which BUGS is built (Cowell *et al.*, 1999). Figure 2 is a directed acyclic graph representation of the model:

$$\begin{aligned} \text{logit}\{P(y_i = 1|\mathbf{u})\} &= \beta_0 + \beta_1 x_i + \sum_{k=1}^K z_k(x_i) = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i; & \mathbf{u}|\sigma_u &\sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}) \\ [\beta_0, \beta_1] &\sim N(\mathbf{0}, \sigma_\beta^2 \mathbf{I}); & [\sigma_u] &= \frac{2A}{\pi(\sigma_u^2 + A^2)}, \quad \sigma_u > 0. \end{aligned} \tag{9}$$

where the data are $(x_i, y_i) \in \mathbb{R} \times \{0, 1\}$, $1 \leq i \leq n$, the z_k are a spline basis as described in Section 1.1, and $\sigma_\beta, A > 0$ are hyper-parameters. Nodes of the graph correspond to the components of the model, while arrows convey ‘parent-child’ relationships of the hierarchical structure.

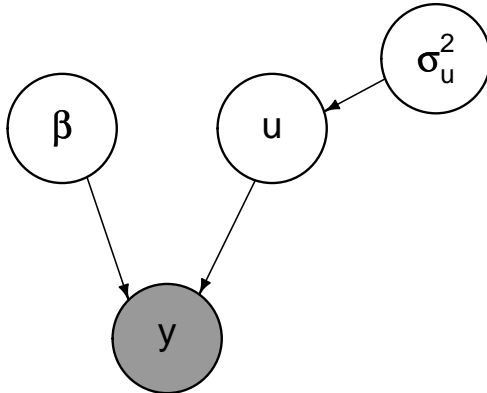


Figure 2: Directed acyclic graph representation of model (9). The shaded node corresponds to the observed data.

Suppose we add the complication that the x_i in (9) are subject to measurement error and that we instead observe $w_i = x_i + z_i$ where the x_i are now modeled to be from a $N(\mu_x, \sigma_x^2)$ distribution and the contaminating variable z_i is from a known fixed distribution. Then an appropriate hierarchical Bayesian model is that represented by Figure 3, a more complex graph with four additional edges and nodes.

MCMC is currently the most common mechanism for approximation of posteriors in the models depicted in Figures 2 and 3. The graphical models setting allows for graph-theoretic structure, such as *Markov blankets*, to be exploited in the design and implementa-

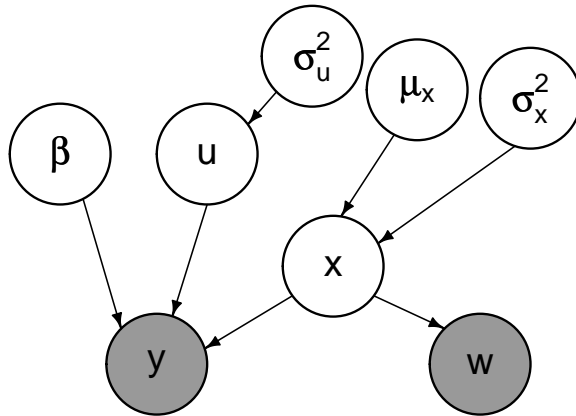


Figure 3: Directed acyclic graph representation of model (9), but with the predictor subject to measurement error. Shaded nodes correspond to the observed data.

tion of MCMC algorithms (Jordan, 2004). An emerging alternative to MCMC is *variational approximation* (e.g. Jordan, Ghahramani, Jaakkola and Saul, 1999). Joint work between the second author and J.T. Ormerod is investigating variational approximations that are specific to semiparametric regression analysis.

Interplay with Computer Science is one of the most exciting recent developments in semiparametric regression. We anticipate this to be an increasingly fruitful area of research.

2.6 Asymptotic theory

Hall & Opsomer (2005) is the first paper to study the asymptotic theory of penalized splines. They replace the nonparametric regression model $y_i = m(x_i) + e_i, i = 1, \dots, n$, where x_i is in a compact interval \mathcal{I} , by the model $y(t) = m(t) + e(t), t \in \mathcal{I}$, where $e(t) = n^{-1/2}v(t)^{-1/2}DW(t)$ and $DW(t)$ is the ‘derivative’ of standard Brownian motion in the sense that $DW(t)dt = dW(t)$. This white noise plus drift model is “asymptotically equivalent” to nonparametric regression meaning that the distribution of y_1, \dots, y_n converges to that of $y(t), t \in \mathcal{I}$, in a metric due to Le Cam (Brown and Low, 1996; Brown, Cai, Low, and Zhang, 2002). Hall and Opsomer (2005) use an idealized version of a penalized spline where there is a continuum of knots, that is, their spline is $\int_{\mathcal{I}} \beta(s)\rho(s)(x-s)_+^p ds$ where $\beta(s)$ is the spline coefficient at knot s , $\rho(s)$ is the knot density, and $(x-s)_+^p$ is the spline basis function with knot s . In this framework, they find asymptotic expressions for the bias and the stochastic part of the penalized spline estimator. These expressions are infinite series with terms depending on the eigenvalues and eigenvectors of a certain functional operator. They show that the mean integrated squared error, which is $\int_{\mathcal{I}} E\{\hat{m}(x) - m(x)\}^2 dx$, is $O(n\lambda^{1/(2p+2)} + \lambda)$, where n is the sample size, p is the degree of the spline and λ is the penalty parameter. Therefore, if λ is a constant multiple of $n^{2(p+1)/(2p+3)}$, then the mean integrated squared error is $O(n^{2(p+1)/(2p+3)})$, which is the optimal rate for functions with $p+1$ square-integrable derivatives (Stone, 1982).

Penalized spline asymptotics with a finite but increasing number of knots can be divided into two cases, depending on the rate at which the number of knots K increases with the sample size n . “Small- K ” asymptotics are similar to those of ordinary least squares regression splines, that is, splines fit by least squares without a penalty. “Large- K ” asymptotics are similar to that of smoothing splines. The bias of a penalized spline has two components, the approximation (or modeling) bias and the shrinkage (or smoothing) bias. Approximation bias is the bias of an ordinary least squares regression spline and is due solely to the approximation of the regression function by a spline. Shrinkage

bias is the difference between the bias of the penalized spline and the approximation bias and so is the additional bias due to the penalty. Under large- K asymptotics, the approximation bias is negligible compared to the shrinkage bias. Hall and Opsomer's (2005) framework is an extreme case of large- K asymptotics. Under small- K asymptotics, the approximation bias converges to zero at the same rate or more slowly than the smoothing bias. The approximation bias is controlled by K , and under small- K asymptotics K is a smoothing parameter. Under large- K asymptotics, the approximation bias is negligible (or exactly zero in Hall & Opsomer's case), the exact value of K has no effect on the asymptotic distribution (provided only that K grows fast enough to be in the large- K case), and the penalty parameter λ is the only smoothing parameter.

We believe that large- K asymptotics are the most relevant to current practice. The original penalized spline methodology proposed by Eilers & Marx (1996) assume that the number of knots is sufficiently large that approximation bias is negligible compared to smoothing bias. Numerical evidence in Ruppert (2002) supports this assumption, as does the current practice of using the data to carefully select the penalty parameter while using some rule of thumb applied to the sample size to select the number of knots. Moreover, under small- K asymptotics, one needs to use a data-based method to select K . One might also need to be careful about the locations as well as the number of knots. These issues have not been investigated, except in the case of pure regression splines with no roughness penalty where both the number and locations of the knots are chosen (Smith & Kohn, 1996; Denison, Mallick & Smith, 1998; Dimatteo, Genovese & Kass, 2001) and the hybrid adaptive splines of Luo & Wahba (1997) that use both adaptive knot selection and a roughness penalty.

Li & Ruppert (2008) use large- K asymptotics. They study Eilers and Marx's (1996) P-splines which involve B-splines and difference penalties on the spline coefficients. Li & Ruppert (2008) obtain simple, explicit expressions for the asymptotic bias and variance. This allows asymptotic distributions of P-splines can be compared with those of kernel regression, local polynomial regression, and smoothing splines. Their results are restricted to the cases of zero-degree or linear splines and a first or second order difference penalty. We say that a penalty is of q th order if the penalty is on the q th derivative (O-splines) or the q th finite difference (P-splines). O-splines (Wand & Ormerod, 2008) use the same penalty as used by smoothing splines, but are similar to the P-splines of Eilers & Marx (1996) in that a reduced set of knots is used.

In a nutshell, Li & Ruppert (2008) found that P-splines with a q th order penalty are asymptotically equivalent to Nadaraya-Watson kernel regression estimators with the equivalent kernel found by Silverman (1984) for smoothing splines with a q th order penalty. The asymptotic distribution does not depend on the degree p , but the minimum rate at which K must increase does depend on p and the rate is slower if higher degree splines are used. First consider the case where the x s are equally-spaced on a finite interval. For first-order penalties, the equivalent kernel in the interior is the double-exponential kernel, which is second order. This is the equivalent kernel for smoothing splines with a first-order penalty (Silverman, 1984). If the penalty parameter λ is chosen at the optimal rate, then the equivalent bandwidth h satisfies $\lambda \sim \{Khn^{-1/5}\}^2$. The asymptotic bias at an interior point x is $\mathcal{B}(x) = h^2 f^{(2)}(x)$, the asymptotic variance is $\mathcal{V}(x) = 4^{-1}h^{-1}\sigma^2(x)$ where $\sigma^2(x)$ is conditional noise variance at x , and $n^{2/5}\{\hat{f}(x) - f(x)\} \rightarrow N(\mathcal{B}(x), \mathcal{V}(x))$ in distribution as $n \rightarrow \infty$. In the boundary regions, which consists of the points within a multiple of $n^{-1/5}$ of the left or right boundaries, the equivalent kernel is of first-order. At the boundaries themselves, the equivalent kernel is an exponential function. If the penalty parameter is chosen to be optimal for the interior, then in the boundary region bias dominates and the convergence rate is $O(n^{-1/5})$; the same as for a Nadaraya-Watson kernel estimator.

For second-order penalties, Li & Ruppert (2008) find that the equivalent kernel in the interior is fourth-order and proportional to $\exp(-|x|)\{\cos(x) + \sin(|x|)\}$, which is

the equivalent kernel for cubic smoothing splines (Silverman, 1984), which also have a second-order penalty. The rate of convergence in the interior is $n^{-4/9}$, which is the same as for a Nadaraya-Watson kernel estimator with fourth-order kernel. In the boundary region the equivalent kernel is only second-order and the rate of convergence is slower.

The results so far assume equally-spaced knots. Li & Ruppert (2008) also study unequally-spaced knots. In this case, the asymptotic bias depends on derivatives of the design density, which means that penalized splines are not “design-adaptive” in the sense of Fan (1992).

Penalized splines have slower convergence at the boundaries than in the interior, whereas local polynomial regression with odd degree polynomials has the same rate of convergence at the boundaries as in the interior. This might seem like an advantage of local polynomial smoothing compared to penalized splines. However, if we compare the widely-used local linear regression smoother with penalized splines with the typical second-order penalty, then what we find is that the local polynomial and penalized spline smoothers have the same boundary rate of convergence. In the interior, the penalized spline has a *faster* rate of convergence. Thus, as typically implemented in practice, penalized splines suffer no disadvantage in rate of convergence relative to local linear estimators and, in fact, have an advantage in the interior region.

Kauermann, Krivobokova & Fahrmeir (2009) study small- K asymptotics for generalized spline modeling, that is, with possibly non-Gaussian responses and a link function relating the expected response to the spline. They put an upper bound on the rate at which the smoothing parameter increases and K is required to grow at a fixed rate, rather than faster than this rate as in Li & Ruppert (2008). The framework is the generalized linear mixed model, and Laplace approximation is used to integrate out the random effects. The authors obtain rates of convergence for the mean squared error and expressions for the asymptotic bias and variance.

Kauermann (2005) is a comparison of the REML and C_p methods of selecting the amount of smoothing for univariate penalized splines. Both asymptotic and finite-sample simulation results are presented. The asymptotics assume that K is fixed. One general conclusion is that REML tends to under-smooth in that REML leads to less smoothing than optimal for minimizing mean squared error. In contrast, C_p targets the mean squared error-optimal amount of smoothing. However, a more detailed look at Kauermann’s results show that REML and C_p have very different behaviors and which one smooths most depends on the underlying regression function, the number of knots, the sample size, and the random sample itself. One advantage of REML can be seen in Kauermann’s (2005) results: the REML choice of the amount of smoothing is less, and often far less, variable compared to that of C_p .

As described more fully in Section 3.1, Bühlmann & Yu (2003) derive asymptotics for boosting in a nonparametric regression context.

Our view of asymptotic theory is that, at least at present, it is mainly of theoretical interest. Penalized spline methodology already had a well-established place in practice before the recent advances in large-sample theory and we have not yet seen cases where asymptotic theory has led to new methodology or changes in practice. It is well-known that, in nonparametric estimation, it often takes extremely large sample sizes before the asymptotics “kick in”. So to be of practical value, asymptotics must be carefully compared with finite-sample results, either exact or by simulation. Nonetheless, asymptotics are important because they show that low-rank penalized splines can achieve the same rates of convergence as full-rank estimators such as smoothing splines.

2.7 Software

Semiparametric regression research is now being conducted at a time of rapid change in computing technology. In particular, the Internet age now facilitates fast and conve-

nient dissemination of code. Software for semiparametric regression is continually being added to the Comprehensive R Archive Network (CRAN) (<http://cran.r-project.org>) allowing free widespread use for anyone who ‘speaks’ R (R Development Core Team, 2008). Developments in commercial packages are also afoot. For example, SAS (SAS Institute, Incorporated, 2007) added PROC GAM for generalized additive model analyses in 2000.

Generalized additive model analysis in R is now well-served by the packages `gam` (Hastie, 2006), `mgcv` (Wood, 2008) and `VGAM` (Yee, 2008). The `mgcv` package is accompanied by the book Wood (2006a), which contains numerous illustrations of its use. It also provides for automatic selection of degrees of freedom values via GCV. The `VGAM` package distinguishes itself by facilitating the ‘vector’ extension of generalized additive models (Yee & Wild, 1996) and now provides for quantile regression (Yee, 2004). Additive semiparametric quantile regression is also available in R’s `quantreg` package (Koenker, 2008).

In Ruppert *et al.* (2003) we mentioned `SemiPar` as a suite of S-PLUS functions to accompany the book’s mixed model-based methodology. It has evolved into a package on CRAN (Wand *et al.* 2007). Other packages with direct links to semiparametric regression include `AdapFit` (Krivobokova, 2007) on spatial adaptive smoothing and `polyspline` (Kooperberg, 2007) on regression spline fitting.

BayesX is a public domain software package that supports Bayesian semiparametric regression analysis using MCMC. It is housed in the Department of Statistics, University of Munich, Germany, and its current Internet address is www.stat.uni-muenchen.de/~bayesx/. Brezger, Kneib & Lang (2005) provides an overview of the capabilities of BayesX. They also demonstrate superior mixing and speed of their MCMC implementations in comparison to WinBUGS.

Several other software modules indirectly benefit semiparametric regression analysis through their support of related methodology such as geostatistics, kernel machines and mixed models. While they exist in a variety of forms, we will mainly confine discussion to those available on CRAN.

The geostatistical packages `fields` (Nychka, 2007), `geoR` (Ribeiro & Diggle, 2008), `geoRglm` (Christensen & Ribeiro, 2008) and `spectralGP` (Paciorek, 2007a, 2007b) each support bivariate smoothing. There is also some support for smoothing in higher dimensions. For example, the `Tps()` function of `fields` allows thin plate spline smoothing of arbitrary dimension.

As we explain later in Section 3.2 kernel machines have fundamental connections with semiparametric regression. The R packages `e1071` (Dimitriadou *et al.* 2008) and `kernlab` (Karatzoglou *et al.* 2007) provide for kernel machine fitting, including support vector machines.

As demonstrated by Ngo & Wand (2004), mixed model software can be very useful for semiparametric regression analysis. A key feature is the support of general random effects design matrices (Zhao *et al.* 2006). The SAS procedure PROC MIXED and the R package `nlme` (Pinheiro *et al.* 2008), each support general design matrices. The function `glmmPQL()` in the package `MASS` (Venables & Ripley, 2008) has structure similar to that of `lme()` and `lme4()` and facilitates generalized response semiparametric regression analyses via penalized quasi-likelihood. In `lmeSplines` (Ball, 2008) mixed model-based splines are the main focus. Exact likelihood ratio tests for semiparametric regression analysis, as discussed in Section 3.6, is supported by the R package `RLRsim` (Scheipl, 2007).

As we discussed in Section 2.3, practical Bayesian inference has benefited enormously from the BUGS software project (Lunn *et al.*, 2000). The employment of BUGS is currently the fastest way to get hierarchical Bayesian models fitted – or at least proto-typed. Ruppert *et al.* (2003) and Crainiceanu, Ruppert & Wand (2005) demonstrate the use of BUGS for Bayesian semiparametric regression analysis. A brief example, which incorporates

the variance component prior recommendations of Gelman (2006), is the Bayesian logistic nonparametric regression model given at (9). Figure 2 provides a graphical representation of this model. Implementation in BUGS involves the model specification code:

```

model
{
  for(i in 1:n)
  {
    logit(mu[i]) <- beta0 + beta1*x[i] + inprod(u[],Z[i,])
    y[i] ~ dbern(mu[i])
  }
  for (k in 1:K)
  {
    u[k] ~ dnorm(0,tauU)
  }
  beta0 ~ dnorm(0,tauBeta) ; beta1 ~ dnorm(0,tauBeta)
  numerU ~ dnorm(0,1) ; denomU ~ dnorm(0,tauA)
  tauU <- pow(numerU/denomU,2)
}

```

where `tauBeta` and `tauA` are the reciprocals of the hyper-parameters σ_β^2 and A^2 . WinBUGS, the most popular version BUGS, can generate samples from posteriors of interest from the above code via a graphical user interface. However, a major breakthrough for efficient and well-managed analyses is the R package `BRugs` (Thomas, O’Hara, Ligges and Sturtz, 2006; Ligges *et al.* 2007), which was first released in 2005, and its predecessor `R2WinBUGS` (Sturtz, Ligges and Gelman, 2005; Sturtz *et al.* 2007), first released in 2004. These packages allow for a single R script to (1) set up the data, spline basis functions, and various tweaking factors; (2) write a BUGS script and call BUGS; and then (3) produce summaries of interest using the vast graphical capabilities of R. These important facets are not available if WinBUGS is used alone. Crainiceanu *et al.* (2005) illustrated this approach with `R2WinBUGS`. However, our most recent Bayesian semiparametric regression work (as yet unpublished) has employed `BRugs`.

3 Advancement of Models and Methods

After reviewing the semiparametric regression literature from 2003–2007 we then categorized the various contributions according to broad themes, with regarding to advancement of models and methods. The following subsections emerged, and are presented alphabetically.

3.1 Boosting

Boosting uses an ensemble of classification or regression fits as a means of improving their performance. The ensemble elements are obtained iteratively, after application of standard procedures to weighted versions of the data. Boosting began within the field of machine learning during the 1990s. Early references are Schapire (1990), Freund (1995) and Freund & Schapire (1996). Far-reaching statistical connections, involving gradient descent methods and additive models, were discovered by Breiman (1998) and Friedman, Hastie & Tibshirani (2000). These acted as catalysts for a great deal of statistical research on boosting, including its interplay with smoothing techniques. Section 2.1 of Tutz & Binder (2006) describes the evolution from boosting as a means to improve classification procedures to a powerful tool for semiparametric regression analysis. We now describe some of these developments.

Bühlmann & Yu (2003) provides an excellent introduction to the main ideas of boosting, by working with linear smoothers and the simplest version of boosting, known as

L_2 boosting. Let \mathbf{y} be the response vector and $\hat{\mathbf{y}}_\lambda$ be the vector of fitted values obtained from a linear smoother (e.g. penalized spline, kernel-based local linear) with smoothing parameter λ . Then the smoother matrix \mathbf{S}_λ is given by $\hat{\mathbf{y}} = \mathbf{S}_\lambda \mathbf{y}$. The L_2 boosting fit at iteration m is one with fitted values $\hat{\mathbf{y}}_{\lambda,m} = \mathbf{I} - (\mathbf{I} - \mathbf{S}_\lambda)^{m+1}$. The case $m = 1$ corresponds to the ‘twicing’ methodology of Tukey (1977). Boosting, in general, involves repeated fitting of ‘weak’ classification or regression procedures. Bühlmann & Yu use asymptotics to explain a new type of bias-variance trade-off that arises from the interplay between m and λ . A very interesting result is that, for optimal values of m , the optimal smoothing parameter is larger than for the ordinary ($m = 0$) case. This is consistent with the boosting ‘folklore’ which says that iteration of weak procedures leads to better performance. For linear smoothers, ‘weakness’ can also be achieved by replacing \mathbf{S} by $\mathbf{S}_{\lambda,\nu} = \nu \mathbf{S}_\lambda$, $0 < \nu \leq 1$. Bühlmann & Yu (2003) provide simulation results that show very small ν can be optimal and quite large m can be optimal. One example has $(\nu, m) = (0.01, 1691)$ as the optimal configuration, showing how slow convergence in boosting can be.

Tutz & Reithinger (2007) integrated the ideas of boosting with semiparametric mixed models based on penalized splines. Their `BoostMixed` algorithm works with weak versions of the smoothers obtained by inflation of the smoothing parameters. Versions of AIC and BIC are used as stopping criteria. Liutenstorfer & Tutz (2007a) use boosting for knot selection in a regression spline approach to smoothing.

The fitting of generalized additive models via likelihood-based boosting is developed by Tutz & Binder (2006), resulting in their `GAMBoost` algorithm. Advantages are found in the case of very many predictors. Binder & Tutz (2008) use a large-scale simulation study to show that `GAMBoost` compares favorably with other methods for fitting generalized additive models where there are many candidate predictors.

A comprehensive account of the statistical aspects of boosting is provided by Bühlmann & Hothorn (2007) and accompanying discussion. Important contributions include connections to smoothing splines and the lasso, asymptotic theory, degrees of freedom and implementation in the R computing environment.

As discussed in Section 3.16, monotone smoothing with boosting is developed in Tutz & Liutenstorfer (2007). Liutenstorfer & Tutz (2007b) apply this methodology to air pollution data.

3.2 Connections with kernel machines

Given data $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathbb{R}$, $1 \leq i \leq n$, and a convex loss function \mathcal{L} , kernel machine estimation of $f = \operatorname{argmin}_g E\{\mathcal{L}(\mathbf{y}, g(\mathbf{x}))\}$ involves modeling f to be of the form $f(\mathbf{x}) = b + \sum_{i=1}^n c_i K(\mathbf{x}, \mathbf{x}_i)$, where the kernel $K(\mathbf{s}, \mathbf{t})$ is a symmetric positive definite bivariate function on $\mathcal{X} \times \mathcal{X}$, and obtaining the coefficients according to

$$\min_{b, \mathbf{c}} \{\mathcal{L}(\mathbf{y}, \mathbf{1}b + \mathbf{K}\mathbf{c}) + \lambda \mathbf{c}^T \mathbf{K}\mathbf{c}\}. \quad (10)$$

Here \mathbf{y} and \mathbf{c} are the vectors containing the y_i and c_i , $\lambda > 0$ is the regularization parameter, $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]$ is the Gram matrix and $\mathbf{1}$ is a vector of ones of length n . There are several ways by which (10) can be derived; including reproducing kernel Hilbert space projection theory (e.g. Kimeldorf & Wahba, 1971), best linear prediction of stationary spatial processes (e.g. Stein, 1999), maximum a posterior estimation in Gaussian processes (e.g. Rasmussen & Williams, 2006) and Tikhonov regularization of ill-posed problems (Tarantola, 2005). Support vector machines (e.g. Cristianini & Shawe-Taylor, 2000) are a special type of kernel machine, in which $\mathcal{L}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^n (1 - a_i b_i)_+$ for $y_i \in \{-1, 1\}$.

Hastie & Zhu (2006) show that kernel machine methods, such as support vector machines for classification, are no different in substance from many statistical methods involving penalization. Their second section provides some revealing connections via the use of spectral decomposition of the Gram matrix of kernel machines.

Pearce & Wand (2006) elucidate connections between the penalized spline and kernel machine literatures. Particular attention is paid to support vector machines. Computational aspects of the resulting penalized spline support vector classifiers are studied by Ormerod, Wand and Koch (2008).

Takeuchi, Le, Sears & Smola (2006) exemplifies research from the machine learning community on nonparametric regression problems. They tackle the nonparametric quantile regression problem using kernel machines. Included are solutions to the quantile crossing problem and incorporation of monotonicity constraints.

Gianola, Fernando & Stella (2006) combine the ideas of linear mixed models and kernel machines to predict total genetic value for quantitative traits. Random effects are used for genetic effects, while kernel machines are used for expression of single-nucleotide polymorphisms. Liu, Lin & Ghosh (2007) derived similar models, with kernel machines used to handle interactions between expression of several genes. They conclude with some interesting commentary on further opportunities for the use of kernel machine methodology in biostatistical research.

3.3 Epidemiological aspects

In our view, epidemiology is an area for which semiparametric regression has a lot to offer and this is reflected in much of our own research. We now review recent semiparametric regression research having epidemiological aspects.

Kim, Carroll & Cohen (2003) take a penalized spline/mixed model approach to generalized additive model analysis in matched case-control studies. They develop an approximate cross-validation scheme to choose the smoothing parameters and explored both Monte Carlo EM and Bayesian approaches to fitting.

The methodology of Carroll, Ruppert, Crainiceanu, Tosteson & Karagas (2004) (see Section 3.10) is applied to the OPEN (Observing Protein and ENergy intake) nutritional epidemiological study. Doubly labeled water, a biomarker for nutrient intake, is used as a instrument in a nonparametric regression measurement error model that relates true protein intake with that reported via a food frequency questionnaire.

Dominici, McDermott & Hastie (2004) work with smoothing spline-based Poisson additive models to assess the effect of particulate matter air pollution on mortality. The data are daily time series with smooth function components to account for seasonal and meteorological effects. Improved inferential techniques leads to strong evidence of association between short-term exposure to particulate matter less than 10 microns in diameter (PM_{10}) and mortality.

Congdon (2006), MacNab (2007) and MacNab & Gustafson (2007) use semiparametric regression techniques in spatial epidemiological analyses. The first applies the methodology to spatial count data on lip cancer in Scotland and suicide data in London. The second and third of these apply the methodology to spatial count data on adverse medical events to hospitalized children, youth and elderly patients in British Columbia, Canada. Temporal and spatial trends within 84 local health areas are estimated and assessed.

The papers Figueiras, Roca-Pardinas & Cadarso-Suarez (2005), Cadarso-Suarez, Roca-Pardinas & Figueiras (2006) and Roca-Pardinas, Cadarso-Suarez, Nacher & Acuna (2006) are motivated by Spanish epidemiological studies and make use of semiparametric regression methodology in various ways. For example, Figueiras *et al.* (2005) uses a Poisson to assess the effect of black smoke on mortality in Vigo, Spain.

3.4 Functional data analysis

Functional data analysis is concerned with data that are collected at very fine gradations in time or space. It is sometimes referred to colloquially as ‘curves as data’ and, from the outset, has had strong connections with nonparametric regression techniques. Ramsay &

Silverman (1996, 2002) provide a solid foundation for this relatively new field. Functional data analysis is also a close relative of longitudinal data analysis. It is not surprising that some recent work in functional data analysis makes use of modern semiparametric regression methodology.

Cardot, Ferraty & Sarda (2003) consider functional linear models, where the predictor is a random function. They consider approaches involving penalized B-splines and smooth principal components regression and establish L_2 rates of convergence for each.

Coull & Staudenmayer (2004) take a linear mixed model approach to self-modeling regression for multiple response curve data. An Expectation-Conditional Maximization algorithm (Meng & Rubin, 1993) is developed for fitting and inference. Application is made to data on the respiratory effects of residual oil fly ash inhalation in humans.

Motivated by functional data on particulate matter exposure and heart-rate variability, Harezlak, Coull, Laird, Magari & Christiani (2006) extends historical functional linear models (Malfait & Ramsay, 2003) in the direction of mixed model-based splines with REML smoothing parameter selection. L_1 penalties, with AIC smoothing parameters selection, are considered as well.

Qin & Guo (2006) build periodicity into functional mixed-effects models to better model the circadian rhythms of cortisol concentrations. They develop a state space representation of periodic splines and use Kalman filtering for estimation.

Morris, Vannucci, Brown & Carroll (2003) make the initial step of wavelet-based non-parametric modeling on hierarchical data, using Bayesian fitting methods. Morris & Carroll (2006) introduce the notion of wavelet-based functional mixed model. Regularization and smoothing are done within the Bayesian paradigm, with easy-to-use code available at odin.mdacc.tmc.edu/~jeffmo/papers_files/wfmm_supplement.html. Their methods are applied to functional mixed models data of the MGMT DNA repair enzyme in a colon carcinogenesis experiment. Morris *et al.* (2006) extend this method to allow for missing response data and apply it to an accelerometer profile study. Morris, Brown, Herrick, Baggerly & Coombes (2007) use the wavelet-based functional mixed model to analyze mass spectrometry data. Antoniadis & Sapatinas (2007) also work with wavelet-based functional mixed models. Recent work on likelihood ratio testing for penalized splines (e.g. Crainiceanu *et al.* 2005; see Section 3.6) is employed. Risk bounds are established and the methodology is applied to stepping-cycle data from an orthosis study.

Marx & Eilers (2005) extend their earlier work on penalized signal regression (Marx & Eilers, 1999) to two-dimensional signal regressors. Their example of such a regressor involves digitizations along the emission wavelength axis of curves arising from a sugar processing experiment. The second dimension arises from these digitizations being done at several excitation levels. The prediction of ash content and color is of interest. The regression fitting and modeling involves tensor product extensions of P-splines and cross-validation. These leads to an estimated coefficient surface, and an image of 't-like' statistics over the wavelength/excitation plane. In Marx & Eilers (2002) and Eilers & Marx (2003) the authors apply their general approach to other chemometrics data sets, with some tailoring to the problems at hand.

Reiss & Ogden (2007) also treat the signal regression problem. They start by pointing out that there two main approaches to dealing with the multicollinearity problem: smoothing (e.g. Marx & Eilers, 1999) and component selection (e.g. Massy, 1965). They develop functional versions of principal component regression and partial least squares, which combine these two approaches. Selection of the smoothing parameter is studied in depth. Both GCV and REML are considered, the latter arising from a linear mixed model representation of their procedures. Their simulation results show good performance of REML.

Yao & Lee (2006) treat the functional principal component analysis problem (e.g. Ramsay & Silverman, 1997, Chapter 8) using an iterative penalized spline procedure that addresses within-subject correlation in functional data. Consistency results are estab-

lished and application is made to yeast cell cycle gene expression data. Zhou, Huang & Carroll (2008) use a novel low-rank principal components approach to address joint modeling of bivariate functional data and show that a seemingly unrelated regression phenomenon exists.

Baladandayuthapani, Mallick, Hong, Lupton, Turner & Carroll (2008) develop an elaborate hierarchical functional data analytic model for data arising from a colon carcinogenesis study. It is tailored to suit the colonic crypt structure of rats. Bayesian representations of penalized splines are used to model signals as a function of distance within a crypt, while the Matérn covariance family is used to model correlation of signals between the crypts.

3.5 Geoadditive models

Geoadditive models combine the ideas of geostatistics and additive models. An example of an geoadditive model is

$$E(\text{birthweight}_i) = f_1(\text{no. prenatal visits}_i) + f_2(\text{cigarettes per day}_i) + f_3(\text{longitude}_i, \text{latitude}_i)$$

Kammann & Wand (2003) show how linear mixed models could be used for geoadditive model fitting and inference. However, several other papers (e.g. Wood, 2003) have treated the same structure in other ways.

Extensions of geoadditive models in the direction of generalized responses are contained in Fahrmeir & Echavarría (2006) and Zhao, Staudenmayer, Coull & Wand (2006). Zhao *et al.* (2006) deal with exponential family models, whilst Fahrmeir & Echavarría (2006) treat over-dispersed and zero-inflated count data. Each use a Bayesian mixed model framework, with fitting via MCMC, and provide applications.

The extension of geoadditive models to survival data has seen considerable research since 2003. Hennerfeind, Brezger & Fahrmeir (2006) develop geoadditive survival models for both geographical point data and count data. They take a Bayesian P-spline approach and use Gaussian and Markov random fields for the spatial components. Kneib & Fahrmeir (2007) lays out the mathematics underpinning geoadditive hazard regression models. Kneib (2006) extends these models to handle interval censored data. Adebayo & Fahrmeir (2005) develop a geoadditive discrete-time survival model and use it to analyze child mortality data. Ganguli & Wand (2006) also deal with geo-referenced survival data, and use the low-rank radial smoothers of Kammann & Wand (2003).

Geoadditive models have also been adapted to model space-time data. Fahrmeir, Kneib & Lang (2004) and Kneib & Fahrmeir (2006) use low-dimensional smooths involving time and age to model forest health data, in conjunction with Gaussian and Markov random fields for the spatial effects. Gryparis, Coull, Schwartz & Suh (2007) also involves space-time data, but their geoadditive model is an elaborate one that includes latent variable structure for multiple exposures from mobile particulate matter.

Geoadditive models with missing data covariate data is studied by French & Wand (2004). Chen & Ibrahim (2006) extend that work to geoadditive models that allow for specification of the covariate distribution and the missing data mechanism.

Other work that contains extensions of geoadditive models includes Lang, Adebayo, Fahrmeir & Steiner (2003), on seemingly unrelated regression, Lang & Brezger (2004) on spatial adaptation and Augustin, Lang, Musio & von Wilpert (2007) on ordered categorical responses.

3.6 Inference

In its early years, smoothing techniques were developed with little regard to related inferential questions such as linearity versus non-linearity of a particular covariate effect.

This is especially noticeable in the early kernel smoothing literature. In recent years, however, this situation has been redressed and there is now quite a large literature on inference in smoothing contexts. The mixed model representation of smoothing splines and penalized splines offers a particularly attractive framework for this endeavor. This is because the well-established tools of likelihood-based and Bayesian inference are readily available. While there is a great deal of research on inference for other approaches to smoothing since 2002, we confine discussion mainly to smoothers based on mixed models.

A significant portion of the 2003–2007 literature involves likelihood ratio tests for testing departures from linear models. This boils down to tests on variance components being different from zero. The classical reference for tests of this type, in which the null value of the parameter is on the boundary of its space, are Self & Liang (1987) and Stram & Lee (1994). However, their theory assumes that independence under the null and alternative hypotheses. This is not the case for many mixed model scenarios, including penalized splines and several recent papers by C. Crainiceanu and co-authors are concerned with rectifying this situation. The main smoothing paper from this body of work is Crainiceanu, Ruppert, Claeskens & Wand (2005). It builds upon Crainiceanu & Ruppert (2004a), where exact distribution theory for the likelihood ratio statistic in Gaussian linear mixed models is obtained. Crainiceanu *et al.* (2005) also obtain confidence intervals for the smoothing parameter by inverting likelihood ratio tests. Claeskens (2004) contains asymptotic results for this setting, but with the number of knots increasing with the sample size and certain restrictions on the design matrices that are not satisfied by standard penalized spline models. Crainiceanu and Ruppert (2004b) develop likelihood ratio and restricted likelihood ratio tests of goodness-of-fit of nonlinear regression models. Liu & Wang (2004) review various versions of linearity tests based on Bayesian representations of smoothing splines and conduct a simulation study to assess their frequentist properties.

The exact distribution theory used in the papers of the previous paragraph applies only to the situation where there is a single variance component. Extensions to models with multiple covariance components is conducted by Crainiceanu & Ruppert (2004c) and Greven, Crainiceanu, Kuechenhoff & Peters (2008). Remedies to the null distribution problem include use of the parametric bootstrap and approximation of the likelihood ratio statistic by a product of independent χ_1^2 and Bernoulli random variables. Greven *et al.* (2008) demonstrate good performance of the second approach and also propose an approximation to the null distribution of the restricted likelihood ratio statistic using an idea similar to pseudo-likelihood estimation. Crainiceanu, Ruppert, Claeskens and Wand (2005) show via simulation studies that the power properties of the likelihood ratio tests compare favorably those of competing tests.

In the context of least-squares kernel machines, Liu, Lin & Ghosh (2007) develop a score test for non-linearity that relies on a mixed model representation. Satterthwaite's approximation is used to obtain approximate p-values.

Extension of likelihood ratio tests in the generalized response setting is challenging due to the presence of intractable integrals in the likelihoods. Lin (1997) and Lin & Zhang (1999) developed score tests for GLMM settings, the latter reference including generalized additive models through the mixed model representation of smoothing splines. This general approach has since been extended to additive mixed models (Zhang & Lin, 2003), varying coefficient models for longitudinal data (Zhang, 2004) and proportional hazards models (Lin, Zhang & Davidian, 2006).

Wood (2006b) develops approximate Bayesian confidence intervals (see Section 6.4 of Ruppert *et al.* and references given there) for the estimated functions in generalized additive models. He takes advantage of the low-rank aspect of penalized splines so that the distribution theory involves the relatively small random vector of spline basis coefficients. The generalized case is dealt with by using a weight matrix approximation in

the ridge regression expression. It is also explained how inference for functionals of the coefficient vector can be made without time-consuming bootstrap replications. This innovative paper finishes off with proposals on how to avoid MCMC in the ‘fully Bayesian’ case, in which variability due to smoothing parameter choice variance components is taken into account.

The Bayesian mixed model approach to semiparametric regression has immediate benefits regarding inference. For example, non-linearity versus linearity of covariate effects can be assessed through the posterior distributions of variance components. They are several new papers on Bayesian on semiparametric regression, scattered throughout Section 3 of this review article.

Lastly, we mention contribution spline-based approaches to the scale-space approach to feature significance, sometimes known as ‘SiZer’ (Chaudhuri & Marron, 1999), and summarized in Section 6.9 of Ruppert *et al.* (2003). Ganguli & Wand (2004) facilitates feature significance for bivariate smoothing, or geostatistics, by developing the appropriate theory for low-rank thin plate splines. Marron & Zhang (2005) develop the requisite theory for a (full-rank) smoothing spline version of SiZer.

3.7 Latent variable models

The introduction of Skrondal & Rabe-Hesketh (2004) defines a latent variable as a random variable whose realizations are hidden from the analyst and gives, as examples of their utility, data measured with error, hypothetical constructs and latent responses underlying categorical variables. Mixed models play a prominent role in latent variable modeling so, for this reason alone, have common ground with contemporary smoothing techniques. Latent variable modeling is a growth area in Statistics in general, and has had some interplay with semiparametric regression in the last five years.

Tutz & Scholz (2004) use the *principle of maximum random utility* to link multi-category responses to latent utilities. They allow for dependence on covariates via additive and varying coefficient structure, aided by penalized splines. Fahrmeir & Raach (2007) develop Bayesian semiparametric latent variable models, including those that allow spatial effects to be incorporated. They involve measurement models for mixed continuous, binary and ordinal responses. For example, the discrete value of ordinal responses are assumed to be generated through a threshold mechanism.

Elliott (2007) uses smoothing splines and their mixed model representation to build flexibility into latent cluster models. These relate underlying ‘clusters’ of variability to measures of interest. Application is made to data on depression levels for patients recovering from myocardial infarction.

Gryparis, Coull, Schwartz & Suh (2007), described more fully in Section 3.5, has a latent variables aspect for handling multiple exposures.

3.8 Longitudinal data analysis

Mixed models have been a staple of longitudinal data analysis for the last 25 years. This partnership has resulted in a high volume of mixed model methodology and software development over the same time period. The mixed model approach to penalized spline smoothing not only allows one to take advantage of these developments, but means that longitudinal structure is easy to incorporate. Nowadays, a single linear mixed model can be used to perform an elaborate longitudinal data analysis that incorporates nonparametric estimation of several smooth functions (e.g. Zhao *et al.*, 2006).

An component of recent semiparametric longitudinal data analytic research has been concerned with marginally specified models such as (11). We review this research in Section 3.9. The models covered in this subsection differ in that that are defined conditionally.

Ghidey, Lesaffre & Eilers (2004) develop the penalized Gaussian mixture linear mixed model. It involves function estimation via spline basis functions that are Gaussian densities and random effects modeled as mixtures of normal distributions. Particular attention is paid to two-dimensional random effects structure.

Durbán, Harezlak, Wand & Carroll (2005) describe mixed models for fitting subject-specific curves to longitudinal data. Models of this general type have been developed by several other authors (e.g. Donnelly, Laird & Ware, 1995; Verbyla *et al.* 1999). The low-rank spline approach of Durbán *et al.* (2005) is particularly simple and has the ability to handle very large sample sizes with standard mixed model software (code is included in an appendix). Harezlak, Ryan, Giedd & Lange (2005) fit similar models to data from accelerated longitudinal designs where subjects enter the study at different points of their growth trajectory and are observed over a relatively short time period. Application is made to longitudinal magnetic resonance imaging data from an ongoing developmental study. Smith & Wand (2008) focus on the variance calculations required for inference in semiparametric mixed models. They describe streamlined algorithms that yield two orders of magnitude improvements over naïve variance calculations.

Welham *et al.* (2006) and Zhang *et al.* (2007), as detailed in Section 3.16, deals with semiparametric longitudinal models under periodicity constraints. Zhao *et al.* (2006), discussed in Section 3.13, provides quite a general treatment of Bayesian generalized response models that include longitudinal models as a special case.

The likelihood ratio methodology of Crainiceanu & Ruppert (2004) and Greven *et al.* (2008) (Section 3.6), is applied to inference in longitudinal settings. Qu & Li (2006) develop quadratic inference functions for fitting and inference in varying coefficient models for longitudinal data.

Harezlak, Naumova & Laird (2007) devise a bump hunting test for longitudinal data, based on the subject-specific curves model of Durbán *et al.* (2005).

Finally, we note that more extensive reviews of this subsection’s general topic are provided by five chapters under the heading *Nonparametric and Semiparametric Methods for Longitudinal Data* in Fitzmaurice, Davidian, Verbeke & Molenberghs (2008). The chapters are authored by X. Lin & R.J. Carroll, H.-G. Müller, S.J. Welham and B.A. Brumback, L. Brumback & M.J. Lindstrom.

3.9 Marginal longitudinal models

Research on the marginal longitudinal nonparametric regression model (see (11) below) continues at a steady rate. Early contributions to this setting include Zeger & Diggle (1994) and Lin & Carroll (2000). While most early research involved kernel smoothing, more recent approaches involve spline smoothing. Marginal models differ from the conditionally specified models of Section 3.8 in that they do not model the within-subject correlation or the error process.

The simplest setup is as follows. For $1 \leq i \leq m$ subjects we observe $1 \leq j \leq n$ ($n \ll m$) responses y_{ij} and predictors x_{ij} . (Somewhat annoyingly, the m and n notation alternates in the literature between their roles given here and the reversal; i.e. that where n is the number of subjects and m is the number of measurements. In this paper we stick with the notation used by Diggle, Heagarty, Liang & Zeger (2002) and Ruppert, Wand & Carroll (2003).) Let \mathbf{y}_i be the vector of responses for the i th subject and \mathbf{x}_i be defined similarly. The marginal longitudinal nonparametric regression model is then

$$E(y_{ij}|\mathbf{x}_i) = f(x_{ij}), \quad \text{Cov}(\mathbf{y}_i|\mathbf{x}_i) = \Sigma, \quad 1 \leq i \leq m, 1 \leq j \leq n \quad (11)$$

for some smooth function f and $n \times n$ covariance matrix Σ . A noteworthy, somewhat paradoxical, result is that ordinary kernel smoothers are more efficient if so-called working independence is assumed (Lin & Carroll, 2001). Wang (2003) develops a more elab-

orate kernel smoothing strategy that escapes from this paradox and is uniformly more efficient.

Welsh, Lin & Carroll (2002) use equivalent kernel theory to demonstrate that penalized spline estimators are non-local compared with kernel smoothers. This means that the ‘legality’ of working independence justified by Lin & Carroll (2000) for ordinary kernel smoothers does not apply to penalized splines. Lin, Wang, Welsh & Carroll (2004) brings together earlier papers by the authors on (11). Theoretical results include asymptotic equivalence between the Wang (2003) kernel estimator and a smoothing spline-based estimator, and optimality of these two approaches.

Other contributions to theory and methodology for (11), but primarily within the kernel smoothing realm, include Carroll, Hall, Apanasovich & Lin (2004), Chen & Jin (2005), Hu, Wang & Carroll (2004), Wang, Carroll & Lin (2005) and Lin & Carroll (2006). Interestingly, there is little use of low-rank spline modeling in this context. Linton, Mammen, Lin & Carroll (2003) and Carroll *et al.* (2004) discuss a two-stage approach that estimates Σ from the residuals of an unweighted fit and then computes a penalized spline estimator, finding good efficiency. But to the best of our knowledge the low-rank spline mixed model approach has not been implemented in this context. Current research on this approach, led by the second and third authors, is under way.

3.10 Measurement error models and deconvolution

Carroll, Ruppert, Stefanski & Crainiceanu (2006) provides a recent and comprehensive review of non-linear measurement error models. In their preface to this second edition the authors point out that, in 11 years since the book’s first edition, semiparametric regression and Bayesian computation via MCMC have grown enormously. These threads run through much of the contemporary research on nonlinear measurement error models. Chapters 9, 12 and 13 of Carroll *et al.* (2006) summarize most of the relevant literature. We now supplement those with some recent literature that is closest to the Ruppert, Wand & Carroll (2003) genre.

Carroll, Ruppert, Crainiceanu, Tosteson & Karagas (2004) study non-linear and non-parametric regression when there is covariate measurement error and an instrumental variable is available. They consider several approaches to estimation and, in a simulation study, a Bayesian spline estimator similar to the one in Berry, Carroll, and Ruppert (2002) is the most effective.

Ganguli, Staudenmayer & Wand (2005) studied additive model fitting and inference when measurement error is present in one or more predictors. They use a maximum likelihood approach and advocate use of the Monte Carlo EM algorithm for fitting and inference.

The periodicity-constrained functional mixed models of Zhang, Lin & Sowers (2007) (see Section 3.16) handle measurement error in the predictor, follicle stimulating hormone, via a two-stage approach.

Ma & Carroll (2006) show how to estimate nonparametric functions in semiparametric models while making no assumptions about the distribution of the variables measured with error.

Mallick, Hoffman & Carroll (2002) use a Bayesian approach to fitting nonparametric functions when the measurement errors are mixtures of Berkson and classical types. They use a Dirichlet process to estimate the distribution of the mismeasured covariate essentially nonparametrically. Their method is applied to the Nevada Test Site radiation study. Carroll, Delaigle & Hall (2008) use a deconvolution approach in the same context.

Liang, Wu & Carroll (2003) develop mixed effects varying coefficient measurement error models, applying the methods to AIDS data.

Ruppert, Nettleton & Hwang (2007) use penalized B-splines on a deconvolution problem from multiple testing. Assume the i th hypothesis is $H_{0i} : \delta_i = 0, 1 = 1, \dots, n; \delta_i \geq 0$

might be a non-centrality parameter. The problem is to estimate the distribution of δ_i , but δ_i is not observed. To estimate this distribution, π_0 , the proportion of true nulls and G , the distribution of δ_i under the alternative, are estimated. The estimate $\hat{\pi}_0$ is useful to estimate the false discovery rate (Benjamini and Hochberg, 1995) and the authors show that \hat{G} can be used to estimate the expected discovery rate, the true negative rate, and the true positive rate.

Staudenmayer, Ruppert & Buonaccorsi (2008) study density estimation under heteroskedastic measurement error. Deconvolution methods that assume homoskedasticity over (under) correct in regions where the measurement error variance is smaller (greater) than average. To remedy this problem they introduce a variance function and estimate the density and the variance function as splines.

It is our belief that penalized splines are proving to be a very effective, perhaps the most effective, method for deconvolution and correction for measurement error. The reason is that, when given a Bayesian implementation, they utilize the likelihood and achieve high efficiency. In contrast, earlier more *ad hoc* methods extract less of the information available in the data.

3.11 Missing data

Because of space and time limitations, a missing data chapter was missing from Ruppert *et al.* (2003). However, many contemporary methods for handling missing data use likelihood-based or Bayesian inference that is in keeping with our semiparametric regression methodology. While there has been a modest amount of work in this direction, which we now summarise, our feeling is that there is still room for more such research.

French & Wand (2004) develop a likelihood-based model for missing covariate data in geosadditive model (Kammann & Wand, 2003) analyses. Monte Carlo EM and a version of penalized likelihood is used for fitting and inference. An application involving relative cancer mapping, with missingness in smoking status, is presented.

Chen & Ibrahim (2006) develop likelihood-based semiparametric regression models, including those with bivariate smoothing, for specifying the covariate distribution and the missing data mechanism. The EM algorithm is recommended for fitting, and application is made to data from a melanoma clinical trial.

Bivariate smoothing is also considered by Geraci & Bottai (2006). They treat the incorporation of auxiliary data when there are non-ignorable missing responses. Mixed model-based low-rank kriging is used for bivariate smoothing, and Monte Carlo EM for fitting. Application is made to mapping of phytoplankton data.

Penalized splines are used in a missing data situation with clustering by Yuan & Little (2007). Several missing data mechanisms are entertained. Hierarchical Bayesian models are used and Gibbs sampling employed for fitting. Application is made to a childhood obesity study.

3.12 Model selection

In Ruppert, Wand & Carroll (2003) we noted (Section 8.6) that model selection for semiparametric regression was still in its infancy, and provided a handful of references – particularly in the special case of additive models with several candidate predictors. There have been a few developments since 2003 on this problem.

Wager, Vaida & Kauermann (2007) use the mixed model representation of penalized spline semiparametric regression models and versions of AIC to obtain a model selection algorithm for the continuous response case. The smoothing parameters of the fitted models are estimated from the data using (restricted) maximum likelihood.

Avalos, Grandvalet & Ambroise (2007) work with smoothing splines and the lasso (Tibshirani, 1996) to choose among additive models. The lasso has the feature of annih-

lating coefficients rather than shrinking them, resulting in better parsimony. An approximation of the generalized cross-validation is used for smoothing parameter selection. Vandenhende, Eilers, Ledent & Renard (2007) make use of penalized splines, GCV and the lasso to sift through candidate biomarkers in drug development applications.

Model selection via boosting is studied by Tutz & Binder (2006), Tutz & Reithinger (2007) and Binder & Tutz (2008). Further details are given in Section 3.1.

Hens, Aerts & Molenberghs (2006) is on model selection for incomplete and design-based samples based on a weighted AIC. While it is mainly concerned with parametric regression model, the approach is extendable to semiparametric regression.

3.13 Non-Gaussian response models

Fitting and inference for semiparametric regression models when the response is non-Gaussian usually entails an extra layer of complexity due to the non-explicit forms that arise. Research on this front roughly parallels that of generalized linear mixed models, where analytic approximations and MCMC are the main combatants.

The Wood (2006a) book is a contemporary account of generalized additive model fitting and analysis – accompanied by the R package `mgcv` (Wood, 2008). Generalized additive mixed models are also treated. Smoothing parameter selection is achieved through generalized cross-validation, AIC and penalized quasi-likelihood. Zhao, Staudenmayer, Coull & Wand (2006) work with a general form of the generalized linear mixed model that includes most exponential family semiparametric regression models as a special case. They adopt a Bayesian approach and describe MCMC fitting and inference using BUGS (e.g. Lunn *et al.* 2000) software. Skaug & Fournier (2006) investigate the use of automatic differentiation in a general GLMM framework. A semiparametric regression example is included.

The past few years has seen several extensions of semiparametric regression beyond the one-parameter exponential family situation. Nott (2006) works with the double exponential family (Efron, 1986) and shows it to be a good vehicle for handling both mean and variance functions. He calls upon the Single Component Adaptive Metropolis algorithm of Haario, Saksman & Tamminen (2005) to perform fitting. Branscum, Johnson & Thurmond (2007) extend the Bayesian semiparametric regression approach to responses from the Beta family of distributions. The paper revolves around two applications on household expenditure and foot-and-mouth disease. Houseman, Coull & Shine (2006) and Skaug & Fournier (2006) each include models with negative binomial responses. Tutz (2003) and Tutz & Scholz (2004) develop semiparametric regression models for, respectively, ordinal and multinomial responses.

Another area of recent activity for non-Gaussian semiparametric regression is modeling of sample extremes. Chavez-Demoulin & Davison (2005) develop smoothing spline-based generalized additive models for exceedances-above-threshold data. The penalized likelihood corresponds to the generalized Pareto distribution because of its role as a limiting distribution in this context. Yee & Stephenson (2007) work with sample maxima data and the generalized extreme value distribution and develop vector generalized additive models in this context.

Several other papers, published since 2003, deal with semiparametric regression with non-Gaussian response – but are discussed elsewhere in this review article. Non-Gaussian spatial models are dealt with by Fahrmeir & Echavarría (2006), Augustin, Lang, Musio & von Wilpert (2007), Paciorek (2007c) and Crainiceanu, Diggle & Rowlingson (2008) Zhang & Lin (2003) and Zhang (2004) describe hypothesis testing for variance components in GLMM smoothing contexts. Tutz & Binder (2006) and Binder & Tutz (2008) provide boosting-type procedures for fitting generalized additive models.

3.14 Quantile regression

There have been a few new approaches to semiparametric regression that target quantiles, rather than means and variances. None of these use mixed models or hierarchical Bayesian approaches.

Yee (2004) embeds the LMS (i.e. λ, μ, σ) quantile regression method of Cole & Green (1992) in the vector generalized additive models (VGAM) framework. An improvement to the LMS method, based on the Yeo-Johnson transformation is developed. Non-Gaussian responses, such as those from the Gamma family, are also treated.

Bollaerts, Eilers & Aerts (2006) and Takeuchi, Le, Sears & Smola (2006) each use the ideas of constrained smoothing and 'pinball' loss functions to impose non-crossing in quantile regression. Bollaerts *et al.* (2006) uses P-splines, making it more in keeping with traditional semiparametric regression. Takeuchi *et al.* (2006) use the kernel machine approach which, as mentioned in Sections 2.5 and 3.2, is becoming increasingly intertwined with semiparametric regression research.

Choudhary (2007) used Bayesian penalized splines to estimate a quantile function for the problem of assessing agreement between two measurement methods.

3.15 Sample survey aspects

An interesting development in recent survey sampling estimation research, led by F.J. Breidt and J.D. Opsomer, is the incorporation of nonparametric regression methodology. An early reference is Breidt & Opsomer (2000) on the use of local polynomial regression. However, some more recent contributions have used penalized splines. The first of these is Breidt, Claeskens & Opsomer (2005) where it is stated that, unlike for local polynomial regression, the theory for penalized splines closely follows the established survey linear regression theory. Breidt *et al.* (2005) is concerned with the incorporation of auxiliary covariate information in the design-based estimation of finite population totals in complex surveys. Theorems on design root-n consistency of the penalized spline regression estimator are provided.

The previous article uses the fixed-penalty formulation of the penalized spline and primarily considers inference with respect to the sampling design, as is most commonly done in survey estimation. Other work uses the penalized spline's mixed-model representation to develop model-based estimators for survey data. Zheng and Little (2003) estimate a finite population total by predicting the unobserved part of the population based on a model for the relationship between the variable of interest and the inclusion probabilities. This is extended to the two-stage sampling context in Zheng and Little (2004), and further incorporates a random item response model in Yuan and Little (2007). Opsomer, Claeskens, Ranalli, Kauermann & Breidt (2008) consider spline-based small area estimation, a type of modeling widely used for survey estimation problems but relying almost exclusively on linear mean model specifications (Rao, 2003). Opsomer *et al.* (2008) combine univariate and bivariate penalized splines with the commonly used small area random effects model, and they establish a theorem on the predicted mean squared error properties of the resulting REML-based predictor of the small area means.

Opsomer, Breidt, Moisen & Kauermann (2007) is at the applied end of the spectrum. They describe how design-based estimation of quantities, such as forested area or total wood volume over large regions, can be enhanced through the incorporation of geographic auxiliary information such as elevation and slope and of satellite-derived measurements. Generalized additive models are used to incorporate the auxiliary variables.

A recent review article (Breidt and Opsomer, 2009) on nonparametric and semiparametric estimation methods in complex surveys discusses these methodological developments in more detail, and provides further information on the design-based and model-based modes of inference for surveys.

3.16 Smoothing under constraints

Another area of semiparametric regression that has seen vigorous activity during 2003-2007 is smoothing subject to constraints. The predominant types of constraints in this work are monotonicity, periodicity and quantile non-crossing.

Bollaerts, Eilers & van Mechelen (2006) explain how to build several shape constraints into univariate and multivariate P-spline quantile regression. Ghosh (2007) focus on monotonicity in the binary response regression problem, making use of mixed models and the pooled adjacent violators algorithm, geared towards biomarker evaluation. Tutz & Leitenstorfer (2007) take a boosting approach to enforcing monotonicity. They arrive at two algorithms: `MonBoost` for continuous responses and `GMonBoost` for generalized responses.

Driven by data from longitudinal studies, Welham, Cullis, Kenward & Thompson (2006) and Zhang, Lin & Sowers (2007) impose periodicity constraints on their fitting curves. Welham *et al.* (2006) use the notion of L-splines (e.g. Kimeldorf & Wahba, 1971; Ansley, Kohn & Wong, 1993) in the penalized spline/mixed model set-up, using specifically designed differential operators that annihilate sine and cosine functions. Zhang *et al.* (2007) work with smoothing splines, and also account for measurement error, in work motivated by a hormone study. Eilers, Gampe, Marx & Rau (2008) build periodicity-type constraints into models for data from seasonal incidence tables.

The quantile regression articles of Bollaerts *et al.* (2006) and Takeuchi *et al.* (2006), outlined in Section 3.14, allow for the imposition of monotonicity.

Other constrained smoothing research includes Eilers (2005), in which unimodality is the focus, and Gluhovsky & Vengerov (2007) in which penalized splines are used to do multivariate constrained extrapolation.

3.17 Spatial adaptivity

Each of the main smoothing techniques (e.g. local polynomials, smoothing splines, wavelets) have an accompanying literature on methods by which improved spatial adaptivity can be achieved. The idea is to perform differing amounts of smoothing at different locations and better recover spatially heterogeneous signals. Chapter 17 of Ruppert, Wand & Carroll (2003) describes spatially adaptive extensions of penalized splines. However, there has been some further work in this area.

Lang & Brezger (2004) develop spatially adaptive Bayesian penalized splines for univariate and bivariate smoothing by allowing smoothing parameters to be locally adaptive. Baladandayuthapani, Mallick & Carroll (2005) take a different approach that involves incorporation of a penalized spline estimate of the variance function into the penalty. Extension is made to additive models. Crainiceanu, Ruppert, Carroll, Joshi & Goodner (2007) develop a Bayesian approach to spatially-adaptive penalized splines in the presence of heteroscedastic errors. They combine three spline models: one for the regression function, a second for the logarithm of the locally varying penalty on the regression function, and a third for the logarithm of the variance function. The authors also generalize their model to multivariate smoothing using low-rank thin-plate splines. In Baladandayuthapani *et al.* (2005) and Crainiceanu *et al.* (2007), special Metropolis-Hastings schemes are developed for implementation. Particular attention paid to improved mixing via innocuous model modifications.

Krivobokova, Crainiceanu & Kauermann (2008) use similar models to those used by Baladandayuthapani *et al.* (2005) and Crainiceanu *et al.* (2007). However, they use Laplace approximation rather MCMC and thereby obtain big speed improvements. Non-normal response is also treated. An R package named `AdaptFit` accompanies the paper.

Leitenstorfer & Tutz (2007a) also achieve spatial adaptive via model selection on the knots and a version of boosting.

Paciorek & Schervish (2006) introduce a new class of non-stationary covariance functions for spatial smoothing via Gaussian processes. Non-stationarity essentially equates to spatial adaptivity.

3.18 Spatial and other high-dimensional data

Section 2.2 covers advancement of fundamental principles for multivariate smoothing. In this section we review new semiparametric regression models and methodology that have a multivariate smoothing component. Excluded, however, are geoadaptive models, which are treated in Section 3.5.

Wager, Coull & Lange (2004) develop an approach labeled “mixed model intensity kriging” based on inhomogeneous Poisson spatial processes. A low-rank version of kriging is achieved through Voronoi tessellation of the plane. Application is made to spatial data arising from brain imaging studies.

Sain, Jagtap, Mearns & Nychka (2006) develop a new multivariate spatial model, utilizing splines and mixed models, for soil water profiles. A particularly novel aspect is bivariate smoothing of the *soil-texture triangle* – where the relative proportions of sand, silt and clay are plotted.

Brezger, Fahrmeir & Hennerfeind (2007) use the ideas of Bayesian semiparametric regression in the analysis of functional magnetic resonance imaging data. Space-varying coefficient models are developed, with the goal of improving upon the voxel-by-voxel approaches of earlier functional magnetic resonance imaging papers. Heim, Fahrmeir, Eilers & Marx (2007) apply the same class of models to diffusion tensor images, also arising from magnetic resonance techniques. Penalized splines are used at all stages of a three-step cascade of data processing: voxel-wise regression, smoothing and interpolation.

Dean, Nathoo & Nielsen (2007) use penalized splines as a component of multi-state models for longitudinal panel count data, where the processes corresponding to different subjects may be spatially correlated. Application is made to weevil infestation in white spruce trees.

Crainiceanu, Diggle & Rowlingson (2008) use the binary response version of penalized bivariate splines to model *Loa loa* prevalence in tropical Africa. A Bayesian/MCMC approach to fitting and inference is adopted. A fast method for approximate predictive inference, based on a calibration model, is developed.

Apanasovich *et al.* (2008) investigate low-rank spline smoothing in a spatial context. They use penalized regression splines and develop a novel method for smoothing parameter selection that overcomes the well-known biases of cross-validation with correlated data. Li *et al.* (2007) show how to estimate a correlation function in longitudinal and spatial data. Both papers give applications to colon carcinogenesis experiments.

Several other papers involving spatial data appear elsewhere in this review: Lang & Brezger (2004) and Crainiceanu *et al.* (2007) (Section 3.17), Eilers *et al.* (2008) (Section 3.16), Geraci & Bottai (2006) (Section 3.11), Jank & Shmueli (2007) (Section 3.20) and Currie *et al.* (2004) (Section 3.19)

3.19 Survival analysis

The extension of parametric survival models for survival data to accommodate non-linear covariate and geographical effects continues to be a vibrant area of semiparametric regression research.

Cai, Hyndman & Wand (2002) show how Poisson mixed models and penalized splines facilitate natural and convenient hazard function estimators. Cai & Betensky (2003) extended this approach to hazard regression with interval censored survival data. Time-varying coefficient models of this general type are developed by Tutz & Binder (2004),

Lambert & Eilers (2005), Kauermann & Khomski (2006) and Brown, Kauermann & Ford (2007).

A variety of methods for fitting, inference and smoothing parameter type are proposed. For example, Lambert & Eilers (2005) call upon the Langevin-Hastings algorithm, while Brown *et al.* (2007) develop a hybrid smoothing parameter selector, based on AIC and penalized quasi-likelihood.

Lin, Zhang & Davidian (2006) work with mixed model and spline-based extensions of the proportional hazard model. Score-test tests for the proportional hazards assumption and covariate effects are developed.

Namata *et al.* (2007) develop GLMM-based methodology for current status data, geared towards an infectious diseases application.

Another interesting development is the integration of penalized spline smoothing into actuarial science – as exemplified by Currie, Durbán & Eilers (2004). In this case, the data take the form of mortality tables. The raw mortality table data used here, obtained from a United Kingdom insurance and pensions database, takes the form of two 53×90 matrices corresponding to the calendar years 1947–1999 and males between 11 and 100 years of age. One matrix is number of deaths; the other is number of years lived. The raw hazards matrix is the ratio of the first matrix to the second. Univariate and bivariate penalized spline smoothing is applied to the raw hazards to arrive at forecasts of mortality rates up to 2050.

As detailed in Section 3.5 geoadditive models for survival data are developed by Adebayo & Fahrmeir (2005), Hennerfeind, Brezger & Fahrmeir (2006), Kneib (2006), Kneib & Fahrmeir (2007) and Ganguli & Wand (2006).

3.20 Temporal data

The use of smoothing techniques in the analysis of temporal (time series) data has flourished in the past two decades – see, for example, Fan & Yao (2003). However, most of this work has involved local polynomial kernel smoothing. The permeation of these ideas to spline-based semiparametric regression is still quite mild.

Houseman, Coull & Shine (2006) develop negative binomial time series models for modeling enterococcus counts in Boston Harbor, utilizing penalized splines and mixed model representations. Jank & Shmueli (2007) use the same general approach to model concurrency of events in on-line auctions.

General correlation structures for mixed model-based smoothing are considered by Durbán & Currie (2003) and Krivobokova & Kauermann (2007). The latter reference contains asymptotic theory for the smoothing parameter chosen via AIC and REML, and application to finance time series data.

As discussed in Section 3.5 Fahrmeir, Kneib & Lang (2004) and Kneib & Fahrmeir (2006) use geoadditive models to handle temporal and spatial effects.

As mentioned in Section 3.3, Dominici *et al.* (2004) use and modify generalized additive model technology for air pollution time series data. Gryparis *et al.* (2007), discussed in Sections 3.5 and 3.7, has a temporal data aspect.

3.21 Miscellanea

A few 2003-2007 papers involving semiparametric regression do not fall into any of the categories corresponding to the previous subsections.

Yee & Hastie (2003) extends reduced-rank regression (e.g. Izenman, 1975) to the class of vector generalized linear models. While this work is mainly parametric, some non-linear modeling based on regression splines is used.

Yu & Ruppert (2004) build on their earlier work (Yu & Ruppert, 2002) on partially

linear single-index models using penalized splines. In particular, they remove the assumption of compactness and establish root- n consistency of the regression coefficients.

Wood (2004) is a rare instance of a semiparametric regression contribution that delves deeply into numerical issues. For example, pivoted QR decomposition is used to make GCV parameter choice in generalized additive (mixed) models more stable and efficient. Later releases of the author's R package, `mgcv` (Wood, 2008), make use of this methodology.

Banerjee, Maiti & Mukhopadhyay (2006) use penalized splines to build classification rules for the pathological state of prostate cancer patients. In Choudhary & Ng (2006), penalized spline estimates of both mean and variance functions are employed to assess agreement between two methods of measurement.

Piepho & Ogutu (2007) explains how simple state-space models can be expressed as linear mixed models. Estimation via REML as an alternative to the Kalman filter is investigated and some advantages are found. It is also explained how smoothing is achieved via integration of state-space components and how the class of covariance structures for modeling serial correlation is broadened via state-space representations.

Lee & Oh (2007) develop robust semiparametric regression procedures based on M-type penalized spline smoothers. Extension is made to additive mixed models, with a robust modification of REML for variance component estimation.

Eilers (2007) uses the discrete Whittaker smoother in meta-analysis. His approach includes nonparametric estimation of the latent distribution of event probabilities in control and treatment groups, and a smoothed EM algorithm with improved convergence to maximum likelihood estimates of the parameters in the latent distribution model.

3.22 Review articles

A few articles have reviewed aspects of semiparametric regression in the last few years. We briefly mention some of them here.

Tutz (2004) reviews semiparametric mixed models in the case of generalized responses. Generalized linear mixed models are shown to play a central role. Maximum likelihood is the main fitting tool. Techniques for dealing with the intractable integrals, such as Gauss-Hermite quadrature and the EM algorithm are described. Similar structures, although within the Bayesian framework and MCMC are treated by Zhao *et al.* (2006).

Brezger & Lang (2006) reviews Bayesian penalized spline approaches to generalized additive models. Pointers to implementation in the authors' `BayesX` package is included.

In Section 3.8 we mentioned the five-chapter component of Fitzmaurice *et al.* (2008). Together, these provide a detailed account of recent semiparametric regression research involving longitudinal data.

Finally, we mention two books from the last few years that have strong semiparametric regression themes. Wood (2006a) presents a thorough account of generalized additive models, with emphasis on implementation in R. Wu & Zhang (2006) focuses on semiparametric regression for longitudinal data, with emphasis on mixed model approaches.

4 Applications

Ruppert *et al.* (2003) emphasize the *modularity* of low-rank spline smoothers; a spline can be embedded as a nonparametric module into a larger model with parametric components. This type of use of such splines in applications has become increasingly sophisticated, as the following selection of applied papers show.

4.1 Blood lead exposure on intellectual impairment

Canfield *et al.* (2003) present an interesting application of semiparametric modeling to an important health problem. The authors study the intellectual impairment in children due to blood lead concentrations below 10 μg per deciliter, the “level of concern” as defined by the Centers for Disease Control and the World Health Organization. They measured blood lead concentration in 172 children at 6, 12, 18, 24, 36, 48, and 60 months of age and modeled longitudinal effects with a mixed model. Maternal intelligence quotient (IQ), quality of the home environment, and other potential confounders are adjusted linearly. Preliminary data analysis suggest that the dose-response curve for IQ might be steeper, that is, IQ decreases more rapidly, in the 0–10 μg per deciliter range compared to blood lead concentrations above 10 μg per deciliter. To model the nonlinear dose-response, the authors used a penalized spline. This semiparametric analysis corroborates the preliminary finding that IQ declines more rapidly with blood lead concentration at low doses compared to dose above 10 μg per deciliter. This result is in disagreement with the previous belief that 10 μg per deciliter is the “level of concern,” and the authors suggest that considerably more children are adversely affected by lead exposure than previously believed.

4.2 Spatial and temporal distribution of particulate air pollution

Gryparis, Coull, Schwartz & Suh (2007) model the spatial and temporal distribution of particulate air pollution in the greater Boston area. Data are available mostly from three Boston area monitoring studies, and there are two surrogates of mobile source pollution, black carbon (BC) and elemental carbon. The authors use a semiparametric latent variable model for combining these multiple surrogates for a common mobile source of pollution. The measurement error model is

$$\mathbf{y}_{ij} = \mathbf{g}(\mathbf{\Lambda}_i, \eta_{ij}) + \boldsymbol{\varepsilon}_{ij},$$

where \mathbf{y}_{ij} is a vector of measurements at location i and day j , \mathbf{g} is a known function, $\mathbf{\Lambda}_i$ is an unknown matrix of factor loadings, η_{ij} is a latent variable and $\boldsymbol{\varepsilon}_{ij}^y$ is an error vector. The loadings matrix $\mathbf{\Lambda}_i$ is modeled as having a linear regression on known covariates. Interest centers on the latent variable η_{ij} , and a geoadditive model is used to express η_{ij} as the sum of a linear function of certain covariates, univariate functions of other covariates, a bivariate function of longitude and latitude, and error. As is typical of factor models, constraints are needed to achieve identifiability. The model is fit separately to summer and winter data. The authors performed a Bayesian analysis and use a Gibbs sampler with Metropolis-Hastings steps. The geoadditive model facilitates visual presentation of the results. There is an obvious non-linear day of the year effect on particulate pollution. Maps of median predicted log-BC show the distribution of mobile source pollution in the greater Boston area during the summer and winter seasons.

4.3 Time series of enterococcus counts in a harbor

Houseman, Coull and Shine (2006) use semiparametric regression methods to develop a new type of model that was particularly suited for their application: enterococcus counts in Boston Harbor. A noteworthy aspect of their model is that it handles a non-stationary time series of counts. The aim of their research was to understand the effects of changes in sewage treatment that were initiated to improve water quality. The authors assume that counts, y_t , are observed on a finite set of time points in an interval \mathcal{T} , and they depend on random effects Q_t and fixed covariate effects μ_t so that $y_t|Q_t$ is $\text{Poisson}(Q_t\mu_t)$. Here the Q_t are independent Gamma with shape and rate parameters both σ^{-1} and induce overdispersion, while μ_t models covariate effects and time effects. Specifically, μ_t

is equal to $\exp\{\mathbf{x}_t^T \boldsymbol{\beta} + f(t)\}$ where $\mathbf{x}_t^T \boldsymbol{\beta}$ is a parametric model for covariate effects, and $f(t)$ is a nonparametric penalized spline model for possibly non-stationary time trends. Because the covariates are time dependent, this semiparametric model is non-identifiable without constraints that are carefully explained by the authors. The time-varying covariates includes four variables that characterized sewage flows, temperature, tide height, salinity, and a sinusoidal seasonal effect. This model is fit separately at each spatial location. Then the authors use a geoaddivitive fit (Kammann & Wand, 2003) to create a spatial summary of the effects of major interest, those of flows from the Nut Island and Deer Island Treatment Plants.

4.4 Concurrency of events in on-line auctions

On-line auctioning is a relatively new and rich source of challenging statistical problems. Jank & Shmueli (2007) investigate concurrency in on-line auctions. They define concurrency as the effect upon an event of the same or similar events at or near the same time. The on-line auction web-site eBay conducts hundreds of simultaneous auctions for similar items. In addition, eBay makes available information on auctions that have closed in the last 15 days. It is expected that both types of information will affect the final price of an item being auctioned. To study these effects, Jank & Shmueli (2007) use the model

$$y_t = g_{AC}(\mathbf{x}_t) + g_{SC}(\mathbf{x}_t) + g_{TC}(\mathbf{x}_{t:(t-1)}) + \varepsilon_t,$$

where y_t is the log-price of an item sold at time t , \mathbf{x}_t is covariate information available at time t , and $\mathbf{x}_{t:(t-1)}$ is covariate information over the time period from $t - 1$ to t . Time is modeled continuously since a auction can close at any time. The three components of the model are: g_{AC} , the ‘‘auction component’’; g_{SC} , the ‘‘spatial component’’; and g_{TC} , the ‘‘temporal component’’. In their example, they use the prices of laptop computers. The auction component is modeled linearly, but the other components are modeling non-parametrically. ‘‘Spatial’’ refers to a feature space where distance measures similarly between laptops in terms of their features, e.g., screen size, memory size, and presence of a DvD drive. Therefore, the authors estimate g_{SC} using a radial penalized spline in 7-dimensional space. The temporal component requires a more complex model than the other two components. The covariates are the prices from time $t - 1$ to t and the features of those laptops. Various functions of the prices, e.g., mean, median, minimum, maximum, and slope of the time trend, are computed for laptops most similar, least similar, and of average similarity to the laptop sold at time t . The result is 18 time-lag variables which are reduced to three principal components. The temporal component is an additive spline model in these principal components. Jank & Shmueli (2007) test their model on a hold-out sample of 30% of the laptops sold, with the remaining 70% used to train the model. They compared the performance of their model with that of linear parametric models and totally nonparametric models that use regression trees. They find that the nonparametric models outperform the linear models, but that their semiparametric model outperforms the nonparametric models.

4.5 Genomic-assisted prediction of genetic value

Gianola, Fernando & Stella (2006) use a semiparametric model for the genetic value of single nucleotide polymorphisms (SNP) and other genetic markers. Let y_i be a measurement, such as height of a plant or milk production of a cow, and let \mathbf{x}_i be a vector of dummy genetic marker variables, e.g., the indicators of the presence of SNP or microsatellite covariates. Gianola *et al.* (2006) use the model

$$y_i = \mathbf{w}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \mathbf{u} + g(\mathbf{x}_i) + \varepsilon_i, \quad (12)$$

where w_i and z_i are known incidence vectors, β is a vector of nuisance location effects, u is a $q \times 1$ vector of additive genetic effects of q individuals, which are modeled as random effects, and g is an unknown function. They present several estimation methods for this model. One of these methods converts (12) into a mixed model by using a type of radial basis function model for $g(x)$. Specifically, they follow Mallick, Ghosh & Ghosh (2005) and use a reproducing kernel mixed model that assumes that

$$g(\mathbf{x}) = \sum_{j=1}^n \alpha_j \exp\{-\|\mathbf{x} - \mathbf{x}_j\|^2/h\}$$

where the α_j are iid $N(0, \sigma_\alpha^2)$ and h is a non-negative smoothing parameter. Their model has one knot at each \mathbf{x}_j , but it should be possible to use only a subset of these knots, say, chosen by a space-filling design. The authors use a simulation experiment to compare the reproducing kernel mixed model method with a parametric mixed model approach. They find that the two estimators have nearly equal performance when the parametric model holds and that the semiparametric method outperforms the parametric method when the linear model does not hold.

4.6 Carbon sequestration in agricultural soils

Sequestration of carbon in soils has the potential to reduce greenhouse gases. This was the motivation for a study by Breidt, Hsu, and Ogle (2007) who use a semiparametric mixed model to compare carbon sequestering under no-till and traditional tillage. Their main conclusion is that more carbon is sequestered under no-tillage than under traditional tillage, especially in wet climates but also in dry climates. Their data come from soil cores. A core is divided into increments, e.g., from 0 to 15cm depth, and total carbon is measured in each increment. They use 63 paired (no-till versus traditional tillage) studies, with 211 increments in total. The boundaries of the increments varied from study to study, making increment-wise comparisons impossible. Therefore, the authors used a varying coefficient penalized spline model for the concentration of carbon as a function of depth, so that the total carbon in any increment is the integral of this function over the increment. These “instantaneous” carbon sequestration functions can be estimated from the increment data and then compared between the two types of tillage. A varying coefficient model of the instantaneous function is needed to accommodate the effects of covariates such as soil type, climate factors, and the number of years since change to no-till. More specially, the model for difference between no-till and traditional tillage in carbon concentration at soil depth t in the i th study is

$$\sum_{\ell=1}^q \alpha_\ell(t) w_{i\ell} \tag{13}$$

where $\alpha_\ell(t)$ is a spline, $w_{\ell,i}$ is the value of the ℓ th covariate in the i th study, and q is the number of covariates. Several choices of covariates are considered and the final choice was to use the indicator of wet climate and the number of years since the change in soil management. Two models are considered for covariance function within a core. The first has i.i.d. random intercepts, which implies a correlation matrix with compound symmetry. The second, which uses a non-homogeneous Ornstein-Uhlenbeck model, allowed heteroscedasticity and a more general type of correlation. The second model fit significantly better than the first and is used by the authors. By plotting the fitted models given by (13) for different values of the covariates, the authors show the effects of no-tillage. Under no-tillage, there is more sequestered carbon in the upper soil layers and less in the lower layers compared to traditional tillage, but the former effect is dominant so that

overall more carbon is sequestered under no-tillage. This suggests that a change to no-till, which has a number of other advantages, also has the beneficial effect of reducing the amount of CO₂ in the atmosphere.

4.7 Time series of air pollution and mortality

In studies of the effects of air pollution on mortality, confounders that are unmeasured, and perhaps even unknown, can bias the estimates. To circumvent this problem, analysts often include in the model a smooth function $f(t)$ of time (t) to capture the effects of confounders that vary smoothly in time. An example, the Milan study of air pollution and mortality, can be found in Ruppert *et al.* (2003). The technique of including $f(t)$ in the model is given careful study by Peng, Dominici & Louis (2006). An issue of primary concern is selecting the degrees of freedom for estimation of $f(t)$. Peng *et al.* (2006) find that the estimator of $f(t)$ should be undersmoothed to reduce the bias in the estimate of the effect of pollution, which is modeled linearly with coefficient β . This finding agrees with asymptotic theory for partially linear models (Rice, 1986; Speckman, 1988). The authors find that the method for selecting the degrees of freedom for $\hat{f}(t)$ that is most accurate for estimating β is to use GCV to find the degrees of freedom that best predicts the pollution series. Then one estimates $f(t)$ with the same degrees of freedom. The function $f(t)$ can be estimated by either an ordinary least squares fit with a natural cubic spline basis or by a penalized spline. The latter requires more degrees of freedom for $\hat{f}(t)$ to achieve approximate unbiasedness of $\hat{\beta}$. Peng *et al.* (2006) include an interesting example involving a 100-city study of the effect of suspended particulate matter on mortality. Data are available from 1987 to 2000. They use an over-dispersed Poisson model with a log link for the daily number of deaths. Known confounders are accounted for explicitly: there are age-specific intercepts, a day of week effect, and smooth functions of temperature and dewpoint. Particulate matter enters the model linearly and the estimate of its coefficient β is studied as the degrees of freedom for $f(t)$ varies. The ordinary least squares natural cubic spline needs about 9 degrees of freedom per year before $\hat{\beta}$ stabilizes. For the penalized spline about 15 degrees of freedom are needed.

4.8 The cosmic microwave background

Genovese, Miller, Nichol, Arjunwadkar & Wasserman (2004) address an important problem in cosmology. They study the peaks in the temperature power spectrum of the cosmic microwave background radiation. Let y_ℓ be the estimated spectrum at multipole index ℓ . The model is $y_\ell = f(x_\ell) + \epsilon_\ell$ where $x_\ell = \ell / \max(\ell)$. A parametric model for f has three peaks and the existence of the third (rightmost) peak would provide the clearest support for the existence of dark matter. The response y_ℓ is highly heteroscedastic with its variance increasing rapidly in ℓ . This complicates inference, especially for higher values of x_ℓ which is precisely where the third peak should be located. Genovese *et al.* (2004) estimate f by a truncated cosine expansion. To construct a uniform confidence set, they extend the methodology of Beran and Dümbgen (1998) to accommodate the heteroscedasticity. The result is a 900-dimensional confidence ball which is, of course, difficult to visualize. To explore the ball, they create targeted “probes” which are functionals of interest. Using the probes they can, for example, find 95% confidence intervals for the heights and widths of the first two peaks. The nonparametric fit is compared with the so-called concordance model, which maximizes the joint likelihood under the parametric model of five independent data sets. The nonparametric fit does not have the third peak although the concordance model does, since the third peak is an intrinsic part of the parametric model. The lack of the third peak in the nonparametric fit does not mean that the third peak does not exist. Rather, more precise data would be needed in order to establish its existence.

This paper is noteworthy both for addressing a very interesting scientific question and for its novel use of simultaneous inference.

4.9 Needle losses of Norway spruces

Augustin, Lang, Musio & von Wilpert (2007) study needle loss which is an indicator of tree vitality. They work with survey data on Norway spruces (*Picea abies*) in the southwestern region of Germany. One novel aspect of the paper is that the response is ordered categorical. The categories are healthy, intermediate, or damaged, defined, respectively, as less than 10%, 10–25%, or more than 25% needle loss. Augustin *et al.* use a geoadditive model for a latent continuous variable U such that

$$U = f_1(x_1) + \cdots + f_P(x_P) + f_{\text{spat}}(c_1, c_2) + \mathbf{w}^T \boldsymbol{\gamma} + \varepsilon$$

where x_1, \dots, x_P are continuous covariates, (c_1, c_2) is spatial location, \mathbf{w} is a vector of covariates that enter linearly, and ε is $N(0, 1)$. The categorical response is a discretized version of U with cutoffs $\theta_1 < \theta_2$. Univariate P-splines are used to model f_1, \dots, f_P and a tensor product of B-splines to model f_{spat} . Including f_{spat} in the model accommodates unknown covariates, but also acts as a partial surrogate for known covariates and reduces the size of their effects. The analysis is Bayesian using MCMC. One important problem is prediction of needle loss at locations not covered by the surveys. The model can be used for prediction, but a complication is that some covariates are also unknown at these locations. To circumvent this problem, the authors use a spatial model for these covariates and draw multiple imputations from their posterior distributions.

4.10 Capture-recapture studies

Mark-recapture studies are a common means of assessing animal abundances and survival probabilities. Frequently, survival probabilities depend on covariates. For example, Gimenez, Crainiceanu, Barbraud, Jenouvrier & Morgan (2006) describe a case study where the survival probabilities of snow petrels nesting at Petrels Island, Terre Adélie, Antarctica, depend upon the Southern Oscillation Index (SOI). SOI is negatively related to temperature and can be used as an index of overall climate condition. Gimenez *et al.* (2006) assume that there are $I + 1$ sampling occasions at times t_1, \dots, t_{I+1} . They define ϕ_i to be the probability that an animal survives to time t_{i+1} given that it is alive at time t_i . The data consist of the number of animals captured, marked, and released at each sampling occasion and the number marked at time t_i and recaptured for the first time at t_j . The authors begin with the Cormack-Jolly-Seber model, which has among its parameters ϕ_1, \dots, ϕ_I . Then they use a semiparametric model with a logit link function for the dependence of ϕ_i upon covariates. The nonparametric dependencies are modeled by splines. They propose a Bayesian analysis with computations by MCMC. In the snow petrel case study, they use WinBUGS. They find that survival probabilities of snow petrels decrease, possibly in a nonlinear way, with increasing values of the SOI. The estimated rate of decrease is high at low values of the SOI but diminishes at higher values of the SOI. Because the data are sparse, there is too much uncertainty to conclude that the effect of SOI is nonlinear. However, Gimenez *et al.* (2006) note that a nonlinear effect of SOI is biologically plausible; lower values of SOI might increase access to prey but prey abundance may increase with higher values of SOI (Loeb *et al.*, 1997).

References

Adebayo, S.B. & Fahrmeir, L. (2005). Analysing child mortality in Nigeria with geoadditive discrete-time survival models. *Statistics in Medicine*, **24**, 709–728.

- Ansley, C.F., Kohn, R. & Wong, C.-M. (1993). Nonparametric spline regression with prior information. *Biometrika*, **80**, 75–88.
- Antoniadis, A. & Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Computational Statistics and Data Analysis*, **51**, 4793–4813.
- Apanasovich, T.V., Ruppert, D., Lupton, J.R., Popovic, N., Turner, N.D., Chapkin, R.S. & Carroll, R. J. (2008). Semiparametric longitudinal-spatial binary regression, with application to colon carcinogenesis. *Biometrics*, **64**, 490–500.
- Augustin, N.H., Lang, S., Musio, M. & von Wilpert, K. (2007). A spatial model for the needle losses of pine-trees in the forests of Baden-Württemberg: an application of Bayesian structured additive regression. *Applied Statistics*, **56**, 29–50.
- Avalos, M. & Grandvalet, Y. & Ambroise, C. (2007). Parsimonious additive models. *Computational Statistics and Data Analysis*, **51**, 2851–2870.
- Bachrach, L.K., Hastie, T., Wang, M.-C., Narasimhan, B. & Marcus, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth. A longitudinal study. *J. Clin. Endocrin. Metab.* **84**, 4702–12.
- Baladandayuthapani, V., Mallick, B.K. & Carroll, R.J. (2005). Spatially adaptive Bayesian regression splines. *Journal of Computational and Graphical Statistics*, **14**, 378–394.
- Baladandayuthapani, V., Mallick, B.K., Hong, M.Y., Lupton, J.R., Turner, N.D. & Carroll, R.J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, **64**, 64–73.
- Banerjee, T., Maiti, T. & Mukhopadhyay, P. (2006). Classification of pathological stage of prostate cancer patients using penalized splines. *Computational Statistics and Data Analysis*, **51**, 1147–1155.
- Beaumont, M.A., Zhang, W. & Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.
- Beran, R., and Dümbgen, L. (1998). Modulation of estimators and confidence sets. *The Annals of Statistics*, **26**, 1826–1856.
- Berry, S.A., Carroll, R.J. & Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, **97**, 160–169.
- Binder, H. & Tutz, G. (2008). A comparison of methods for the fitting of generalized additive models. *Statistics and Computing*, **18**, 87–99.
- Bollaerts, K., Eilers, P.H.C. & Aerts, M. (2006). Quantile regression with monotonicity restrictions using P-splines and the L_1 norm. *Statistical Modelling*, **6**, 189–207.
- Bollaerts, K., Eilers, P.H.C. & van Mechelen, I. (2006). Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, **59**, 451–469.
- Branscum, A. J., Johnson, W. O. & Thurmond, M.C. (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian and New Zealand Journal of Statistics*, **49**, 287–301.
- Breidt, F.J., Claeskens, G. & Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, **92**, 831–846.

- Breidt, F.J., Hsu, N.J. & Ogle, S. (2007). Semiparametric mixed models for increment-averaged data with application to carbon sequestration in agricultural soils. *Journal of the American Statistical Association*, **102**, 803–812.
- Breidt, F.J. & Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, **28**, 1026–1053.
- Breidt, F.J. & Opsomer, J.D. (2009). Nonparametric and semiparametric estimation in complex surveys. In: *Sample Surveys: Theory, Methods and Inference*, Handbook of Statistics, Vol. 29, C.R. Rao and D. Pfeiffermann (Editors), North Holland.
- Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, **26**, 801–824.
- Brezger, A., Fahrmeir, L. & Hennerfeind, A. (2007). Adaptive Gaussian Markov random fields with applications in human brain mapping. *Applied Statistics*, **56**, 327–345.
- Brezger, A., Kneib, T. & Lang, S. (2005). BayesX: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software*, Volume 14, Issue 11.
- Brezger, A. & Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, **50**, 967–991.
- Brown, D., Kauermann, G. & Ford, I. (2007). A partial likelihood approach to smooth estimation of dynamic covariate effects using penalised splines. *Biometrical Journal*, **49**, 441–452.
- Brown, L.D., Cai, T.T., Low, M.G. & Zhang, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *The Annals of Statistics*, **30**, 688–707.
- Brown, L.D. & Low, M.G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, **24**, 2384–2398.
- Bühlmann, P. & Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science*, **22**, 477–522.
- Bühlmann, P. & Yu, B. (2003). Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, **98**, 324–339.
- Cadarso-Suarez, C., Roca-Pardinas, J. & Figueiras, A. (2006). Effect measures in non-parametric regression with interactions between continuous exposures. *Statistics in Medicine*, **25**, 603–621.
- Cai, T. & Betensky, R.A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics*, **59**, 570–579.
- Cai, T., Hyndman, R.J. & Wand, M.P. (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics*, **11**, 784–798.
- Canfield, R.L., Henderson, C.R., Cory-Slechta, D.A., Cox, C., Jusko, T.A. & Lanphear, B.P. (2003). Intellectual impairment in children with blood lead concentrations below 10 μg per deciliter. *The New England Journal of Medicine*, **348**, 1517–1526.
- Cantet, R.J.C., Birchmeier, A.N., Cayo, A.W.C. & Fioretti, C. (2005). Semiparametric animal models via penalized splines as alternatives to models with contemporary groups. *Journal of Animal Science*, **83**, 2482–2494.
- Cardot, H., Ferraty, F. & Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, **13**, 571–591.
- Carroll, R.J., Delaigle, A. & Hall, P. (2008). Nonparametric regression estimation from data contaminated by a mixture of Berkson and classical errors. *Journal of the Royal Statistical Society*,

Series B, **69**, 859–878.

- Carroll, R.J., Hall, P., Apanasovich, T.V. & Lin, X. (2004). Histospline method in nonparametric regression models with application to clustered/longitudinal data. *Statistica Sinica*, **14**, 649–674.
- Carroll, R.J., Ruppert, D., Stefanski, L.A. & Crainiceanu, C.M. (2006). *Measurement Error in Non-linear Models (Second Edition)*. Boca Raton, Florida: Chapman & Hall/CRC.
- Carroll, R.J., Ruppert, D., Tosteson, T.D., Crainiceanu, C. & Karagas, M.R. (2004). Nonparametric regression and instrumental variables. *Journal of the American Statistical Association*, **99**, 736–750.
- Casella, G. & Robert, C. (2004). Introduction to the special issue: Bayes then and now. *Statistical Science*, **19**, 1–2.
- Chaudhuri, P. & Marron, J.S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807–823.
- Chavez-Demoulin, V. & Davison, A.C. (2005). Generalized additive modelling of sample extremes. *Applied Statistics*, **54**, 207–222.
- Choudhary, P.K. (2007). Semiparametric regression for assessing agreement using tolerance bands. *Computational Statistics and Data Analysis*, **51**, 6229–6241.
- Choudhary, P.K. & Ng, H.K.T. (2006). Assessment of agreement under nonstandard conditions using regression models for mean and variance. *Biometrics*, **62**, 288–296.
- Chen, Q.X. & Ibrahim, J.G. (2006). Semiparametric models for missing covariate and response data in regression models. *Biometrics*, **62**, 177–184.
- Chen, K. & Jin, Z. (2005). Local polynomial regression analysis of clustered data. *Biometrika*, **92**, 59–74.
- Christensen, O.F. & Ribeiro, P.J. (2008) `geoRglm 0.8`. R package. <http://cran.r-project.org>.
- Claeskens, G. (2004). Restricted likelihood ratio lack-of-fit tests using mixed spline models. *Journal of the Royal Statistical Society, Series B*, **66**, 909–926.
- Cole, T.J. & Green, P.J. (1992). Smooth reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305–1319.
- Congdon, P. (2006). A model for non-parametric spatially varying regression effects. *Computational Statistics and Data Analysis*, **50**, 422–445.
- Cook, R.D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman & Hall.
- Coull, B.A. & Staudenmayer, J. (2004). Self-modeling regression for multivariate curve data. *Statistica Sinica*, **14**, 695–711.
- Cowell, R.G., Dawid, A.P., Lauritzen, S.L. and Spiegelhalter, D.J. (1999). *Probabilistic Networks and Expert Systems*. New York: Springer.
- Crainiceanu, C.M., Diggle, P.J. & Rowlingson, B. (2008). Bivariate binomial spatial modelling Loa loa prevalence in tropical Africa (with discussion). *Journal of the American Statistical Association*, **103**, 21–43.
- Crainiceanu, C. & Ruppert, D. (2004a). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B*, **66**, 165–185.

- Crainiceanu, C. & Ruppert, D. (2004b). Restricted likelihood ratio tests for longitudinal models. *Statistica Sinica*, **14**, 713–729.
- Crainiceanu, C. & Ruppert, D. (2004c). Likelihood ratio tests for goodness-of-fit of a nonlinear regression model. *Journal of Multivariate Analysis*, **91**, 35–42.
- Crainiceanu, C.M., Ruppert, D., Carroll, R.J., Joshi, A. & Goodner, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computational Graphical Statistics*, **16**, 265–288.
- Crainiceanu, C., Ruppert, D., Claeskens, G. & Wand, M.P. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika*, **92**, 91–103.
- Crainiceanu, C., Ruppert, D. & Wand, M.P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, Volume 14, Article 14.
- Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.
- Currie, I.D. & Durbán, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, **4**, 333–349.
- Currie, I.D., Durbán, M. & Eilers, P.H.C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**, 279–298.
- Currie, I.D., Durbán, M. & Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, **68**, 259–280.
- Dean, C.B., Nathoo, F. & Nielsen, J.D. (2007). Spatial and mixture models for recurrent event processes. *Environmetrics*, **18**, 713–725.
- Del Moral, P., Doucet, A. & Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B*, **68**, 411–436.
- Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society, Series B*, **60**, 333–350.
- Diggle, P.J., Heagerty, P., Liang, K.-L. & Zeger, S. (2002). *Analysis of Longitudinal Data (Second Edition)*. Oxford: Oxford University Press.
- Dimatteo, I., Genovese, C.R. & Kass, R.E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika*, **88**, 1055–1071.
- Dimitriadou, E., Hornik, K., Leisch, F., Mayers, D. & Weingessel, A. (2008). e1071 1.5 R package. <http://cran.r-project.org>
- Dominici, F., McDermott, A. & Hastie, T. (2004). Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association*, **99**, 938–948.
- Donnelly, C.A., Laird, N.M. & Ware, J.H. (1995). Prediction and creation of smooth curves for temporally correlated longitudinal data. *Journal of the American Statistical Association*, **90**, 984–989.
- Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis (Third Edition)*. New York: John Wiley & Sons.
- Durbán, M. & Currie, I. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics*, **18**, 251–262.

- Durbán, M., Harezlak, J., Wand, M.P. & Carroll, R.J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**, 1153–1167.
- Efron, B. (1986). Double exponential families and their use in generalized linear regression. **81**, 709–721.
- Eilers, P.H.C. (2005). Unimodal smoothing. *Journal of Chemometrics*, **19**, 317–328.
- Eilers, P.H.C. (2007). Data exploration in meta-analysis with smooth latent distributions. *Statistics in Medicine*, **26**, 3358–3368.
- Eilers, P.H.C., Currie, I.D. & Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, **50**, 61–76.
- Eilers, P.H.C., Gampe, J., Marx, B.D. & Rau, R. (2008). Modulation models for seasonal incidence tables. *Statistics in Medicine*, **NA**, NA–NA.
- Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.
- Eilers, P.H.C. & Marx, B.D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, **66**, 159–174.
- Elliott, M.R. (2007). Identifying latent clusters of variability in longitudinal data. *Biostatistics*, **8**, 756–771.
- Fahrmeir L. & Echavarría, L.O. (2006). Structured additive regression for overdispersed and zero-inflated count data. *Applied Stochastic Models in Business and Industry*, **22**, 351–369.
- Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 715–745.
- Fahrmeir, L. & Raach, A. (2007). A Bayesian semiparametric latent variable model for mixed responses. *Psychometrika*, **72**, 327–346.
- Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.
- Fan, J. & Gijbels, I. (1995). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.
- Fan, J. & Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.
- Fan, Y., Leslie, D.S. & Wand, M.P. (2006). Comment on paper by Del Moral, Doucet & Jasra. *Bayesian Statistics 8*, Oxford University Press.
- Figueiras, A., Roca-Pardinas, J. & Cadarso-Suarez, C. (2005). A bootstrap method to avoid the effect of concavity in generalised additive models in time series studies of air pollution. *Journal of Epidemiology and Community Health*. **59**, 881–884.
- Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (Eds.) (2008). *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. Boca Raton, Florida: Chapman & Hall/CRC.
- Floyd, R.W. (1962). Algorithm 97: shortest path. *Communications of the Association for Computing Machinery*, **5**, 345.
- French, J.L. & Wand, M.P. (2004). Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics*, **5**, 177–191.

- Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, **121**, 256–285.
- Freund, Y. & Schapire, R.E. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, San Francisco: Morgan Kaufman, pp. 148–156.
- Friedman, J., Hastie, T. & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *The Annals of Statistics*, **28**, 337–407.
- Ganguli, B., Staudenmayer, J. & Wand, M.P. (2005). Additive models with predictors subject to measurement error. *Australia and New Zealand Journal of Statistics*, **47**, 193–202.
- Ganguli, B. & Wand, M.P. (2004). Feature significance in geostatistics. *Journal of Computational and Graphical Statistics*, **13**, 954–973.
- Ganguli, B. & Wand, M.P. (2006). Additive models for geo-referenced failure time data. *Statistics in Medicine*, **25**, 2469–2482.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.
- Genovese, C.R., Miller, C.J., Nichol, R.C., Arjunwadkar, M. & Wasserman, L. (2004). Nonparametric inference for the cosmic microwave background. *Statistical Science*, **19**, 308–321.
- Geraci, M. & Bottai, M. (2006). Use of auxiliary data in semi-parametric spatial regression with nonignorable missing responses. *Statistical Modelling*, **6**, 321–336.
- Ghidey, W., Lesaffre, E. & Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, **60**, 945–953.
- Ghosh, D. (2007). Incorporating monotonicity into the evaluation of a biomarker. *Biostatistics*, **8**, 402–413.
- Gianola, D., Fernando, R.L. & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, **173**, 1761–1776.
- Gimenez, O., Crainiceanu, C., Barbraud, C., Jenouvrier, S. & Morgan, B.J.T. (2006). Semiparametric regression in capture-recapture modeling. *Biometrics*, **62**, 691–698.
- Gluhovsky, I. & Vengerov, D. (2007). Constrained multivariate extrapolation models with application to computer cache rates. *Technometrics*, **49**, 129–137.
- Green, P.J. (1985). Linear models for field trials, smoothing and cross-validation. *Biometrika*, **72**, 523–537.
- Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.
- Greven, S., Crainiceanu, C.M., Kuechenhoff, H. & Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, **17**, 8700–891.
- Gryparis, A., Coull, B.A., Schwartz, J. & Suh, H.H. (2007). Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area. *Applied Statistics*, **56**, 183–209.
- Gurrin, L.C., Scurrah, K.J. & Hazelton, M.L. (2005). Tutorial in biostatistics: spline smoothing with linear mixed models. *Statistics in Medicine*, **24**, 3361–3381.

- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383–393.
- Haario, H., Saksman, E., Tamminen, J. (2005). Componentwise adaptation for high dimensional MCMC. *Computational Statistics*, **20**, 265–273.
- Hall, P. & Opsomer, J.D. (2005). Theory for penalised spline regression, *Biometrika*, **92**, 105–118.
- Harezlak, J., Coull, B.A., Laird, N.M., Magari, S.R. & Christiani, D.C. (2006). Penalized solutions to functional regression problems. *Computational Statistics and Data Analysis*, **51**, 4911–4925.
- Harezlak, J., Naumova, E. & Laird, N.M. (2007). LongCrisp: A test for bump hunting in longitudinal data. *Statistics in Medicine*, **26**, 1383–1397.
- Harezlak, J., Ryan, L.M., Giedd, J.N. & Lange, N. (2005). Individual and population penalized regression splines for accelerated longitudinal designs. *Biometrics*, **61**, 1037–1048.
- Hastie, T. (2006). `gam 0.98`. R package. <http://cran.r-project.org>.
- Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman and Hall.
- Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.
- Hastie, T. & Zhu, J. (2006). Comment on paper by Moguerza and Muñoz. *Statistical Science*, **21**, 352–357.
- Heim, S., Fahrmeir, L., Eilers, P.H.C. & Marx, B.D. (2007). 3D space-varying coefficient models with application to diffusion tensor imaging. *Computational Statistics and Data Analysis*, **51**, 6212–6228.
- Hennerfeind, A., Brezger, A. & Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, **101**, 1065–1075.
- Hens, N., Aerts, M. & Molenberghs, G. (2006). Model selection for incomplete and design-based samples. *Statistics in Medicine*, **25**, 2502–2520.
- Hickernell, F.J., Lemieux, C. & Owen, A.B. (2005). Control variates for quasi-Monte Carlo. *Statistical Science*, **20**, 1–31.
- Houseman, E.A., Coull, B.A. & Shine, J.P. (2006). A nonstationary negative binomial time series with time-dependent covariates: enterococcus counts in Boston Harbor. *Journal of the American Statistical Association*, **101**, 1365–1376.
- Hu, Z.H., Wang, N. & Carroll, R.J. (2004). Profile-kernel versus backfitting in the partially linear models for longitudinal/clustered data. *Biometrika*, **91**, 251–262.
- Izenman, A.J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, **5**, 248–264.
- Jank, W. & Shmueli, G. (2007). Modelling concurrency of events in on-line auctions via spatiotemporal semiparametric models. *Applied Statistics*, **56**, 1–27.
- Jordan, M.I. (2004). Graphical models. *Statistical Science*, **19**, 140–155.
- Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. & Saul, L.K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, **37**, 183–233.
- Jullion, A. & Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics and Data Analysis*,

51, 2542–2558.

- Kammann, E.E. & Wand, M.P. (2003). Geoadditive models. *Applied Statistics*, **52**, 1–18.
- Karatzoglou, A., Smola, A. & Hornik, K. (2007). kernlab 0.9. R package. <http://cran.r-project.org>
- Kauermann, G. (2005). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning Inference*, **127**, 53–69.
- Kauermann, G. & Khomski, P. (2006). Additive two-way hazards model with varying coefficients. *Computational Statistics and Data Analysis*, **51**, 1944–1956.
- Kauermann, G., Krivobokova, T. & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B*, **71**, 487–503.
- Kim, I., Carroll, R.J. & Cohen, N.D. (2003). Semiparametric regression splines in matched case-control studies. *Biometrics*, **59**, 1158–1169.
- Kimeldorf, G.S. & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Its Application*, **33**, 82–95.
- Kneib, T. (2006). Mixed model-based inference in geoadditive hazard regression for interval-censored survival times. *Computational Statistics and Data Analysis*, **51**, 777–792.
- Kneib, T. & Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: a mixed model approach. *Biometrics*, **62**, 109–118.
- Kneib, T. & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, **34**, 207–228.
- Koenker, R. (2008). Quantile regression in R: a vignette. <http://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf>
- Kooperberg, C. (2007). polyspline 1.0 R package. <http://cran.r-project.org>
- Krivobokova, T. (2007). AdaptFit 0.2 R package. <http://cran.r-project.org>
- Krivobokova, T., Crainiceanu, C.M. & Kauermann, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, **17**, 1–20.
- Krivobokova, T. & Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, **102**, 1328–1337.
- Kuo, F., Dunsmuir, W.T.M., Sloan, I.H., Wand, M.P. & Womersley, R.S. (2008). Quasi-Monte Carlo for highly structured generalised response models. *Methodology and Computing in Applied Probability*, **10**, 239–275.
- Lambert, P. & Eilers, P.H.C. (2005). Bayesian proportional hazards model with time-varying regression coefficients: a penalized Poisson regression approach. *Statistics in Medicine*, **24**, 3977–3989.
- Lang, S., Adebayo, S.B., Fahrmeir, L. & Steiner, W.J. (2003). Bayesian geoadditive seemingly unrelated regression. *Computational Statistics*, **18**, 163–192.
- Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.
- Lee, T.C.M. & Oh, H.S. (2007). Robust penalized regression spline fitting with application to additive mixed modeling. *Computational Statistics*, **22**, 159–171.

- Leitenstorfer, F. & Tutz, G. (2007a). Knot selection by boosting techniques. *Computational Statistics and Data Analysis*, **51**, 4605–4621.
- Leitenstorfer, F. & Tutz, G. (2007b). Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics*, **8**, 654–673.
- Li, Y. & Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, **95**, 415–436.
- Li, Y., Wang, N., Hong, M., Turner, N.D., Lupton, J.R. & Carroll, R.J. (2007). Nonparametric estimation of correlation functions in longitudinal and spatial data, with application to colon carcinogenesis experiments. *The Annals of Statistics*, **35**, 1608–1643.
- Liang, H., Wu, H. & Carroll, R.J. (2003). The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying-coefficient semiparametric models with measurement error. *Biostatistics*, **4**, 297–312.
- Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K. & Sturtz, S. (2007). BRugs 0.4. R package. <http://cran.r-project.org>.
- Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, **84**, 309–326.
- Lin, X. & Carroll, R.J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, **95**, 520–534.
- Lin, X. & Carroll, R.J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*, **96**, 1045–1056.
- Lin, X. & Carroll, R.J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B*, **68**, 68–88.
- Lin, X., Wang, N., Welsh, A.H. & Carroll, R.J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika*, **91**, 177–193.
- Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B*, **61**, 381–400.
- Lin, J., Zhang, D.W. & Davidian, M. (2006). Smoothing spline-based score tests for proportional hazards models. *Biometrics*, **62**, 803–812.
- Linton, O.B., Mammen, E., Lin, X. & Carroll, R.J. (2003). Correlation in marginal longitudinal nonparametric regression. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, Eds. D. Y. Lin and P. J. Heagerty, pp. 23–33, New York: Springer.
- Liu, D., Lin, X. & Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*, **63**, 1079–1088.
- Liu, A. & Wang, Y.D. (2004). Hypothesis testing in smoothing spline models. *Journal of Statistical Computing and Simulation*, **74**, 581–597.
- Loeb, V., Siegel, V., Holm-Hansen, O., Hewitt, R., Fraser, W., Trivelpiece, W., & Trivelpiece, S. (1997). 'Effects of sea-ice extent and krill or salp dominance on the Antarctic food web. *Nature*, **387**, 897–900.
- Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.
- Luo, Z. & Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association*,

- Ma, Y. & Carroll, R.J. (2006). Locally efficient estimators for semiparametric models with measurement error. *Journal of the American Statistical Association*, **101**, 1465–1474.
- MacNab, Y.C. (2007). Spline smoothing in Bayesian disease mapping. *Environmetrics*, **18**, 727–744.
- MacNab, Y.C. & Gustafson, P. (2007). Regression B-spline smoothing in Bayesian disease mapping: with an application to patient safety surveillance. *Statistics in Medicine*, **26**, 4455–4474.
- Malfait, N. & Ramsay, J.O. (2003). The historical functional linear model. *The Canadian Journal of Statistics*, **31**, 115–128.
- Mallick, B. K., Ghosh, D. & Ghosh, M. (2005). Bayesian classification of tumours by using gene expression data. *Journal of the Royal Statistics Society, Series B*, **67**, 219–234.
- Mallick, B., Hoffman, F.O. & Carroll, R.J. (2002). Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada Test Site. *Biometrics*, **58**, 13–20.
- Marjoram, P. Molitor, J., Plagnol, V. & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences in the United States of America*, **100**, 15324–15328.
- Marron, J.S. & Zhang, J.-T. (2005). SiZer for smoothing splines. *Computational Statistics*, **20**, 481–502.
- Marx, B.D. & Eilers, P.H.C. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics*, **41**, 1–13.
- Marx, B.D. & Eilers, P.H.C. (2002). Multivariate calibration stability: a comparison of methods. *Journal of Chemometrics*, **16**, 129–140.
- Marx, B.D. & Eilers, P.H.C. (2005). Multidimensional penalized signal regression. *Technometrics*, **47**, 13–22.
- Massy, W.F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, **60**, 234–256.
- McCulloch, C.E., Searle, S.R. & Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models, Second Edition*. New York: John Wiley & Sons.
- Meng, X.-L., Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.
- Moguerza, J.M. & Muñoz, A. (2006). Support vector machines with applications (with discussion). *Statistical Science*, **21**, 322–362.
- Morris, J.S., Arroyo, C., Coull, B., Ryan, L.M., Herrick, R. & Gortmaker, S.L. (2006). Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *Journal of the American Statistical Association*, **101**, 1352–1364.
- Morris, J.S., Brown, P.J., Herrick, R.C., Baggerly, K.A. & Coombes, K.R. (2007). Bayesian analysis of mass spectrometry data using wavelet based functional mixed models. *Biometrics*, **64**, 479–489.
- Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, **68**, 179–199.
- Morris, J.S., Vannucci, M., Brown, P.J. & Carroll, R.J. (2003). Wavelet-based nonparametric mod-

- eling of hierarchical functions in colon carcinogenesis. *Journal of the American Statistical Association*, **98**, 573–597.
- Namata, H., Shkedy, Z., Faes, C., Aerts, M., Molenberghs, G., Theeten, H., Van Damme, P. & Beutels, Ph. (2007). Estimation of the force of infection from current status data using generalized linear mixed models. *Journal of Applied Statistics*, **34**, 923–939.
- Neal, R. M. (2003). Slice sampling (with discussion) *Annals of Statistics*, **31**, 705–767.
- Ngo, L. & Wand, M.P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, Volume 9, Article 1.
- Nott, D. (2006). Semiparametric estimation of mean and variance functions for non-Gaussian data. *Computational Statistics*, **21**, 603–620.
- Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2 . *The Annals of Statistics*, **13**, 984–997.
- Nychka, D. (2007). `fields` 4.1. R package. <http://cran.r-project.org>.
- O’Connell, M.A. & Wolfinger, R.D. (1997). Spatial regression models, response surfaces, and process optimization. *Journal of Computational and Graphical Statistics*, **6**, 224–241.
- Ogden, R.T. (1996). *Essential Wavelets for Statistical Applications and Data Analysis*. Boston: Birkhauser.
- Opsomer, J.D., Breidt, F.J., Moisen, G.G. & Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models (with discussion). *Journal of the American Statistical Association*, **102**, 400–416.
- Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. & Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, **70**, 265–286.
- Ormerod, J.T., Wand, M.P. & Koch, I. (2008). Penalised spline support vector classifiers: computational issues. *Computational Statistics*, <http://dx.doi.org/10.1007/s00180-007-0102-8>.
- O’Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, **1**, 505–527.
- Paciorek, C.J. (2007a). Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package. *Journal of Statistical Software*, Volume 19, Issue 2.
- Paciorek, C.J. (2007b). `spectralGP` 1.1. R package. <http://cran.r-project.org>.
- Paciorek, C.J. (2007c). Computational techniques for spatial logistic regression with large data sets. *Computational Statistics and Data Analysis*, **51**, 3631–3653.
- Paciorek, C.J. & Schervish, M.J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, **17**, 483–506.
- Parker, R.L. & Rice, J.A. (1985). Comment on paper by Silverman. *Journal of the Royal Statistical Society, Series B*, **47**, 40–42.
- Pearce, N.D. & Wand, M.P. (2006). Penalized splines and reproducing kernel methods. *The American Statistician*, **60**, 233–240.
- Peng, R.D., Dominici, F. & Louis, T.A. (2006). Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistic Society, Series A*, **169**, 179–203.

- Piepho, H.P. & Ogutu, J.O. (2007). Simple state-space models in a mixed model framework. *The American Statistician*, **61**, 224–232.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and the R Core team. (2008). nlme 3.1. R package. <http://cran.r-project.org>.
- Qin, L. & Guo, W.S. (2006). Functional mixed-effects model for periodic data. *Biostatistics*, **7**, 225–234.
- Qu, A. & Li, R.Z. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics*, **62**, 379–391.
- R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org>.
- Ramsay, J.O. & Silverman, B. W. (1997). *Functional Data Analysis*. New York: Springer-Verlag.
- Ramsay, J.O. & Silverman, B. W. (2002). *Applied Functional Data Analysis*. New York: Springer-Verlag.
- Rasmussen, C.E. & Williams, K.I. (2006). *Gaussian Processes for Machine Learning*, The MIT Press.
- Reiss, P.T. & Ogden, P.T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, **102**, 984–996.
- Ribeiro, P.J. & Diggle, P.J. (2008). geoR 1.6. R package. <http://cran.r-project.org>.
- Rice, J. (1986). Convergence rates for partially splined models. *Statistical and Probability Letters*, **4**, 203–208.
- Roca-Pardinas, J., Cadarso-Suarez, C., Nacher, V. & Acuna, C. (2006). Bootstrap-based methods for testing factor-by-curve interactions in generalized additive models: assessing prefrontal cortex neural activity related to decision-making. *Statistics in Medicine*, **25**, 2483–2501.
- Ruppert, D., Nettleton, D., and Hwang, J. T. G. (2008). Exploring the information in p-values for the analysis and planning of multiple-test experiments, *Biometrics*, **63**, 483–495.
- Ruppert, D., Wand, M. P. & Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.
- Sain, S.R., Jagtap, S., Mearns, L. & Nychka, D. (2006). A multivariate spatial model for soil water profiles. *Journal of Agricultural, Biological and Environmental Statistics*, **11**, 462–480.
- SAS Institute, Incorporated (2008). Cary, North Carolina, USA.
- Schapire, R.E. (1990). The strength of weak learnability. *Machine Learning*, **5**, 197–227.
- Scheipl, F. (2007). RLRsim 1.0 R package. <http://cran.r-project.org>
- Self, S.G. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.
- Sergeant, J.C. & Firth, D. (2006). Relative index of inequality: definition, estimation, and inference. *Biostatistics*, **7**, 213–224.
- Silverman, B.W. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, **12**, 898–916.

- Sisson, S.A., Fan, Y. & Tanaka, M.M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences in the United States of America*, **104**, 1760–1765.
- Skaug, H.J. & Fournier, D.A. (2006). Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics and Data Analysis*, **51**, 699–709.
- Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, Florida: Chapman & Hall.
- Smith, A.D.A.C. & Wand, M.P. (2008). Streamlined variance calculations for semiparametric mixed models. *Statistics in Medicine*, **27**, 435–448.
- Smith, M. & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317–344.
- Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 413–436.
- Speed, T. (1991). Comment on paper by Robinson. *Statistical Science*, **6**, 42–44.
- StataCorp LP (2008). College Station, Texas, USA.
- Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. P. (2008). Density estimation in the presence of heteroskedastic measurement error. *Journal of the American Statistical Association*, **103**, 726–736.
- Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.
- Stone, C.J. (1982). Optimal rate of convergence for nonparametric regression. *The Annals of Statistics*, **10**, 1040–1053.
- Stone, C.J., Hansen, M.H., Kooperberg, C. and Truong, Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, **25**, 1371–1425.
- Stram, D.O. & Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171–1177.
- Sturtz, S., Gelman, A., Ligges, U., Gorjanc, G. & Kerman, J. (2007). R2WinBUGS 2.1. R package. <http://cran.r-project.org>.
- Sturtz, S., Ligges, U. & Gelman, A. (2005). R2WinBUGS: A packages for running WinBUGS from R. *Journal of Statistical Software* Volume 12, Issue 3.
- Takeuchi, I., Le, Q. V., Sears, T. D. & Smola, A.J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, **7**, 1231–1264.
- Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia: Society of Industrial and Applied Mathematics.
- Thomas, A., O'Hara, B., Ligges, U. & Sturtz, S. (2006). Making BUGS open. *R News*, Volume 6/1. 12–17.
- Thompson, R. (1985). Comment on paper by Silverman. *Journal of the Royal Statistical Society, Series B*, **47**, 43–44.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological*, **58**, 267–288.
- Tukey, J.W. (1977). *Exploratory Data Analysis*, Reading, Massachusetts: Addison-Wesley.

- Tutz, G. (2003). Generalized semiparametrically structured ordinal models. *Biometrics*, **59**, 263–273.
- Tutz, G. (2004). Generalized semiparametrically structured mixed models. *Computational Statistics and Data Analysis*, **46**, 777–800.
- Tutz, G. & Binder, H. (2004). Flexible modelling of discrete failure time including time-varying smooth effects. *Statistics in Medicine*, **23**, 2445–2461.
- Tutz, G. & Binder, H. (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics*, **62**, 961–971.
- Tutz, G. & Leitenstorfer, F. (2007). Generalized smooth monotonic regression in additive modelling. *Journal of Computational and Graphical Statistics*, **16**, 165–188.
- Tutz, G. & Reithinger, F. (2007). A boosting approach to flexible semiparametric mixed models. *Statistics in Medicine*. **26**, 2872–2900.
- Tutz, G. & Scholz, T. (2004). Semiparametric modelling of multicategorical data. *Journal of Statistical Computation and Simulation*, **74**, 183–200.
- Vandenhende, F., Eilers, P., Ledent, E. & Renard, D. (2007). Joint detection of important biomarkers and optimal dose-response model using penalties. *Statistics in Medicine*, **26**, 4876–4888.
- Verbyla, A.P. (1994). Testing linearity in generalized linear models. *Contributed Paper, 17th International Biometrics Conference, Hamilton, Canada.*, 177.
- Verbyla, A.P., Cullis, B.R., Kenward, M.G. & Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Applied Statistics*, **48**, 269–312.
- Wager, C.G., Coull, B.A. & Lange, N. (2004). Modelling spatial intensity for replicated inhomogeneous point patterns in brain imaging. *J. Royal Statistical Soc. Series B-statistical Methodology*, **66**, 429–446.
- Wager, C., Vaida, F. & Kauermann, G. (2007). Model selection for penalized spline smoothing using Akaike information criteria. *Australian and New Zealand Journal of Statistics*, **49**, 173–190.
- 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. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.
- Wahba, G. (2006). Comment on paper by Moguerza and Muñoz. *Statistical Science*, **21**, 347–351.
- Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, **18**, 223–249.
- Wand, M.P., Coull, B.A., French, J.L., Ganguli, B., Kammann, E.E., Staudenmayer, J. & Zanobetti, A. (2007). SemiPar 1.0. R package. <http://cran.r-project.org>
- Wand, M.P. & Ormerod, J.T. (2008). On O’Sullivan penalised splines and semiparametric regression. *Australian and New Zealand Journal of Statistics*, **50**, 179–198.
- Wang, H.N. & Ranalli, M.G. (2007). Low-rank smoothing splines on complicated domains. *Biometrics*, **63**, 209–217.
- Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. **90**, *Biometrika*, 43–52.

- Wang, N., Carroll, R.J. & Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, **100**, 147–157.
- Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, **60**, 159–174.
- Welham, S.J. (2008). Smoothing spline models for longitudinal data. In Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (Eds.) (2008). *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. Boca Raton, Florida: Chapman & Hall/CRC.
- Welham, S.J., Cullis, B.R., Kenward, M.G. & Thompson, R. (2006). The analysis of longitudinal data using mixed model L-splines. *Biometrics*, **62**, 392–401.
- Welham, S.J., Cullis, B.R., Kenward, M.G. & Thompson, R. (2007). A comparison of mixed model splines for curve fitting. *Australian and New Zealand Journal of Statistics*, **49**, 1–23.
- Welsh, A. H., Lin, X. & Carroll, R.J. (2002). Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. *Journal of the American Statistical Association*, **97**, 482–493.
- Wikle, C. (2002). Spatial modeling of count data: a case study in modeling breeding bird survey data on large spatial domains. In A. Lawson, D. Denison (eds.) *Spatial Cluster Modelling*, pp. 199–209. Chapman & Hall.
- Wood, S.N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B*, **65**, 95–114.
- Wood, S.N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99**, 673–686.
- Wood, S.N. (2006a). *Generalized Additive Models: An Introduction with R*. Boca Raton, Florida: Chapman & Hall/CRC.
- Wood, S.N. (2006b). On confidence intervals for generalized additive models based on penalized regression splines. *Australian and New Zealand Journal of Statistics*, **48**, 445–464.
- Wood, S.N. (2006c). Low-Rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, **62**, 1025–1036.
- Wood, S.N. (2008). mgcv 1.3. R package. <http://cran.r-project.org>.
- Wu, H. & Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*. Hoboken, New Jersey: Wiley.
- Yao, F. & Lee, T.C.M. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society, Series B*, **68**, 3–25.
- Yee, T.W. (2004). Quantile regression via vector generalized additive models. *Statistics in Medicine*, **23**, 2295–2315.
- Yee, T.W. (2008). VGAM 0.7 R package. <http://cran.r-project.org>
- Yee, T.W. & Hastie, T.J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, **3**, 15–41.
- Yee, T.W. & Stephenson, A.G. (2007). Vector generalized linear and additive extreme value models. *Extremes*, **10**, 1–19.
- Yee, T.W. & Wild, C.J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society, Series B*, **58**, 481–493.

- Yu, Y. & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, **97**, 1042–1054.
- Yu, Y. & Ruppert, D. (2004). Root-n consistency of penalized spline estimator for partially linear single-index models under general Euclidean space. *Statistica Sinica*, **14**, 449–456.
- Yuan, Y. & Little, R.J.A. (2007). Parametric and semiparametric model-based estimates of the finite population mean for two-stage cluster samples with item nonresponse. *Biometrics*, **63**, 1172–1180.
- Zeger, S. & Diggle, P.J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689–699.
- Zhang, D. (2004). Generalized linear mixed models with varying coefficients for longitudinal data. *Biometrics*, **60**, 8–15.
- Zhang, D. & Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, **4**, 57–74.
- Zhang, D., Lin, X. & Sowers, M. (2007). Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome. *Biometrics*, **63**, 351–362.
- Zhao, Y., Staudenmayer, J., Coull, B.A. & Wand, M.P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science*, **21**, 35–51.
- Zheng, H. & Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, **19**, 99–117.
- Zheng, H. & Little, R.J.A. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology*, **30**, 209–218.
- Zhou, L., Huang, J.Z. & Carroll, R.J. (2008). Joint modeling of paired sparse functional data using principal components. *Biometrika*, **95**, 601–619.
- Zhou, S., Shen, X. & Wolfe, D.A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, **26**, 1760–1782.
- Wand, M. P. (2009). Semiparametric regression and graphical models. *Australian and New Zealand Journal of Statistics*, **51**, 9–41.