*Aust. N. Z. J. Stat.* **51**(1), 2009, 9–41 doi: 10.1111/j.1467-842X.2009.00538.x

# SEMIPARAMETRIC REGRESSION AND GRAPHICAL MODELS

M. P. WAND[1]

*University of Wollongong*

## Summary

Semiparametric regression models that use spline basis functions with penalization have graphical model representations. This link is more powerful than previously established mixed model representations of semiparametric regression, as a larger class of models can be accommodated. Complications such as missingness and measurement error are more naturally handled within the graphical model architecture. Directed acyclic graphs, also known as Bayesian networks, play a prominent role. Graphical model-based Bayesian 'inference engines', such as BUGS and VIBES, facilitate fitting and inference. Underlying these are Markov chain Monte Carlo schemes and recent developments in variational approximation theory and methodology.

*Key words*: additive models; Bayesian networks; BUGS; directed acyclic graphs; Markov chain Monte Carlo; measurement error models; missing data; mixed models; penalized splines; variational approximation; variational inference; VIBES.

## 1. Introduction

The main thrust of two of my publications from five years ago, Ruppert, Wand & Carroll (2003) and Wand (2003), was that *mixed models* are a very useful framework for carrying out *semiparametric regression* analyses. The thrust of this paper is that the more general *graphical models* framework is also very useful for semiparametric regression, especially when the problem is *non-standard*.

Semiparametric regression is an embellishment of parametric regression that uses *penalized spline* basis functions to achieve greater flexibility. Ruppert *et al.* (2003) surveyed the field up to about 2002. The mixed model aspects of semiparametric regression, and antecedents such as *smoothing splines*, have been known for some time (e.g. Wahba 1978). However, the advent of formal mixed model software in the 1990s led to a surge in research on mixed model approaches to semiparametric regression, mainly in the last decade. A recent survey of semiparametric regression for the period 2003–2007, to be published as Ruppert, Wand & Carroll (2009), revealed more than 150 research articles making use of the mixed model-based semiparametric regression. Sophisticated semiparametric regression analyses are now being routinely carried out, with the 'work' being done by established mixed model software such as lme() (Pinheiro *et al.* 2008) and PROC MIXED (SAS Institute, Inc. 2008),

[1] School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522, Australia. e-mail: mwand@uow.edu.au

*Acknowledgments*. I would like to thank the editors for inviting me to write this paper. I am grateful to John Ormerod for his assistance with the variational approximation examples and to him and Raymond Carroll for their valuable feedback. This research was partially supported by Australian Research Council Discovery Project DP0877055.

or with BUGS (Lunn *et al.* 2000) if a Bayesian approach is adopted. Recent examples are Harezlak *et al.* (2005) and Crainiceanu, Diggle & Rowlingson (2008).

In many applied situations, however, complications such as non-Gaussian response and missingness prevent the use of standard mixed model methodology and software. I will argue that *graphical models* are a better vehicle for fitting and inference in this case. Software for graphical models is less mature but, in 2008, some reasonable options exist. In this article I make use of BUGS (Lunn *et al.* 2000) and VIBES (Bishop, Spiegelhalter & Winn 2003), each of which are *Bayesian inference engines* built upon graphical model architecture. Methodology and software of this type is an ongoing active area of research.

Research into graphical models is currently a very vibrant area, although much of it is taking place in Computer Science rather than in Statistics. The recent book *Pattern Recognition and Machine Learning*, by C.M. Bishop (2006), stated that 'graphical models have emerged as a general framework for describing and applying probabilistic models' in the areas of machine learning and pattern recognition.

The central thesis of the present article is that (non-standard) semiparametric regression can be embedded in graphical model architecture and benefit from ongoing graphical model research. There is also the potential for new applications for semiparametric regression in areas of research that are intrinsically graphical model-based. Some examples are causal inference, social networks and phylogenetic trees.

The proposed marriage of semiparametric regression and graphical models is in keeping with a current general trend that is seeing ideas being exchanged between Statistics and Computer Science much more freely than in earlier days of each discipline. The foreword of a recent special issue of *Statistical Science* on Bayesian Statistics described 'the dissolving of the frontier between Statistics and Computer Science' (Casella & Robert 2004). The special issue contained two review articles, Jordan (2004) and Titterington (2004), of recent Computer Science literature involving Bayesian Statistics. Each of these has had a strong influence on the present article. In 2006, *Statistica Sinica* had a special issue entitled *Challenges in Statistical Machine Learning*.

Semiparametric regression has already benefited from other areas of Computer Science. One spectacular example is boosting (Schapire 1990; Freund 1995; Freund & Schapire 1996). Tutz & Binder (2006) described the evolution from boosting as a means to improve classification procedures to a powerful tool for semiparametric regression analysis and provide relevant references. Also see Bühlmann & Hothorn (2007) and accompanying discussion. Kernel machine research (e.g. Schölkopf & Smola 2002) is another area in which there is a great deal of common ground; see the recent Statistics articles by Pearce & Wand (2006), Wahba (2006) and Hastie & Zhu (2006). To date, there seems to be have been very little interplay between graphical models and semiparametric regression. A rare example of such interplay is Liang, Truong & Wong (2001), who used graphical models in their Bayesian nonparametric regression procedure.

Section 2 summarizes semiparametric regression, focussing on mixed model and hierarchical Bayesian representations. In Section 3 a brief summary of graphical models is provided. A graphical models viewpoint of semiparametric regression is put forward in Section 4. Section 5 then pays special attention to non-standard variants of semiparametric regression. Sections 4 and 5 both work with Bayesian inference engines based on Markov chain Monte Carlo (MCMC) and BUGS software. Section 6 describes an alternative type of inference engine, based on *variational approximation*. A case study involving relative cancer mapping,

when some auxiliary data are missing, is described in Section 7. In Section 8 I add some brief discussion on what might potentially be new areas of application for semiparametric regression, in light of this article's central thesis. Concluding remarks are given in Section 9.

### 1.1. Notation and conventions

Column vectors with entries consisting of subscripted variables are denoted by a bold-faced version of the letter for that variable. For example, the vector containing $x_1, \ldots, x_n$ is denoted by $\mathbf{x}$. Scalar functions applied to vectors are evaluated element-wise. For example, $\tanh(a_1, a_2, a_3) = (\tanh(a_1), \tanh(a_2), \tanh(a_3))$.

The density function of a random vector $\mathbf{x}$ is denoted by $[\mathbf{x}]$. The conditional density of $\mathbf{y}$ given $\mathbf{x}$ is denoted by $[\mathbf{y} \mid \mathbf{x}]$. A random variable $x$ has an inverse gamma distribution with parameters $A$, $B > 0$, denoted by $x \sim \mathrm{IG}(A, B)$, if its density function is $[x] = B^A \Gamma(A)^{-1} x^{-A-1} e^{-B/x}$, $x > 0$. For a general random vector $\mathbf{v}$, $\mathbf{v} \sim (\mu, \Sigma)$ is shorthand for $\mathrm{E}(\mathbf{v}) = \mu$ and $\mathrm{cov}(\mathbf{v}) = \Sigma$, the covariance matrix of $\mathbf{v}$. If, for $1 \leqslant i \leqslant n$, $y_i$ has distribution $D_i$ and the $y_i$ are independent, then I write $y_i \overset{\text{ind.}}{\sim} D_i$.

There are several directed graphs in this article. I use the same conventions as Bishop (2006). Random nodes are denoted by open circles. Non-random nodes are shown as small solid circles. Observed ('evidence') nodes are distinguished from parameter ('hidden') nodes using shading.

All Bayesian models are fitted using standardized versions of continuous variables. Unless otherwise stated, MCMC examples use a burn-in period of 5000 iterations and then retain 5000 iterations. They are then thinned by a factor of 5, resulting in samples of size 1000 being retained for inference.

## 2. Semiparametric regression

Three examples of semiparametric regression models are

$$y_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\left[\text{logit}^{-1}\{\beta_1 x_{1i} + f_2(x_{2i}) + f_{34}(x_{3i}, x_{4i})\}\right], \quad 1 \leqslant i \leqslant n, \tag{1}$$

$$y_i \overset{\text{ind.}}{\sim} \text{Poisson}\left[\exp\{\beta_0(x_{1i}) + \beta_1(x_{1i})x_{2i}\}\right], \quad 1 \leqslant i \leqslant n, \tag{2}$$

$$\begin{aligned} y_{ij} \mid u_{i,\text{sbj}} &\overset{\text{ind.}}{\sim} N\left(u_{i,\text{sbj}} + f_1(x_{1i}) + \beta_2^\top \mathbf{x}_{2i}, \sigma_\varepsilon^2\right), \\ u_{i,\text{sbj}} &\overset{\text{ind.}}{\sim} N\left(0, \sigma_{\text{sbj}}^2\right), \ 1 \leqslant j \leqslant n_i, \ 1 \leqslant i \leqslant m. \end{aligned} \tag{3}$$

Model (1) is an extension of the generalized additive model paradigm that allows non-parametric bivariate components. If $(x_{3i}, x_{4i})$ corresponds to geographic position, then (1) is sometimes called a *geoadditive model* (e.g. Kammann & Wand 2003). In Model (2), $\beta_0$ and $\beta_1$ are smooth functions of the $x_1$ variable. This model is known as a Poisson *varying coefficient model*. Model (3) is usually called an *additive mixed model*, as it represents the fusion of an additive model and a linear mixed model.

An example data set that might benefit from (3) is shown in Figure 1. The source and description of the data are given in Section 4. A question of interest is how the response variable, spinal bone mineral density, differs among the four ethnicity groups. However, the

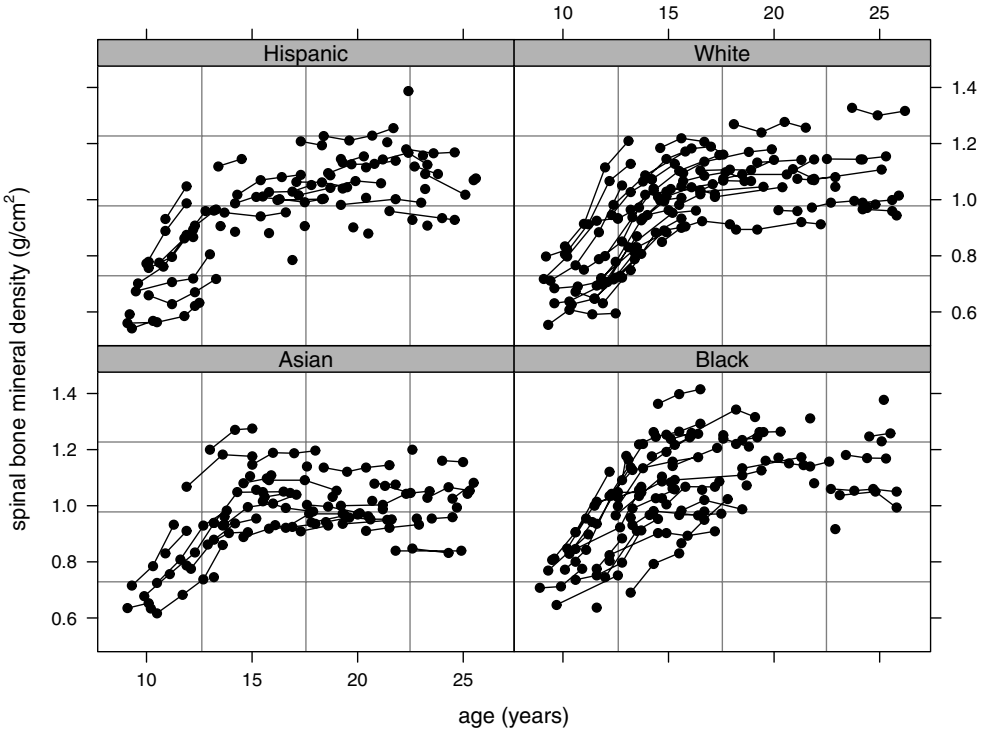


Figure 1. Data on spinal bone mineral density versus age, data broken down according to ethnicity of the subjects. Points for the same subject are connected by lines.

analysis is complicated by (a) the non-linear effect of age, and (b) correlation arising from repeated measurements on the same subject.

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

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

Here $z_1, \ldots, z_K$ are a set of spline basis functions. The simplest example is $z_k(x) = (x - \kappa_k)_+$ for some knot sequence $\kappa_1, \ldots, \kappa_K$. Here $u_+$ equals $u$ for $u \geqslant 0$ and equals 0 otherwise. However, more sophisticated options now exist: see, for example, Wood (2003), Welham *et al.* (2007) and Wand & Ormerod (2008). Most of the spline bases described in these three references 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 2006).

The upshot of (4) is that most semiparametric regression models are expressible as

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

Here $g$ is a scalar 'link' function. The fixed effects term, $\mathbf{X}\boldsymbol{\beta}$, handles covariates that enter the model linearly, whereas the random effects component $\mathbf{Zu}$, with corresponding covariance matrix $\mathbf{G}$, handles non-linear effects, random subject effects and other spatial correlation structure. There will often be other parameters arising, for example, in the variance structure (e.g. $\mathbf{R} = \text{cov}(\mathbf{y} \mid \mathbf{u})$), but I will ignore this in the current discussion.

The hierarchical Bayesian version of (5) takes the form

$$[\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{u}] = f_1(\mathbf{y}; \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}); \quad [\mathbf{u} \mid \mathbf{G}] = f_2(\mathbf{u}; \mathbf{G})$$
$$[\boldsymbol{\beta}] = f_3(\boldsymbol{\beta}; \mathbf{A}_{\boldsymbol{\beta}}); \qquad [\mathbf{G}] = f_4(\mathbf{G}; \mathbf{A}_{\mathbf{G}}), \tag{6}$$

where $\mathbf{A}_{\beta}$ and $\mathbf{A}_{\mathbf{G}}$ are hyperparameters, $f_1, \ldots, f_4$ are fixed conditional density functions and $[\mathbf{v} \mid \mathbf{w}]$ denotes the conditional density of $\mathbf{v}$ given $\mathbf{w}$. Inference is based on posterior densities for parameters of interest, in particular

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

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

Semiparametric regression, especially by means of mixed models and hierarchical Bayesian approaches, is now a major branch of Statistics. In Ruppert *et al.* (2009) we reviewed literature on the topic for the period 2003–2007 and found about 300 papers with connections to semiparametric regression. About 100 of these were in non-Statistics journals. Applications include quantitative trait prediction (Gianola, Fernando & Stella, 2006), modelling of on-line auctions (Jank & Shmueli 2007) and disease mapping (Crainiceanu *et al.* 2008).

Graphical models, described in the next section, can be used for both frequentist and Bayesian statistical models. For the remainder of this article I will restrict attention to Bayesian semiparametric regression. This is in keeping with the graphical model software used in the examples.

## 3. Graphical models

The field of graphical models is a relatively young branch of mathematics that combines ideas from graph theory and probability. Sometimes known as *probabilistic graphical models*, they facilitate the visualization of probability models. Graph-theoretic results have been established for determining conditional independence relationships and for devising efficient algorithms for inference. The fundamental components of a graph are *nodes* and *edges*, which link pairs of nodes. Directed graphs add an arrow-head to each link, conveying a *parent–child* relationship. Several examples of graphs are given later in this section.

Even though there are instances of graphical models in probabilistic and statistical contexts going back several decades (e.g. Wright 1934; Besag 1974; Geman & Geman 1984), the 1980s saw the emergence of substantive theoretical advances and their use in applications. Much of this occurred outside mainstream Statistics, and was driven mainly by applications in Machine Learning and Pattern Recognition. Pearl (1988) is a watershed book on modern graphical models. It was soon followed by a number of others: Whittaker (1990), Jensen (1996), Lauritzen (1996), Castillo, Gutiérrez & Hadi (1997), Jordan (1999) and Cowell *et al.*

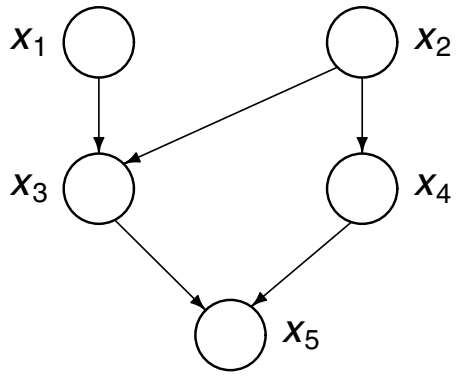


Figure 2. A directed acyclic graph involving five random variables: $x_1, \ldots, x_5$.

(1999). The preparation of this paper has been aided by summaries of the field contained in Jordan (2004), Wasserman (2004, Chapter 17) and Bishop (2006, Chapter 8), and each is highly recommended as background reading for the current article. I also adopt the conventions of Bishop (2006) for displaying graphs.

There are two main types of graphical models: *directed acyclic graphs (DAGs)*, also known as *Bayesian networks*, and *undirected graphs*, also known as *Markov random fields*. Of these, DAGs are more immediately relevant to semiparametric regression, and attention will be restricted to this subclass of graphical models.

An elementary example of a DAG is given in Figure 2. The $x_1, \ldots, x_5$ are random variables corresponding to each of the nodes. The joint density of the $x_i$s defined by this graph takes the form

$$[x_1, x_2, x_3, x_4, x_5] = \prod_{i=1}^{5} [x_k \mid \text{parents of } x_k] = [x_1][x_2][x_3 \mid x_1, x_2][x_4 \mid x_2][x_5 \mid x_3, x_4].$$

More generally, a DAG with $N$ nodes corresponding to the random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ has its joint distribution given by

$$[\mathbf{x}_1, \ldots, \mathbf{x}_N] = \prod_{k=1}^{N} [\mathbf{x}_k \mid \text{parents of } \mathbf{x}_k].$$

Of particular relevance to this article is the DAG representation of hierarchical Bayesian models. Consider a Bayesian version of simple linear regression:

$$\begin{aligned} y_i \mid \beta_0, \beta_1, \sigma^2 &\overset{\text{ind.}}{\sim} N\left(\beta_0 + \beta_1 x_i, \sigma^2\right), \quad 1 \leqslant i \leqslant n, \\ \beta_0 \sim N\left(\mu_{\beta_0}, \sigma^2_{\beta_0}\right), &\quad \beta_1 \sim N\left(\mu_{\beta_1}, \sigma^2_{\beta_1}\right), \quad \sigma^2 \sim \text{IG}(A, B). \end{aligned} \tag{7}$$

Then Figure 3 shows (7) represented as a DAG. Constant nodes, corresponding to the hyperparameters and $x_i$s, are shown as small solid circles. The shading of the $y_i$ nodes indicates replacement by observed values. Bayesian inference involves conditioning on these

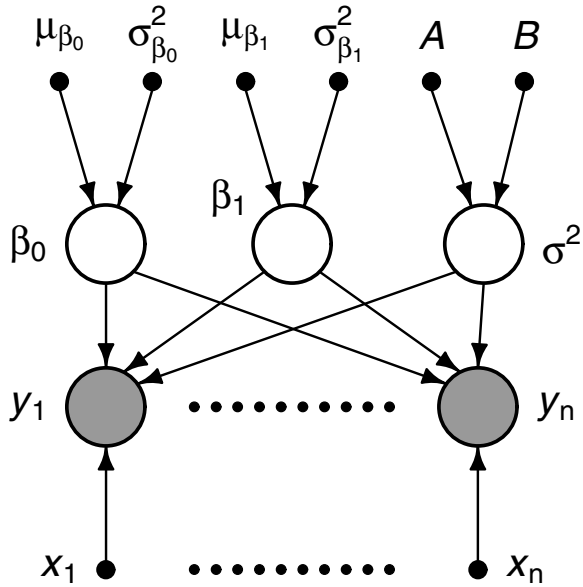


Figure 3. Directed acyclic graph representation of a hierarchical Bayesian simple linear regression model. Shaded nodes correspond to observed data.

nodes to obtain posterior densities. For example, the posterior density of the slope parameter $\beta_1$ is $[\beta_1 \mid y_1, \ldots, y_n]$.

More compact versions of Figure 3 are shown in Figure 4. In Figure 4(a) we introduce a *plate*, shown here as a rectangle, for the $x_i$ and $y_i$ nodes. The plate convention is that all subscripted nodes inside the plate represent several nodes corresponding to the subscript ranging from 1 to the number in the bottom right-hand corner of the plate. Panel (b) of Figure 4 suppresses the constant nodes, and conveys only the essential probabilistic structure of the model. In Figure 4(c) we replace the $y_i$ nodes by a single node for the random vector $\mathbf{y} = (y_1, \ldots, y_n)$ and the $\beta_0$ and $\beta_1$ nodes by one for $\boldsymbol{\beta} = (\beta_0, \beta_1)$. This graph hides the fact

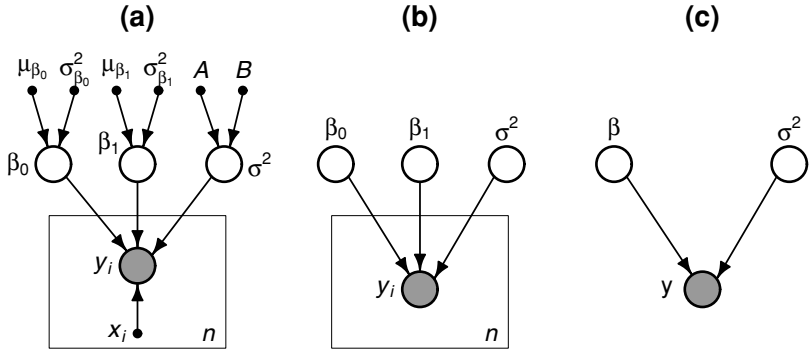


Figure 4. Compact graphical representations of a hierarchical Bayesian logistic regression model. Panel (a) uses the plate convention for the subscripted nodes. In panel (b) the constant nodes are suppressed. Panel (c) treats the vectors $\mathbf{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)$ as entities.

that the $y_i$ are conditionally independent given the parameters, but provides a particularly succinct summary of (7).

The nodes corresponding to the parameters of a hierarchical Bayesian model are often referred to as *hidden* in the literature on graphical models. The observed data correspond to *evidence* nodes. In Figure 4(c) the evidence node is

$$\mathcal{E} = \{\mathbf{y}\},$$

and the set of hidden nodes is

$$\mathcal{H} = \{\boldsymbol{\beta}, \sigma^2\}.$$

Bayesian inference relies upon

$$[\mathcal{H} \mid \mathcal{E}] = \frac{[\mathcal{H}, \mathcal{E}]}{[\mathcal{E}]} = \frac{[\mathbf{y}, \boldsymbol{\beta}, \sigma^2]}{[\mathbf{y}]} = \frac{[\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2][\boldsymbol{\beta}][\sigma^2]}{\int_0^\infty \int_{\mathbb{R}^2} [\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2][\boldsymbol{\beta}][\sigma^2]\, d\boldsymbol{\beta}\, d\sigma^2},$$

the posterior density of the parameters given the data.

For general hierarchical Bayesian models, the probability calculus required to make inference about parameters of interest can be aided by DAG representation and graph-theoretic results. An example of such a result is that, conditional on its parents, each node is independent of the rest of the graph except for its descendants.

Another example of graph theory involves the notion of *Markov blankets* (Pearl 1988). The Markov blanket of a node is defined to be the set of its parents, co-parents and children. The Markov blanket of a node separates it from the rest of the graph, in that conditioning on it renders that node independent from the rest of the graph. An illustration of the use of Markov blankets is provided in Section 4. There is also the theory of *d-separation* (Geiger, Verma & Pearl 1990), which provides necessary and sufficient conditions for two sets of nodes in a DAG to be independent after conditioning on a third set of nodes.

DAG representation of hierarchical Bayesian models has had a profound influence on Bayesian inference since the early 1990s. As pointed out in Jordan (2004), systematic application of graph-theoretic algorithms to Bayesian inference problems has led to so-called Bayesian 'inference engines' (Cowell *et al.* 1999), and is exploited by the popular BUGS software.

Numerous packages in the R language are concerned with graphical models and are summarized on the web-site *CRAN Task View: gRaphical Models in R*. At the time of writing, this web-site has the address cran.r-project.org/web/views/gR.html.

## 4. Graphical models viewpoint of semiparametric regression

As discussed in Section 2, many semiparametric regression analyses can be couched in the framework of hierarchical Bayesian models. These, in turn, have natural representations as DAGs. As an illustration, consider the Bayesian additive mixed model for the data shown in Figure 1. It consists of longitudinal measurements on the spinal bone mineral density (SBMD) of a cohort of young female subjects (source: Bachrach *et al.* 1999). The number of subjects is $m = 230$. Let $n_i$, $1 \leqslant i \leqslant m$, denote the number of measurements for the $i$th subject. One question of interest concerns differences in mean SBMD among the four ethnic groups, Asian, Black, Hispanic and White, after accounting for age. An appropriate model is the *Bayesian additive mixed model*:

$$
\begin{aligned}
&y_{ij} \mid \boldsymbol{\beta}, \mathbf{u}_{\text{sbj}}, \mathbf{u}_{\text{spl}}, \sigma^2_{\text{sbj}}, \sigma^2_{\text{spl}}, \sigma^2_{\varepsilon} \overset{\text{ind.}}{\sim} N\left(\beta_{\text{x}}^{\top}\mathbf{x}_i + u_{i,\text{sbj}} + f\left(\texttt{age}_{ij}; \beta_0, \beta_1, \sigma^2_{\text{spl}}\right), \sigma^2_{\varepsilon}\right),\\
&\mathbf{u}_{\text{sbj}} \,\big|\, \sigma^2_{\text{sbj}} \sim N\left(\mathbf{0}, \sigma^2_{\text{sbj}}\mathbf{I}\right), \quad \mathbf{u}_{\text{spl}} \,\big|\, \sigma^2_{\text{spl}} \sim N\left(\mathbf{0}, \sigma^2_{\text{spl}}\mathbf{I}\right), \quad \beta_0, \beta_1 \sim N\left(\mathbf{0}, \sigma^2_{\beta}\mathbf{I}\right)\\
&\boldsymbol{\beta}_{\text{x}} \sim N\left(\mathbf{0}, \sigma^2_{\beta}\mathbf{I}\right),\ \sigma^2_{\text{sbj}} \sim \text{IG}(A_{\text{sbj}}, B_{\text{sbj}}),\ \sigma^2_{\text{spl}} \sim \text{IG}(A_{\text{spl}}, B_{\text{spl}}),\ \sigma^2_{\varepsilon} \sim \text{IG}(A_{\varepsilon}, B_{\varepsilon}). \qquad (8)
\end{aligned}
$$

Here $y_{ij}$ denotes the $j$th ($1 \leqslant j \leqslant n_i$) SBMD measurement on subject $i$ ($1 \leqslant i \leqslant m$), $\texttt{age}_{ij}$ is the age in years at which $y_{ij}$ was recorded, $\mathbf{x}_i = (1, \texttt{black}_i, \texttt{hispanic}_i, \texttt{white}_i)$, where $\texttt{black}_i$, $\texttt{hispanic}_i$ and $\texttt{white}_i$ are indicator variables for ethnicity (Asian ethnicity is taken as the baseline). In addition, the $u_{i,\text{sbj}}$ are independent and identically distributed (i.i.d.) $N(0, \sigma^2_{\text{sbj}})$ random subject intercepts, and the $\varepsilon_{ij}$ are i.i.d. $N(0, \sigma^2_{\varepsilon})$, independent of the $u_{i,\text{sbj}}$s, and account for within-subject variability. We will model the smooth function for the age effect using penalized splines:

$$
f\left(\texttt{age}; \beta_0, \beta_1, \sigma^2_{\text{spl}}\right) = \beta_0 + \beta_1\texttt{age} + \sum_{k=1}^{K} u_k z_k(\texttt{age}), \qquad u_{k,\text{spl}} \text{i.i.d. } N\left(0, \sigma^2_{\text{spl}}\right),
$$

where the $z_k$ are the spline basis functions described in Wand & Ormerod (2008). A graphical representation of (8) is displayed in Figure 5. In panel (a) the predictors and hyperparameters are included as constant nodes. These are suppressed in Figure 5(b).

In graphical model phraseology, $\mathcal{E} = \{\mathbf{y}\}$ is the evidence node, and

$$
\mathcal{H} = \left\{\beta, \mathbf{u}_{\text{sbj}}, \mathbf{u}_{\text{spl}}, \sigma^2_{\text{sbj}}, \sigma^2_{\text{spl}}, \sigma^2_{\varepsilon}\right\} \qquad (9)
$$

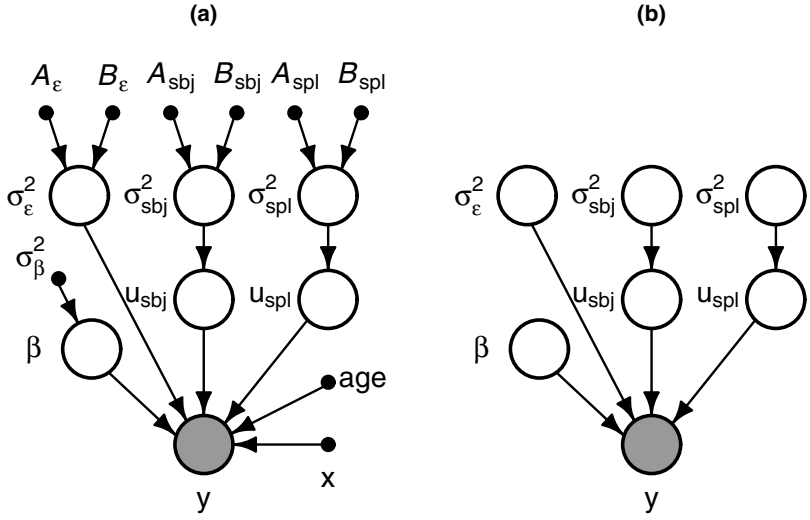


Figure 5. Directed acyclic graph representation of the hierarchical Bayesian model for the spinal bone mineral density data (source: Bachrach *et al.* 1999). In panel (a) hyperparameters and the vector of age values (taken to be deterministic) are shown as small solid circles. In panel (b) these non-random nodes are suppressed.

is the set of the hidden nodes. We wish to learn

$$[\mathcal{H} \,|\, \mathcal{E}] = \left[\beta, \mathbf{u}_{\text{sbj}}, \mathbf{u}_{\text{spl}}, \sigma^2_{\text{sbj}}, \sigma^2_{\text{spl}}, \sigma^2_{\varepsilon} \,\middle|\, \mathbf{y}\right]. \tag{10}$$

The probability calculus required to obtain (10) is somewhat difficult because one gets stuck with intractable integrals that arise from integrating out the variance components. Useful inferential statements, such as credible intervals for the components of $\beta$, are therefore burdensome via direct calculation. *Gibbs sampling* (e.g. Robert & Casella 2004), a special case of MCMC, circumvents this problem by providing samples of arbitrary size from (10). There are several Gibbs sampling options as a result of the various ways in which $\mathcal{H}$ can be partitioned. For the partition corresponding to the nodes of Figure 5, and ordering as in (9), the Gibbs sampling strategy is:

Initialize: $\beta^{[0]}$, $\mathbf{u}_{\text{sbj}}{}^{[0]}$, $\mathbf{u}_{\text{spl}}{}^{[0]}$, $(\sigma^2_{\text{sbj}})^{[0]}$, $(\sigma^2_{\text{spl}})^{[0]}$, $(\sigma^2_{\varepsilon})^{[0]}$.

Cycle: $g = 1, \ldots, B + G$:

$$\begin{aligned}
\beta^{[g]} &\sim \left[\beta^{[g-1]} \,\middle|\, \mathbf{u}_{\text{sbj}}{}^{[g-1]}, \mathbf{u}_{\text{spl}}{}^{[g-1]}, \left(\sigma^2_{\text{sbj}}\right)^{[g-1]}, \left(\sigma^2_{\text{spl}}\right)^{[g-1]}, \left(\sigma^2_{\varepsilon}\right)^{[g-1]}, \mathbf{y}\right],\\
\mathbf{u}_{\text{sbj}}{}^{[g]} &\sim \left[\mathbf{u}_{\text{sbj}}{}^{[g-1]} \,\middle|\, \beta^{[g]}, \mathbf{u}_{\text{spl}}{}^{[g-1]}, \left(\sigma^2_{\text{sbj}}\right)^{[g-1]}, \left(\sigma^2_{\text{spl}}\right)^{[g-1]}, \left(\sigma^2_{\varepsilon}\right)^{[g-1]}, \mathbf{y}\right],\\
\mathbf{u}_{\text{spl}}{}^{[g]} &\sim \left[\mathbf{u}_{\text{spl}}{}^{[g-1]} \,\middle|\, \beta^{[g]}, \mathbf{u}_{\text{sbj}}{}^{[g]}, \left(\sigma^2_{\text{sbj}}\right)^{[g-1]}, \left(\sigma^2_{\text{spl}}\right)^{[g-1]}, \left(\sigma^2_{\varepsilon}\right)^{[g-1]}, \mathbf{y}\right],\\
\left(\sigma^2_{\text{sbj}}\right)^{[g]} &\sim \left[\left(\sigma^2_{\text{sbj}}\right)^{[g-1]} \,\middle|\, \beta^{[g]}, \mathbf{u}_{\text{sbj}}{}^{[g]}, \mathbf{u}_{\text{spl}}{}^{[g]}, \left(\sigma^2_{\text{spl}}\right)^{[g-1]}, \left(\sigma^2_{\varepsilon}\right)^{[g-1]}, \mathbf{y}\right],\\
\left(\sigma^2_{\text{spl}}\right)^{[g]} &\sim \left[\left(\sigma^2_{\text{spl}}\right)^{[g-1]} \,\middle|\, \beta^{[g]}, \mathbf{u}_{\text{sbj}}{}^{[g]}, \mathbf{u}_{\text{spl}}{}^{[g]}, \left(\sigma^2_{\text{sbj}}\right)^{[g]}, \left(\sigma^2_{\varepsilon}\right)^{[g-1]}, \mathbf{y}\right],\\
\left(\sigma^2_{\varepsilon}\right)^{[g]} &\sim \left[\left(\sigma^2_{\varepsilon}\right)^{[g-1]} \,\middle|\, \beta^{[g]}, \mathbf{u}_{\text{sbj}}{}^{[g]}, \mathbf{u}_{\text{spl}}{}^{[g]}, \left(\sigma^2_{\text{sbj}}\right)^{[g]}, \left(\sigma^2_{\text{spl}}\right)^{[g]}, \mathbf{y}\right].
\end{aligned} \tag{11}$$

For a sufficiently high value of the burn-in sample size $B$, the draws

$$\beta^{[g]}, \mathbf{u}_{\text{sbj}}{}^{[g]}, \mathbf{u}_{\text{spl}}{}^{[g]}, \left(\sigma^2_{\text{sbj}}\right)^{[g]}, \left(\sigma^2_{\text{spl}}\right)^{[g]}, \left(\sigma^2_{\varepsilon}\right)^{[g]}, \quad B + 1 \leqslant g \leqslant B + G,$$

are a sample of size $G$ from (10) and can be used for inference. Implementation of this Gibbs sampling scheme requires the *full conditional* densities

$$[\beta \,|\, \text{rest}], \quad [\mathbf{u}_{\text{sbj}} \,|\, \text{rest}], \quad [\mathbf{u}_{\text{spl}} \,|\, \text{rest}], \quad \left[\sigma^2_{\text{sbj}} \,\middle|\, \text{rest}\right], \quad \left[\sigma^2_{\text{spl}} \,\middle|\, \text{rest}\right] \text{ and } \left[\sigma^2_{\varepsilon} \,\middle|\, \text{rest}\right], \tag{12}$$

where 'rest' denotes the nodes in the graph apart from the node appearing before the vertical bar. Determination of (12) benefits from the Markov blanket result stated in Section 3:

$$[\text{node} \,|\, \text{rest}] = [\text{node} \,|\, \text{Markov blanket of node}].$$

The Markov blanket of $\sigma^2_{\text{sbj}}$ is shown in Figure 6.

For this node we then have

$$\begin{aligned}
\left[\sigma^2_{\text{sbj}} \,\middle|\, \text{rest}\right] &= \left[\sigma^2_{\text{sbj}} \,\middle|\, \text{Markov blanket of } \sigma^2_{\text{sbj}}\right] = \left[\sigma^2_{\text{sbj}} \,\middle|\, \mathbf{u}_{\text{sbj}}\right] \propto \left[\mathbf{u}_{\text{sbj}} \,\middle|\, \sigma^2_{\text{sbj}}\right]\left[\sigma^2_{\text{sbj}}\right]\\
&\propto \left(\sigma^2_{\text{sbj}}\right)^{-m/2} \exp\left\{-\left\|\mathbf{u}_{\text{sbj}}\right\|^2 \big/ \left(2\sigma^2_{\text{spl}}\right)\right\} \left(\sigma^2_{\text{sbj}}\right)^{-A_{\text{sbj}}-1} \exp\left(-B_{\text{sbj}} \big/ \sigma^2_{\text{spl}}\right)\\
&= \left(\sigma^2_{\text{sbj}}\right)^{-(A_{\text{sbj}}+m/2)-1} \exp\left\{-\left(B_{\text{sbj}} + \tfrac{1}{2}\|\mathbf{u}_{\text{sbj}}\|^2\right) \Big/ \sigma^2_{\text{spl}}\right\}\\
&\sim \text{IG}\left(A_{\text{sbj}} + \tfrac{1}{2}m,\, B_{\text{sbj}} + \tfrac{1}{2}\|\mathbf{u}_{\text{sbj}}\|^2\right).
\end{aligned}$$

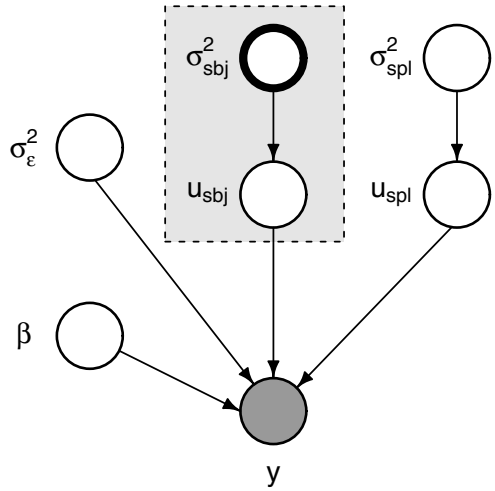


Figure 6. Markov blanket of the node $\sigma^2_{\text{sbj}}$ for the directed acyclic graph in Figure 5.

Continuing in this fashion we obtain the set of full conditionals as:

$$
\begin{aligned}
\boldsymbol{\beta} \mid \text{rest} &\sim N\big(\big\{\mathbf{X}^\top\mathbf{X} + \big(\sigma_\varepsilon^2/\sigma_\beta^2\big)\,\mathbf{I}\big\}^{-1}\mathbf{X}^\top(\mathbf{y} - \mathbf{Z}_{\text{sbj}}\mathbf{u}_{\text{sbj}} - \mathbf{Z}_{\text{spl}}\mathbf{u}_{\text{spl}}),\\
&\qquad \sigma_\varepsilon^2\big\{\mathbf{X}^\top\mathbf{X} + \big(\sigma_\varepsilon^2/\sigma_\beta^2\big)\,\mathbf{I}\big\}^{-1}\big),\\
\mathbf{u}_{\text{sbj}} \mid \text{rest} &\sim N\big(\big\{\mathbf{Z}_{\text{sbj}}{}^\top\mathbf{Z}_{\text{sbj}} + \big(\sigma_\varepsilon^2/\sigma_{\text{sbj}}^2\big)\,\mathbf{I}\big\}^{-1}\mathbf{Z}_{\text{sbj}}{}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_{\text{spl}}\mathbf{u}_{\text{spl}}),\\
&\qquad \sigma_\varepsilon^2\big\{\mathbf{Z}_{\text{sbj}}{}^\top\mathbf{Z}_{\text{sbj}} + \big(\sigma_\varepsilon^2/\sigma_{\text{sbj}}^2\big)\,\mathbf{I}\big\}^{-1}\big),\\
\mathbf{u}_{\text{spl}} \mid \text{rest} &\sim N\big(\big\{\mathbf{Z}_{\text{spl}}{}^\top\mathbf{Z}_{\text{spl}} + \big(\sigma_\varepsilon^2/\sigma_{\text{spl}}^2\big)\,\mathbf{I}\big\}^{-1}\mathbf{Z}_{\text{spl}}{}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_{\text{sbj}}\mathbf{u}_{\text{sbj}}),\\
&\qquad \sigma_\varepsilon^2\big\{\mathbf{Z}_{\text{spl}}{}^\top\mathbf{Z}_{\text{spl}} + \big(\sigma_\varepsilon^2/\sigma_{\text{spl}}^2\big)\,\mathbf{I}\big\}^{-1}\big),\\
\sigma^2_{\text{sbj}} \mid \text{rest} &\sim \text{IG}\left(A_{\text{sbj}} + \frac{1}{2}m,\, B_{\text{sbj}} + \frac{1}{2}\|\mathbf{u}_{\text{sbj}}\|^2\right),\\
\sigma^2_{\text{spl}} \mid \text{rest} &\sim \text{IG}\left(A_{\text{spl}} + \frac{1}{2}K,\, B_{\text{spl}} + \frac{1}{2}\|\mathbf{u}_{\text{spl}}\|^2\right) \text{ and}\\
\sigma^2_{\varepsilon} \mid \text{rest} &\sim \text{IG}\left(A_{\varepsilon} + \frac{1}{2}\sum_{i=1}^{m} n_i,\, B_{\varepsilon} + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}_{\text{sbj}}\mathbf{u}_{\text{sbj}} - \mathbf{Z}_{\text{spl}}\mathbf{u}_{\text{spl}}\|^2\right). \qquad (13)
\end{aligned}
$$

Approximate Bayesian inference can then proceed via implementation of (11) and (13). Implementation in the R language (R Development Core Team 2008) is very straightforward. However, BUGS offers even more immediate results. Note that, in the Windows version of BUGS, known as WinBUGS, there is the option to specify the model using a graphical model drawing facility. Figure 7 is a screen-shot of the graph used for fitting (8) in BUGS.

Figure 8 summarizes the MCMC output and subsequent Bayesian inference for the parameters in (8). In keeping with previously published analyses, a statistically significant difference is found between Black and Asian females in terms of mean spinal bone mineral

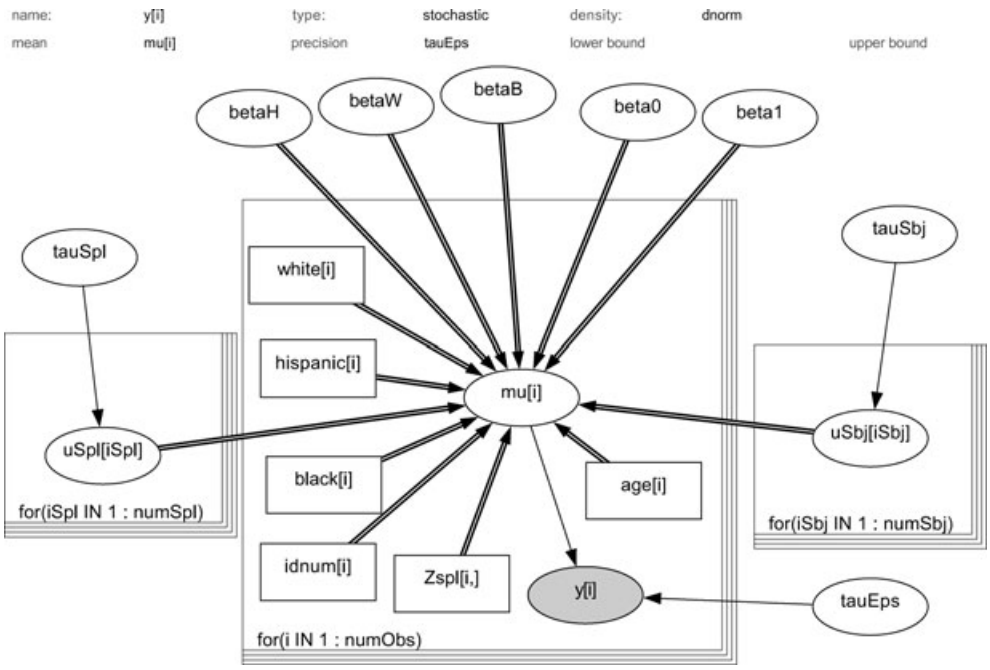


Figure 7. Screen-shot of the directed acyclic graph drawn in WinBUGS for specifying the semiparametric regression model applied to the spinal bone mineral density data.

density. Note that, in Figure 8, $\sigma_{\text{sbj}}$ is replaced by the *(effective) degrees of freedom* for the non-linear age effect (e.g. Buja, Hastie & Tibshirani 1989).

The fitted curves $\widehat{f}(\texttt{age})$, together with 95% pointwise credible sets, are shown in Figure 9.

## 5. Non-standard semiparametric regression

The biggest gains from a graphical models viewpoint of semiparametric regression are realized when the setting is a non-standard one. In 'standard' semiparametric regression, the response variable is approximately Gaussian and all data are cleanly observed. However, in many applications, the data do not conform with this ideal state of affairs, and the analyst has to deal with the likes of categorical response variables, outliers, missingness and measurement error. In this section I focus on three aspects of non-standard semiparametric regression: non-Gaussian response, predictors subject to missingness, and predictors subject to measurement error, and view them through the graphical model prism.

### 5.1. Non-Gaussian response

As is well known, the generalization of regression models to non-Gaussian response variables has its challenges. In semiparametric regression it essentially entails the extension from linear mixed models to generalized linear mixed models. If taking a frequentist likelihood-based approach, then an immediate consequence is intractable integrals. If a Bayesian/MCMC approach is taken, then the full conditionals are no longer standard distributions like those

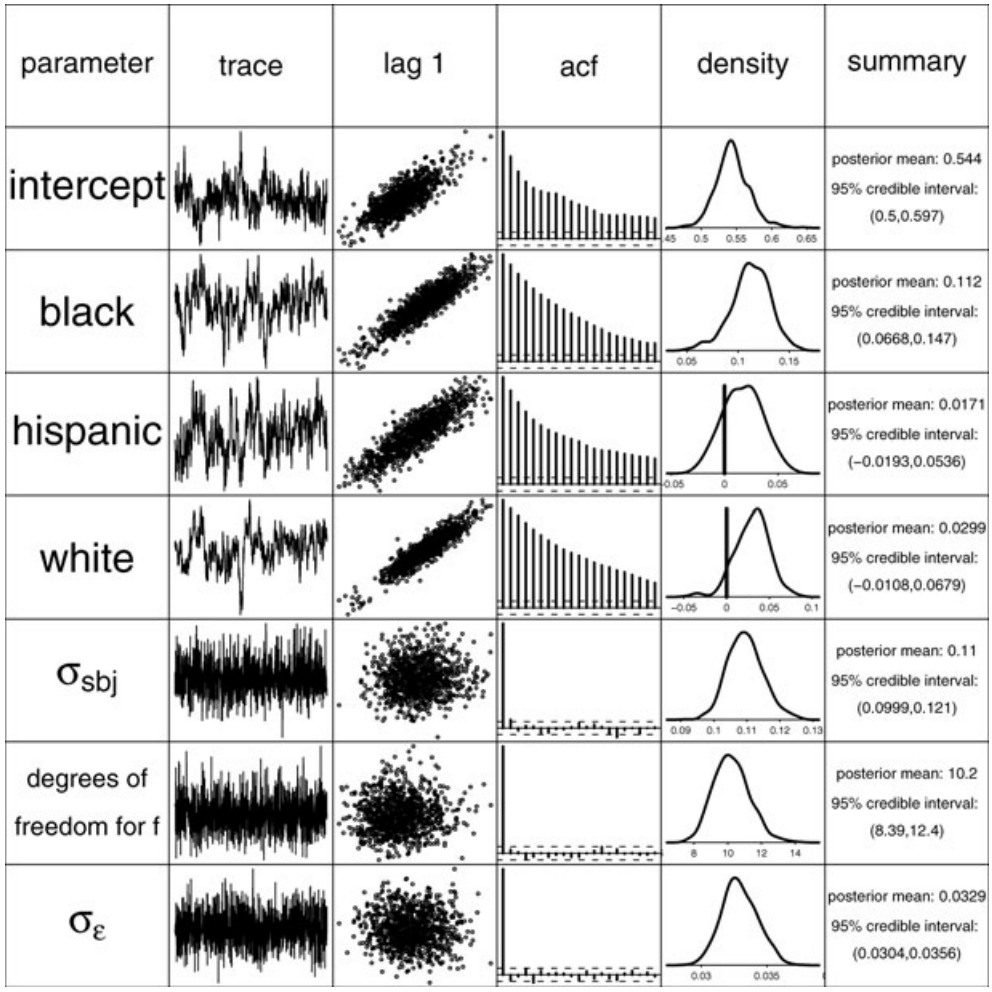


Figure 8. Summary of MCMC-based inference for parameters in the fitted model for the spinal bone mineral density data. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates of posterior density and basic numerical summary.

appearing in (13). More elaborate MCMC schemes, such as Metropolis–Hastings and adaptive rejection sampling (e.g. Robert & Casella 2004), are required. Metropolis–Hastings algorithms can also benefit from the viewpoint of graphical models and Markov blanket theory. As pointed out by Jordan (2004), 'factors that do not appear in the Markov blanket of a set of variables being considered in a proposed update can be neglected.'

Zhao *et al.* (2006) recently studied semiparametric regression for the case in which the response variable distribution is in the one-parameter exponential family. Implementation in BUGS was described, which then means that such semiparametric regression models have a DAG representation. For example, the logistic additive mixed model fitted to data on respiratory infection in Indonesian children in section 4.1 of Zhao *et al.* (2006) has the graphical representation shown in Figure 10. This DAG is very similar to the one in Figure 5

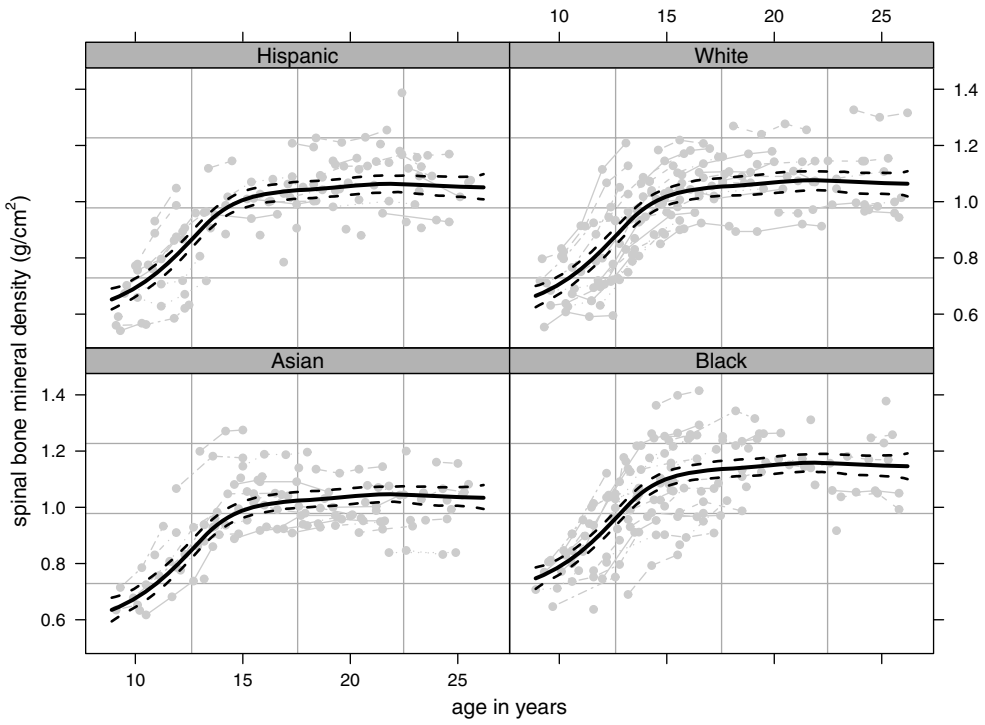


Figure 9. MCMC-based estimate of the non-linear age effect in the spinal bone mineral density example. The dashed lines correspond to pointwise 95% credible sets.

for the spinal bone mineral density example. The main difference is that **y** is now a binary rather than a Gaussian node.

However, many other non-Gaussian response models of interest fall outside the one-parameter exponential family structure. Examples from the semiparametric regression literature include negative binomial (e.g. Thurston, Wand & Weincke 2000), Efron's double exponential family (e.g. Nott 2006), beta (e.g. Branscum, Johnson & Thurmond 2007), Student's $t$ (e.g. Staudenmayer, Lake & Wand 2009), and generalized extreme value (e.g. Yee & Stephenson 2007; Padoan & Wand 2008) distributions. All can be embedded within a graphical models framework. Current joint research with Jennifer K. Marley is investigating BUGS fitting of non-Gaussian response semiparametric regression models such as these.

As the response becomes less standard, the suitability of established mixed model software is less likely, and more general software packages such as BUGS are about the only current option. Essentially, this means that ordinary mixed model architectures are inadequate and that more general graphical model architectures are required. The next two subsections provide even more potent illustrations of this state of affairs.

### 5.2. Predictors subject to missingness

Consider the simple nonparametric regression setting

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d. } N\left(0, \sigma_\varepsilon^2\right), \quad 1 \leqslant i \leqslant n, \tag{14}$$

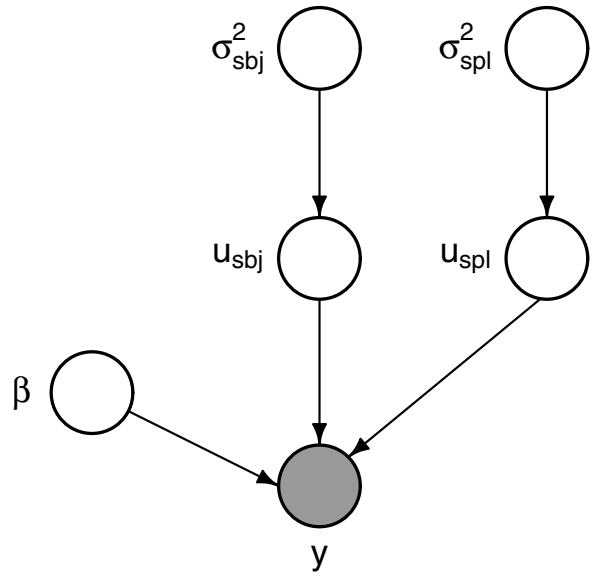


Figure 10. Directed acyclic graph representation of the Bayesian logistic additive mixed model applied to data on respiratory infection in Indonesian children in Zhao *et al.* (2006).

for a smooth function $f$, and assume that the $x_i$s can be modelled as coming from a normal distribution with mean $\mu_x$ and variance $\sigma_x^2$. Suppose, however, that all of the $y_i$s are observed but that some of the $x_i$s are missing. An appropriate hierarchical Bayesian model for this situation is

$$\begin{aligned} &y_i \mid x_i, \boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\left(\beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k z_k(x_i), \sigma_\varepsilon^2\right), \quad \mathbf{u} \,\big|\, \sigma_u^2 \sim N\big(\mathbf{0}, \sigma_u^2 \mathbf{I}\big), \\ &x_i \mid \mu_x, \sigma_x^2 \overset{\text{ind.}}{\sim} N\big(\mu_x, \sigma_x^2\big), \quad \boldsymbol{\beta} \sim N\big(\mathbf{0}, \sigma_\beta^2 \mathbf{I}\big), \quad \mu_x \sim N\big(0, \sigma_{\mu_x}^2\big), \\ &\sigma_u^2 \sim \text{IG}(A_u, B_u), \quad \sigma_\varepsilon^2 \sim \text{IG}(A_\varepsilon, B_\varepsilon), \quad \sigma_x^2 \sim \text{IG}(A_x, B_x). \end{aligned} \tag{15}$$

Let $\mathbf{x}^{\text{obs}}$ be the vector of observed $x_i$s and $\mathbf{x}^{\text{mis}}$ be the missing values. Then the observed data, or evidence nodes, are

$$\mathcal{E} = \{\mathbf{y}, \mathbf{x}^{\text{obs}}\},$$

and the parameters, or hidden nodes, are

$$\mathcal{H} = \big\{\boldsymbol{\beta}, \mathbf{u}, \mathbf{x}^{\text{mis}}, \sigma_u^2, \sigma_\varepsilon^2, \mu_x, \sigma_x^2\big\}.$$

The DAG for (15) is given in Figure 11.

Bayesian inference requires

$$[\mathcal{H} \mid \mathcal{E}] = \big[\boldsymbol{\beta}, \mathbf{u}, \mathbf{x}^{\text{mis}}, \sigma_u^2, \sigma_\varepsilon^2, \mu_x, \sigma_x^2 \,\big|\, \mathbf{y}, \mathbf{x}^{\text{obs}}\big].$$

Note the extra layer of complexity imposed by missingness, because

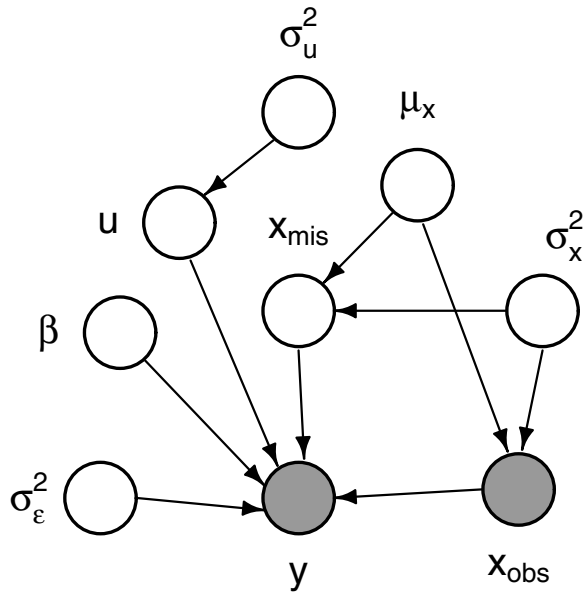


Figure 11. Graphical representation of the penalized spline nonparametric regression model with the predictor subject to missingness. Shading corresponds to the observed, or evidence, nodes.

$$\begin{aligned}&\left[\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2, \mu_x, \sigma_x^2 \,\middle|\, \mathbf{y}, \mathbf{x}^{\text{obs}}\right] \\ &\quad = \frac{\displaystyle\int \left[\mathbf{y} \,\middle|\, \mathbf{x}, \boldsymbol{\beta}, \sigma_\varepsilon^2\right]\left[\mathbf{x} \,\middle|\, \mu_x, \sigma_x^2\right][\boldsymbol{\beta}]\left[\mathbf{u} \,\middle|\, \sigma_u^2\right]\left[\sigma_u^2\right]\left[\sigma_\varepsilon^2\right][\mu_x]\left[\sigma_x^2\right] d\,\mathbf{x}^{\text{mis}}}{\displaystyle\int \left[\mathbf{y} \,\middle|\, \mathbf{x}, \boldsymbol{\beta}, \sigma_\varepsilon^2\right]\left[\mathbf{x} \,\middle|\, \mu_x, \sigma_x^2\right][\boldsymbol{\beta}]\left[\mathbf{u} \,\middle|\, \sigma_u^2\right]\left[\sigma_u^2\right]\left[\sigma_\varepsilon^2\right][\mu_x]\left[\sigma_x^2\right] d\,\mathcal{H}}\end{aligned}$$

now involves integration over the missing data vector $\mathbf{x}^{\text{mis}}$.

I tested BUGS fitting of (15) to simulated data with

$$n = 300, \quad f(x) = \sin(4\pi x), \quad \mu_x = \frac{1}{2}, \quad \sigma_x^2 = \frac{1}{36}, \quad \sigma_\varepsilon^2 = 0.35 \tag{16}$$

and 20% of the $x_i$s missing completely at random. The hyperparameters were set to be

$$\sigma_\beta^2 = \sigma_{\mu_x}^2 = 10^8, \quad A_u = B_u = A_\varepsilon = B_\varepsilon = A_x = B_x = \frac{1}{100}. \tag{17}$$

Because the spline basis functions for the missing $x_i$s have to be computed inside BUGS, I used truncated lines:

$$z_k(x) = (x - \kappa_k)_+ \text{ with } \kappa_k = \left\{(K+1-k)\min\left(x_i^{\text{obs}}\right) + k\max\left(x_i^{\text{obs}}\right)\right\}\Big/(K+1), \;\; 1 \leqslant k \leqslant K$$

and $K = 25$. The relevant BUGS code is

```
model
{
  for(i in 1:nObs)
```

```
    {
    muObs[i] ← beta0 + beta1*xObs[i] + inprod(u[],ZxObs[i,])
    yxObs[i] ~ dnorm(muObs[i],tauEps)
    xObs[i] ~ dnorm(muX,tauX)
    }
    for(i in 1:nMis)
    {
    muMis[i] ← beta0 + beta1*xMis[i] + inprod(u[],ZxMis[i,])
    yxMis[i] ~ dnorm(muMis[i],tauEps)
    xMis[i] ~ dnorm(muX,tauX)
    }
    for (k in 1:K)
    {
    for (i in 1:nMis)
    {
    ZxMis[i,k] ← (xMis[i]-knots[k])*step(xMis[i]-knots[k])
    }
    u[k] ~ dnorm(0,tauU)
    }
    beta0 ~ dnorm(0,1.0E-8) ; beta1 ~ dnorm(0,1.0E-8)
    muX ~ dnorm(0,1.0E-8) ; tauX ~ dgamma(0.01,0.01)
    tauU ~ dgamma(0.01,0.01) ; tauEps ~ dgamma(0.01,0.01)
}
```

where, for example, `yxObs[]` is the vector of $y_i$ values that have an observed $x_i$ partner.

The upper panels of Figure 12 summarize the MCMC output produced by BUGS for the nodes $\mu_x$, $\sigma_x$ and $\sigma_\varepsilon$. The true values, (16), from which the data were simulated are shown as vertical dashed lines in the posterior density plots. The lowest panel monitors the effective degrees of freedom for estimation of $f$. The chains are seen to be reasonably well behaved.

Figure 13 shows the estimate of $f$ as well as pointwise 95% credible intervals. The missing data, known from simulation but hidden from the methodology, are shown as grey circles.

Lastly, we study some of the output for the hidden node $\mathbf{x}^{\text{mis}}$. Five components were chosen at random, and the MCMC summaries are shown in Figure 14. Interestingly, the posterior densities of some of the $\mathbf{x}^{\text{mis}}$ components are multimodal. This arises from the periodic nature of the underlying signal. Knowledge about the ordinate manifests in the posterior of $x_i^{\text{mis}}$ as two or three clumps of probability mass corresponding, roughly, to vertical slicing of $f$ at that ordinate.

We close this subsection by noting that a moderate amount of research on missingness for mixed model-based semiparametric regression now exists. References include French & Wand (2004), Chen & Ibrahim (2006), Geraci & Bottai (2006) and Yuan & Little (2007).

### 5.3. Predictors subject to measurement error

In the previous subsection, the $x_i$s in (15) were subject to missingness. Now suppose instead that they are subject to measurement error. In this case, rather than observing $x_i$ we observe

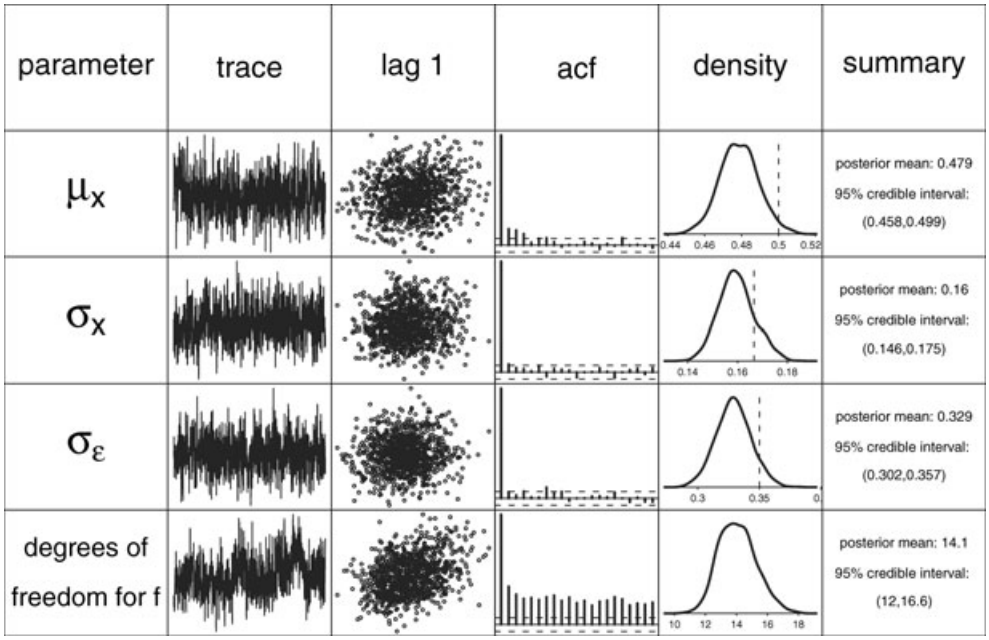


Figure 12. Summary of MCMC-based inference for parameters in the missing predictor nonparametric regression model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates of posterior density and basic numerical summary. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

$$w_i = x_i + z_i, \quad 1 \leqslant i \leqslant n, \tag{18}$$

where the $z_i$ are i.i.d. $N(0, \sigma_z^2)$ and independent of the $x_i$s. The contamination variance, $\sigma_z^2$, is assumed to be known.

This is an instance of nonparametric regression with measurement error. Carroll *et al.* (2006) is a recent survey of this and related topics. A hierarchical Bayesian model for (14) and (18) is

$$\begin{aligned} &y_i \mid x_i, \boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\left(\beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k z_k(x_i), \sigma_\varepsilon^2\right), \mathbf{u} \mid \sigma_u^2 \sim N\left(\mathbf{0}, \sigma_u^2 \mathbf{I}\right), \\ &x_i \mid \mu_x, \sigma_x^2 \overset{\text{ind.}}{\sim} N\left(\mu_x, \sigma_x^2\right), \quad w_i \mid x_i \overset{\text{ind.}}{\sim} N\left(x_i, \sigma_z^2\right), \beta \sim N\left(\mathbf{0}, \sigma_\beta^2 \mathbf{I}\right), \quad \mu_x \sim N\left(0, \sigma_{\mu_x}^2\right), \\ &\sigma_u^2 \sim \text{IG}(A_u, B_u), \quad \sigma_\varepsilon^2 \sim \text{IG}(A_\varepsilon, B_\varepsilon), \quad \sigma_x^2 \sim \text{IG}(A_x, B_x). \end{aligned} \tag{19}$$

The observed data, or evidence nodes, are

$$\mathcal{E} = \{\mathbf{y}, \mathbf{w}\},$$

where $\mathbf{w}$ is the vector of $w_i$s. The set of parameters, or hidden nodes, is

$$\mathcal{H} = \left\{\boldsymbol{\beta}, \mathbf{u}, \mathbf{x}, \sigma_u^2, \sigma_\varepsilon^2, \mu_x, \sigma_x^2\right\}.$$

The graphical representation of (19) is shown in Figure 15.

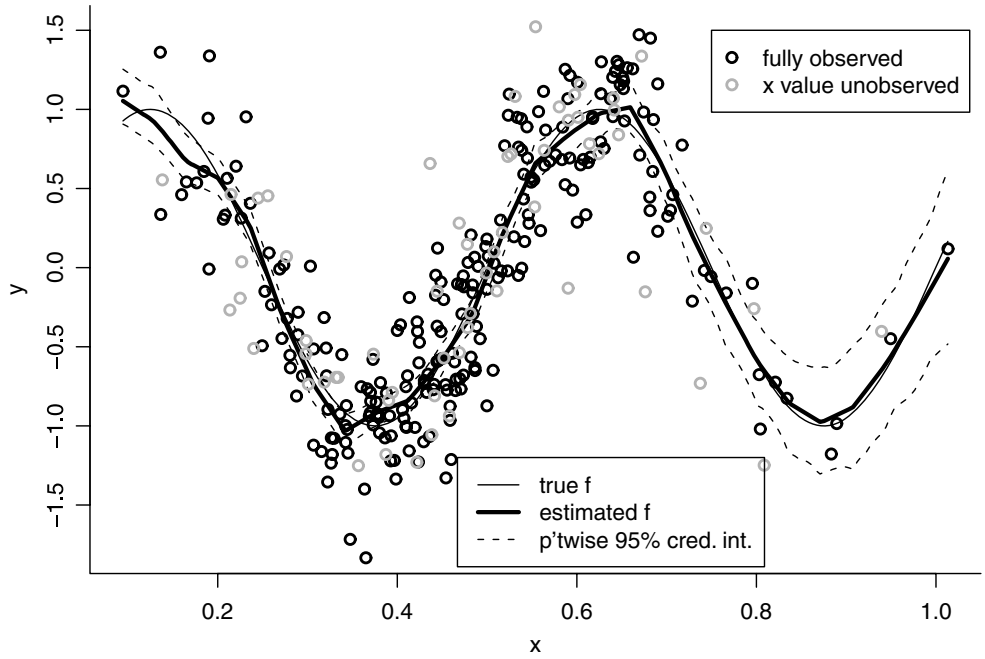


Figure 13. MCMC-based estimate of $f$ in the missing predictor nonparametric regression model. The dashed lines correspond to pointwise 95% credible sets. The grey points are those for which the $x$ values were missing and not used by the fitting procedure.

BUGS fitting of (19) was tested using the parameter settings given by (16) and (17) and with $\sigma_z$ set to be 0.1. As for the missing data example, spline basis functions have to be computed inside BUGS, so I used truncated line basis functions with knots

$$\kappa_k = \{(K+1-k)\min(x_i) + k\max(x_i)\}/(K+1), \quad 1 \leqslant k \leqslant K,$$

(which depend on the hidden **x** node) and $K = 20$. The BUGS code is:

```
model
  {
  for(i in 1:n)
  {
  x[i] ~ dnorm(muX,tauX)
  w[i] ~ dnorm(x[i],tauZ)
  mu[i] <- beta0 + beta1*x[i] + inprod(u[],Z[i,])
  y[i] ~ dnorm(mu[i],tauEps)
  }
  for (k in 1:K)
  {
  knots[k] <- ((K+1-k)*ranked(x[],1)+k*ranked(x[],n))/(K+1)
  for (i in 1:n)
  {
  Z[i,k] <- (x[i]-knots[k])*step(x[i]-knots[k])
```

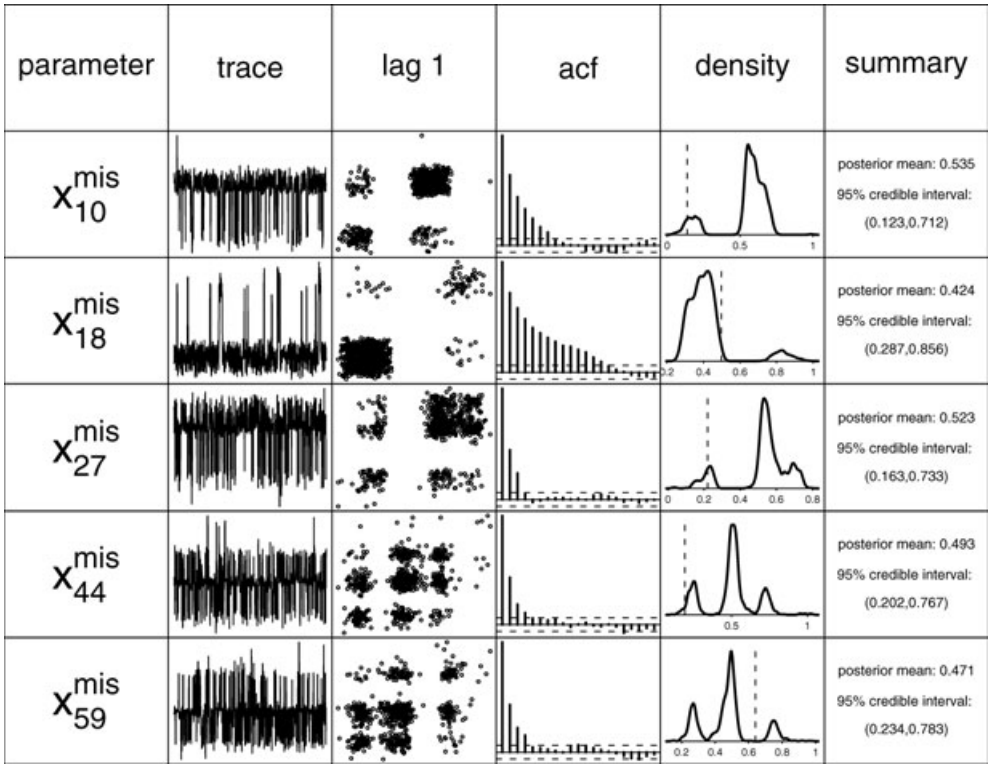


Figure 14. Summary of MCMC-based inference for five randomly chosen missing predictors in the missing predictor nonparametric regression model. The columns are: missing predictor, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates of posterior density and basic numerical summary. The vertical dashed lines in the density plots correspond to the true values of the predictors for the simulation.

```
    }
    u[k] ~ dnorm(0,tauU)
    }
    beta0 ~ dnorm(0,1.0E-8) ; beta1 ~ dnorm(0,1.0E-8)
    muX ~ dnorm(0,1.0E-8) ; tauX ~ dgamma(0.01,0.01)
    tauU ~ dgamma(0.01,0.01) ; tauEps ~ dgamma(0.01,0.01)
}
```

The upper panels of Figure 16 are the analogue of Figure 12 for the current measurement error example. Once again, the chains are seen to be reasonably well behaved, and the true parameters are inside the 95% credible sets.

Figure 17 shows the estimate of $f$ as well as pointwise 95% credible intervals. The grey circles are the unobserved $(x_i, y_i)$ pairs which, because this is a simulation study, are known. The curve estimate is seen to be quite reasonable, despite having to adjust for contamination of the $x_i$s.

Models of type (19) were first formulated by Berry, Carroll & Ruppert (2002).

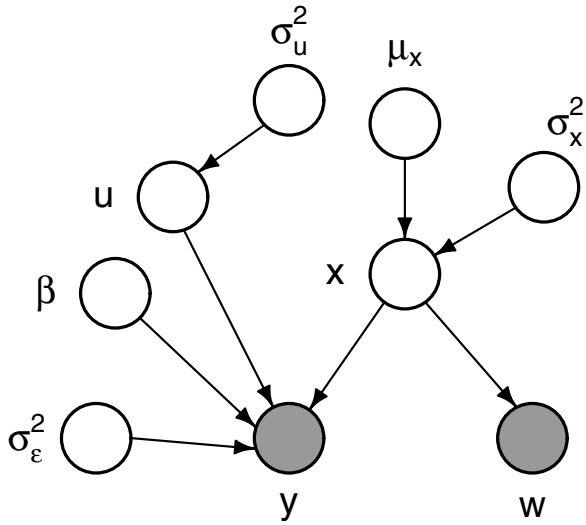


Figure 15. Graphical representation of the penalized spline nonparametric regression model with the predictor subject to measurement error. Shading corresponds to the observed, or evidence, nodes.

## 6. Variational inference engines

Each of the examples in the previous two sections was performed using an MCMC-based Bayesian inference engine, namely BUGS. However, MCMC is computationally intensive, and inference can be quite slow. The example in Section 5.3 involving measurement error took about a day to run on my laptop computer (Mac OS X; 2.33-GHz processor, 3 Gbytes of RAM). An alternative to MCMC, which offers the possibility of much faster approximate inference, is *variational approximation*. So-called *variational inference engines* have emerged in recent years for conducting inference in DAG models. The most prominent is VIBES (Variational Inference for BayESian networks), authored by Bishop *et al.* (2003). Several others are described in Murphy (2007), including a new successor to VIBES named Infer.NET (Minka *et al.* 2008) (current web-site: research.microsoft.com/infernet). An illustration of VIBES is given later in this section. Before that I will provide a brief description of variational approximation.

Variational approximation is an alternative to MCMC that is gathering steam as a means of making inference in complex models when the latter becomes untenable. Most contemporary literature on variational approximation for graphical models is in Computer Science rather than Statistics. Review articles that summarize contemporary variational inference are Jordan *et al.* (1999), Jordan (2004), Titterington (2004) and Bishop (2006).

The essence of variational approximation is the use of variational forms for non-linear functions. An example is

$$\log(x) = \min_{\xi>0}\{\xi x - \log(\xi) - 1\}, \quad \text{for all } x > 0.$$

The fact that $\xi x - \log(\xi) - 1$ is linear in $x$ for every value of the *variational parameter* $\xi > 0$ allows for simplifications of expressions involving the logarithmic function. The value of $\xi$ can then be chosen to make the approximation as accurate as possible.

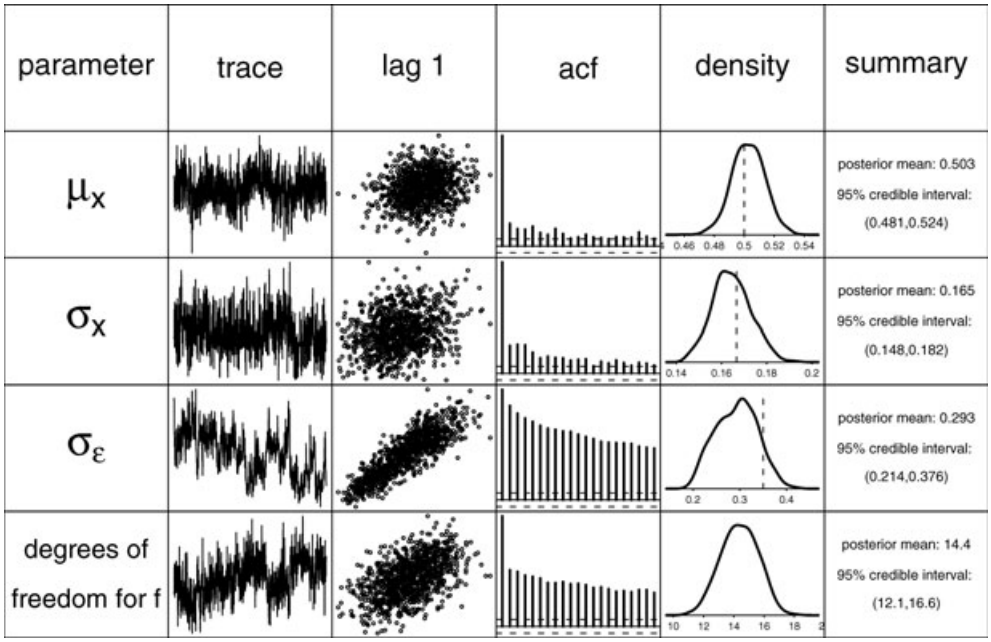


Figure 16. Summary of MCMC-based inference for parameters in the nonparametric regression measurement error model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates of posterior density and basic numerical summary. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

An instructive example of variational inference arises in simple Bayesian logistic regression:

$$[\mathbf{y} \mid \beta_0, \beta_1] = \prod_{i=1}^{n} \frac{e^{y_i(\beta_0+\beta_1 x_i)}}{1+e^{\beta_0+\beta_1 x_i}}, \quad (\beta_0, \beta_1) \sim N\big(\mathbf{0}, \sigma_\beta^2 \mathbf{I}\big), \quad 1 \leqslant i \leqslant n \tag{20}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$, $y_i \in \{0, 1\}$ is the $i$th realization of a binary response variable and $x_i$ is the corresponding predictor. This is a special case of an example given in Jaakkola & Jordan (2000). Inference about the slope parameter requires

$$[\beta_1 \mid \mathbf{y}] \propto e^{-\beta_1^2/(2\sigma_\beta^2)} \int_{-\infty}^{\infty} \exp\left( \sum_{i=1}^{n} [\beta_0 y_i - \log\{1+\exp(\beta_0+\beta_1 x_i)\}] - \beta_0^2 \big/ \big(2\sigma_\beta^2\big) \right) d\beta_0. \tag{21}$$

The presence of $-\log\{1+\exp(\beta_0+\beta_1 x_i)\}$ in the exponent of the integrand makes the integrals irreducible. However, we can make use of the variational form

$$-\log(1+e^x) = \max_{\xi \in \mathbb{R}} \{A(\xi)x^2 + B(\xi)x + C(\xi)\} \quad \text{for all } x \in \mathbb{R}, \tag{22}$$

where

$$A(\xi) = -\tanh(\xi/2)/(4\xi), \;\; B(\xi) = -1/2 \text{ and } C(\xi) = \xi/2 - \log(1+e^{\xi}) + \xi \tanh(\xi/2)/4$$

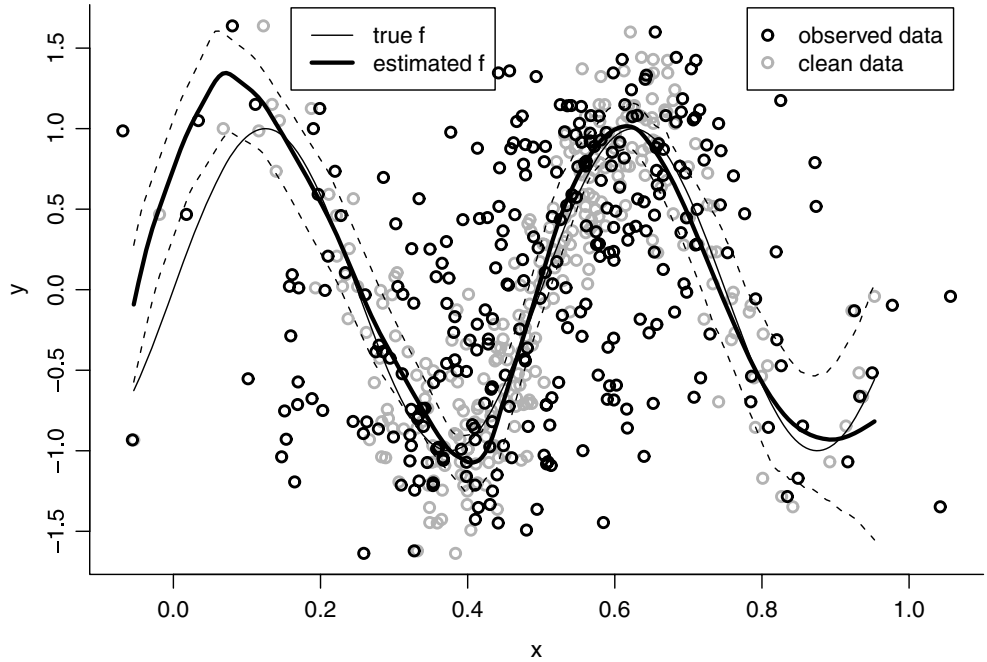


Figure 17. MCMC-based estimate of $f$ in the nonparametric regression with measurement error model. The shaded region corresponds to pointwise 95% credible sets. The black points are the observed $(w, y)$ pairs (contaminated data). The grey points are the unobserved $(x, y)$ pairs (clean data).

(Jaakkola & Jordan 2000). Figure 18 is a graphical representation of (22), in which the function $-\log(1+e^x)$ is seen to be the maximum of a family of parabolas.

In (21) one can then replace

$$-\log\{1+\exp(\beta_0+\beta_1 x_i)\} \quad \text{by} \quad A(\xi_i)x^2 + B(\xi_i)x + C(\xi_i), \quad 1 \leqslant i \leqslant n. \tag{23}$$

This entails the introduction of a vector of $n$ variational parameters $\xi = (\xi_1, \ldots, \xi_n)$. For any choice of $\xi \in \mathbb{R}^n$, one can solve the posterior density problem analytically and arrive at the following family of solutions:

$$\beta_1 \,|\, \mathbf{y}; \xi \sim N(\mu(\xi), \sigma^2(\xi)), \quad \xi \in \mathbb{R}^n,$$

where

$$\mu(\xi) = \frac{\big(2n\overline{\lambda}(\xi) + \sigma_\beta^{-2}\big)(\mathbf{x}^\top \mathbf{y} - \overline{x}/2)}{\big(2n\overline{\lambda}(\xi) + \sigma_\beta^{-2}\big)\{2(\mathbf{x}^2)^\top \lambda(\xi) + \sigma_\beta^{-2} - 4\{\lambda(\xi)^\top \mathbf{x}\}}$$

and

$$\sigma^2(\xi) = \big[2(\mathbf{x}^2)^\top \lambda(\xi) + \sigma_\beta^{-2} - 4\{\lambda(\xi)^\top \mathbf{x}\}^2 \big/ \big\{2n\overline{\lambda}(\xi) + \sigma_\beta^{-2}\big\}\big]^{-1},$$

with $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$ and $\overline{\lambda}(\xi) = \frac{1}{n}\sum_{i=1}^n \lambda(\xi_i)$. The variational parameters $\xi$ should then be chosen to make the approximation (23) as accurate as possible. This involves maximization of the lower bounds on the left-hand side of

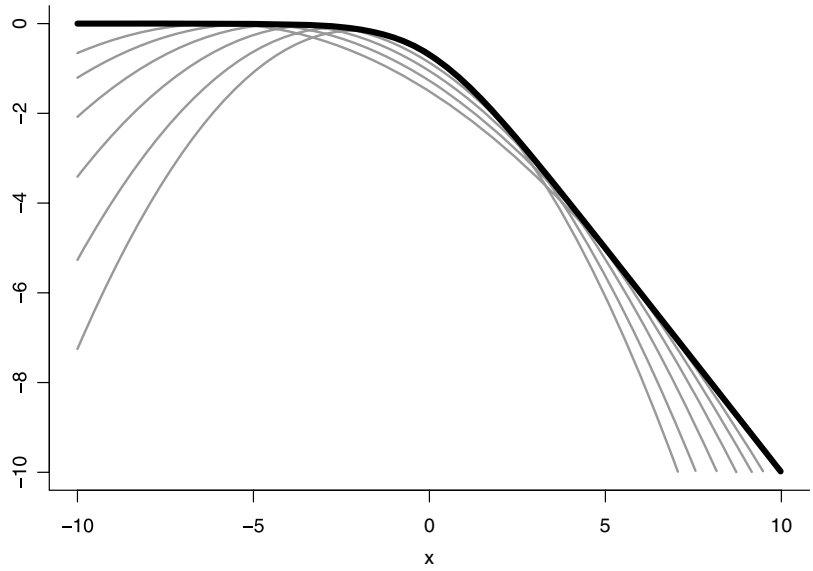


Figure 18. Variational representation of the function $-\log(1+e^x)$, as the maximum of a family of parabolas.

$$A(\xi_i)x^2 + B(\xi_i)x + C(\xi_i) \leqslant -\log\{1+\exp(\beta_0+\beta_1 x_i)\}, \quad 1 \leqslant i \leqslant n.$$

An expectation maximization-type iterative scheme can be devised for carrying this out (Jaakkola & Jordan 2000). Let $\beta_0, \beta_1 \,|\, \mathbf{y}\,; \xi \sim N(\mu(\xi), \Sigma(\xi))$ be the variational approximation to $[\beta_0, \beta_1 \,|\, \mathbf{y}]$ based on $\xi$. Then, with $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{X}$ equal to the $n \times 2$ matrix whose $i$th row is $(1\ x_i)$, the following algorithm usually leads to rapid convergence to the optimum:

Cycle:

1. $\Sigma(\xi) \leftarrow [\sigma_\beta^{-2}\,\mathbf{I} + 2\mathbf{X}^\top \operatorname{diag}\{\lambda(\xi)\}\mathbf{X}]^{-1}$
2. $\mu(\xi) \leftarrow \Sigma(\xi)\mathbf{X}^\top(\mathbf{y} - \frac{1}{2}\mathbf{1})$
3. $\xi \leftarrow \sqrt{\operatorname{diagonal}[\mathbf{X}\{\Sigma(\xi) + \mu(\xi)\mu(\xi)^\top\}\mathbf{X}^\top]}$.

I compared the posterior distribution approximations for $\beta_0$ and $\beta_1$ obtained by means of this variational approach with data on 223 birth-weight measurements (grammes) and the occurrence of *bronchopulmonary dysplasia* (source: Pagano & Gauvreau 1993). Throughout this example I work with the standardized version of the birth-weights rather than with the original birth-weight values, and with $\sigma_\beta^2 = 10^8$. Figure 19 shows the variational approximations to $[\beta_0 \,|\, \mathbf{y}]$, $[\beta_1 \,|\, \mathbf{y}]$ and $[\beta_0, \beta_1 \,|\, \mathbf{y}]$. As a benchmark, I obtained one million realizations from the posterior densities using MCMC and BUGS, and constructed kernel density estimates using direct plug-in bandwidth selectors (available in the R packages KernSmooth, Wand & Ripley 2008, and KS, Duong 2008). We see from Figure 19 that the Jaakkola & Jordan (2000) variational approximations are reasonable, but not extremely accurate.

Although this example, based on (22), provides an illustration of variational approximation, it should be pointed out that many other methods exist. A common general approach to variational approximation involves the theory of Kullback–Liebler divergence; see, for example, Titterington (2004) and Bishop (2006 section 10.1). It should also be pointed out that other analytic approximations exist and can be used for approximate inference in DAG

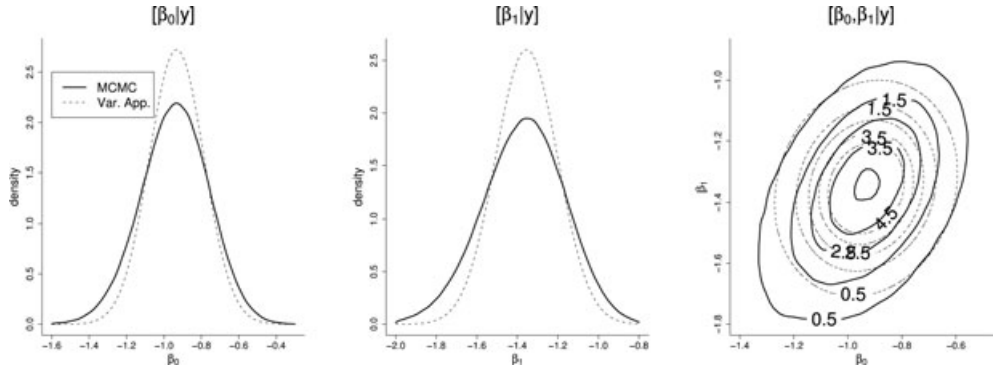


Figure 19. Assessment of the accuracy of the Jaakkola & Jordan (2000) variational approximation method. The dashed curves indicate the variational approximation to $[\beta_0 \mid \mathbf{y}]$, $[\beta_1 \mid \mathbf{y}]$ and $[\beta_0, \beta_1 \mid \mathbf{y}]$ for the Bayesian logistic regression fit to data on birth-weights and the occurrence of bronchopulmonary dysplasia. The solid curves are approximations to the posteriors using MCMC samples of size one million (obtained using BUGS).

models. A particularly simple and popular one is *Laplace approximation*, which, for example, is used by Spiegelhalter & Lauritzen (1990) in DAG models and by Breslow & Clayton (1993) in generalized linear mixed models.

Current joint research with John T. Ormerod involves a Kullback–Liebler divergence approach in which the lower bound on the likelihood is reduced to the calculation of $n$ univariate integrals, which are calculated numerically using adaptive Gauss–Hermite quadrature. The results for model (20) applied to the bronchopulmonary dysplasia data are shown in Figure 20. The accuracy is seen to be very good in this case, and considerably better than that of Jaakkola & Jordan (2000)

I also tested the use of VIBES for Bayesian semiparametric regression by getting it to fit (8) to the spinal bone mineral density data. Figure 21 is a screen-shot of the specified model in VIBES, obtained using its graph-drawing interface.

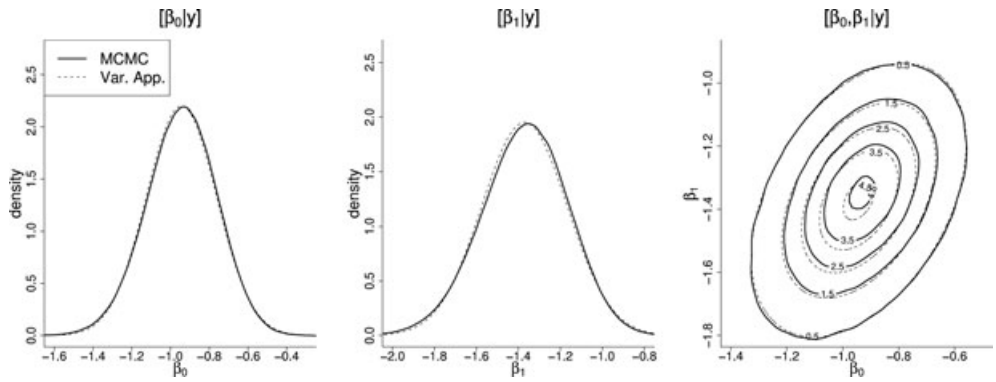


Figure 20. Assessment of the accuracy of the variational approximation method arising from current joint research with John T. Ormerod. The dashed curves indicate the variational approximation to $[\beta_0 \mid \mathbf{y}]$, $[\beta_1 \mid \mathbf{y}]$ and $[\beta_0, \beta_1 \mid \mathbf{y}]$ for the Bayesian logistic regression fit to data on birth-weights and the occurrence of bronchopulmonary dysplasia. The solid curves are approximations to the posteriors using MCMC samples of size one million (obtained using BUGS).

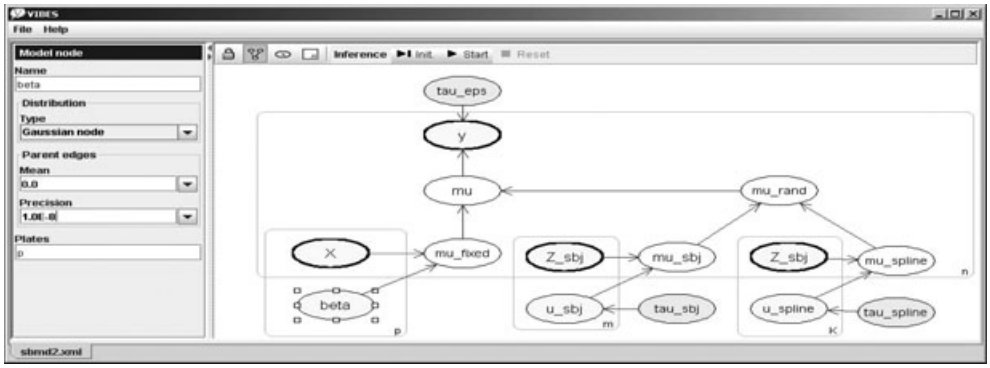


Figure 21. Screen-shot of the directed acyclic graph drawn in VIBES for specifying the semiparametric regression model applied to the spinal bone mineral density data.

Approximate posterior densities for four of the model parameters are shown in Figure 22. The regression coefficients, corresponding to the indicators for Black and Hispanic, have narrower posterior densities compared with those obtained using MCMC via BUGS. The posterior densities for the standard deviation parameters, $\sigma_{\text{spl}}$ and $\sigma_{\varepsilon}$, are quite close to the

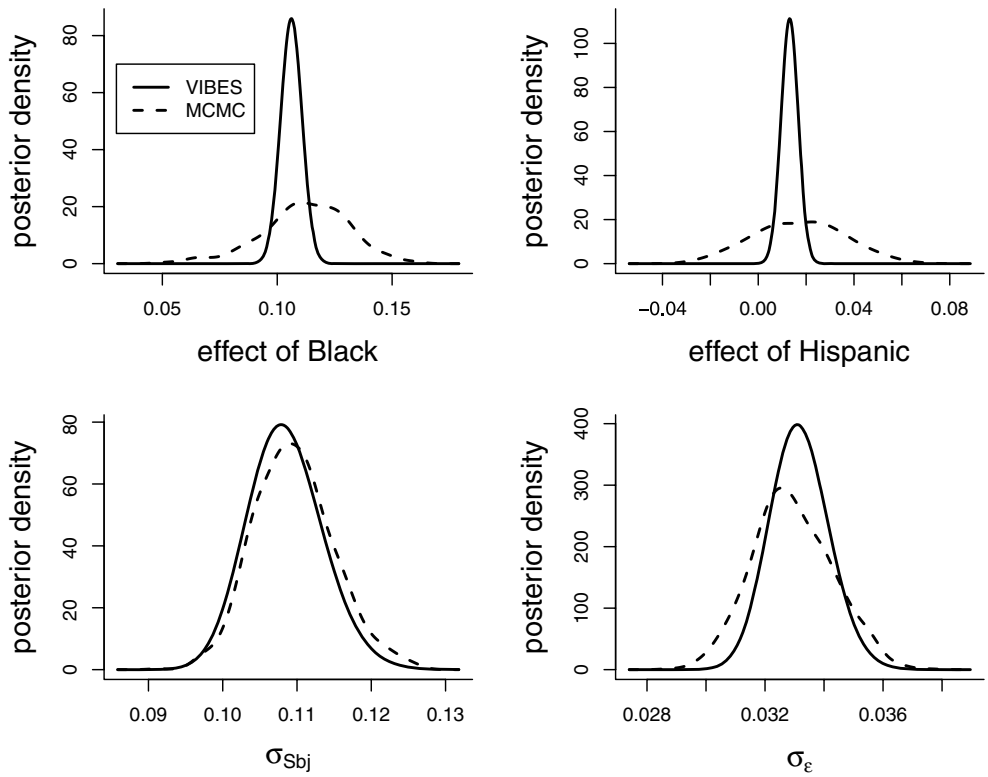


Figure 22. Approximate posterior densities for a selection of parameters in the VIBES fit of an additive mixed model to the spinal bone mineral density data. The MCMC-based approximations are shown for comparison.

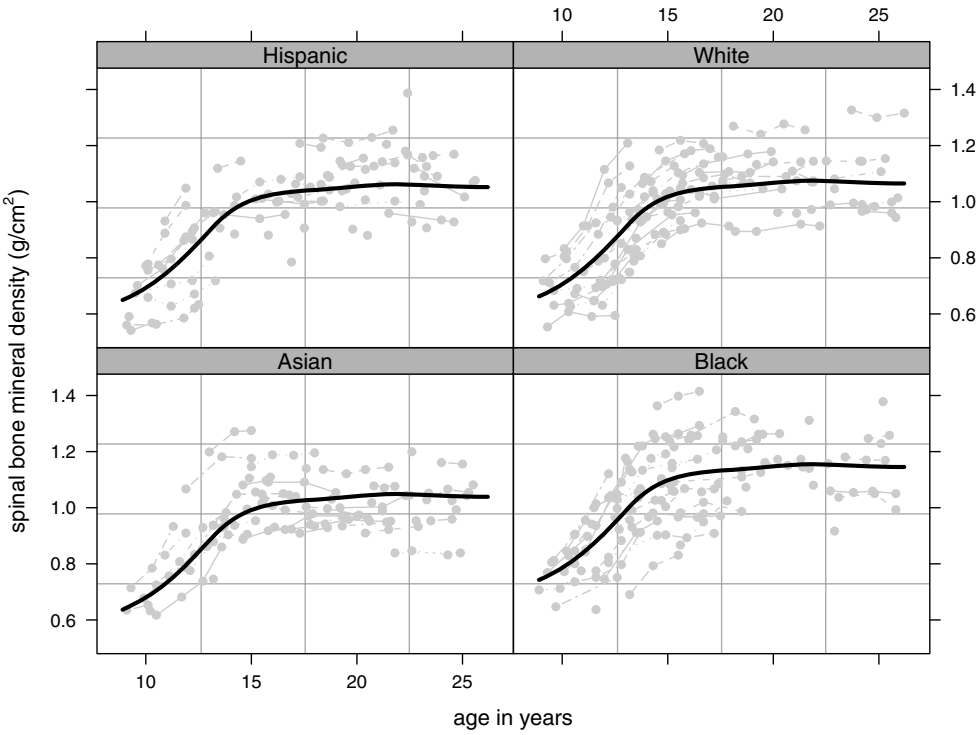


Figure 23. Bayes estimates for the age effect arising from a VIBES fit of an additive mixed model to the spinal bone mineral density data.

BUGS answers. The fitted age effects, shown in Figure 23, are also similar to those obtained via BUGS. However, the credible interval bars cannot be produced from the VIBES output.

Variational inference engines, such as VIBES, constitute a young and emerging field. They have the potential to yield satisfactory solutions to complex graphical model inferential problems much more quickly than what is currently being achieved via MCMC. There is also the question of the statistical properties of variational approximations to quantities such as maximum likelihood estimators and posterior densities. Jordan (2004) states that 'variational inference is still in its infancy' and cites Tatikonda & Jordan (2002) for early work on asymptotics for variational approximation. Several other relevant references are listed in section 3.3 of Jordan (2004). In the Statistics literature, pioneering work on variational approximation theory has been undertaken by D.M. Titterington and co-authors. Examples of published work to date include Hall, Humphreys & Titterington (2002) and Wang & Titterington (2004, 2006)

## 7. Example: relative cancer mapping with missingness

We applied the new Ormerod & Wand variational approximation methodology, mentioned in the previous section, to some real data for which semiparametric regression in the face of missingness is appropriate. The data, corresponding to a female cancer study in Cape Cod, Massachusetts, USA, are described in French & Wand (2004). Of primary interest for these data is *relative cancer mapping*, whereby the geographical variation of a certain cancer

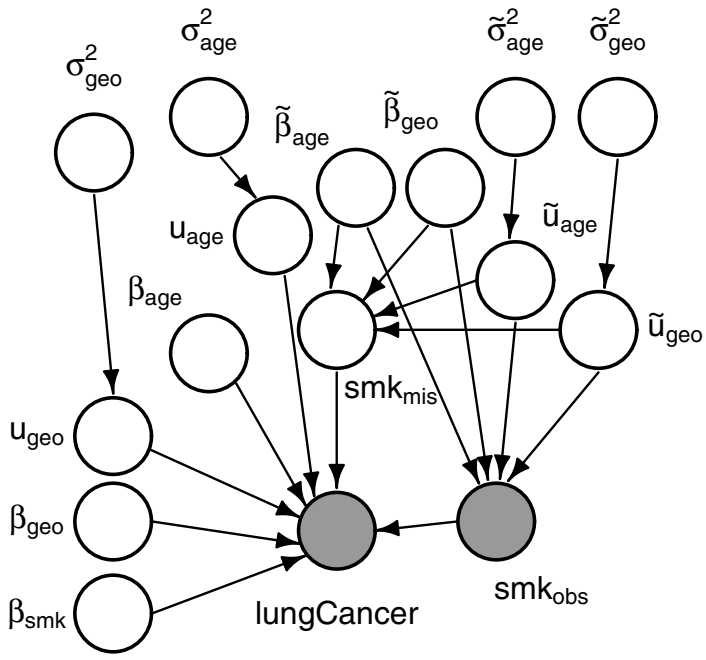


Figure 24. Graphical representation of the semiparametric regression missing data model for the relative cancer mapping example.

type, relative to other cancers, is assessed. In this illustration, the cancer type of interest is lung cancer. Define

$$\texttt{lungCancer}_i = \begin{cases} 1 & \text{female } i \text{ has lung cancer} \\ 0 & \text{female } i \text{ has other type of cancer} \end{cases}$$

for $1 \leqslant i \leqslant n$, where $n = 2540$ is the number of females in the study. For each female in the study we observe the longitude ($\texttt{lon}_i$) and latitude ($\texttt{lat}_i$) values of her residence, her age in years ($\texttt{age}_i$) and whether or not she has ever smoked ($\texttt{smoked}_i$). An exception is the 15.4% of females for which smoking information is missing.

To account for age and smoking, as well as for missingness in smoking, we entertained the hierarchical Bayesian additive models

$$\begin{aligned}
&\text{lungCancer}_i \mid \texttt{smoked}_i, \beta_0, \beta_{\text{smk}}, \beta_{\text{age}}, \beta_{\text{geo}}, \mathbf{u}_{\text{age}}, \mathbf{u}_{\text{geo}}, \sigma^2_{\text{age}}, \sigma^2_{\text{geo}} \\
&\quad \overset{\text{ind.}}{\sim} \text{Bernoulli}\Big[\text{logit}^{-1}\big\{\beta_0 + \beta_{\text{smk}}\texttt{smoked}_i + f\big(\texttt{age}_i; \beta_{\text{age}}, \sigma^2_{\text{age}}\big) \\
&\qquad + g\big(\texttt{lon}_i, \texttt{lat}_i; \beta_{\text{geo}}, \sigma^2_{\text{geo}}\big)\big\}\Big] \\
&\texttt{smoked}_i \mid \widetilde{\beta}_0, \widetilde{\beta}_{\text{age}}, \widetilde{\beta}_{\text{geo}}, \widetilde{\mathbf{u}}_{\text{age}}, \widetilde{\mathbf{u}}_{\text{geo}}, \widetilde{\sigma}^2_{\text{age}}, \widetilde{\sigma}^2_{\text{geo}} \\
&\quad \overset{\text{ind.}}{\sim} \text{Bernoulli}\Big[\text{logit}^{-1}\big\{\widetilde{\beta}_0 + \widetilde{f}\big(\texttt{age}_i; \widetilde{\beta}_{\text{age}}, \widetilde{\sigma}^2_{\text{age}}\big) + \widetilde{g}\big(\texttt{lon}_i, \texttt{lat}_i; \widetilde{\beta}_{\text{geo}}, \widetilde{\sigma}^2_{\text{geo}}\big)\big\}\Big].
\end{aligned}$$

The univariate functions $f$ and $\widetilde{f}$ are handled analogously to that for age in the spinal bone mineral density example of Section 4. The bivariate functions $g$ and $\widetilde{g}$ use penalized thin plate

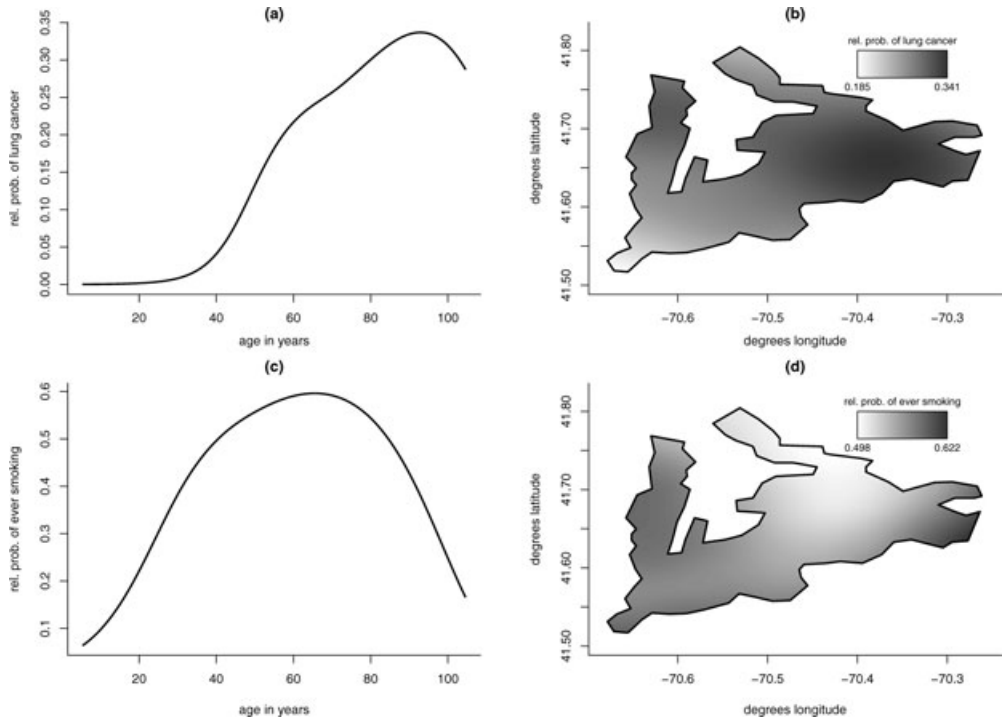


Figure 25. Ormerod & Wand variational approximation fit to the female lung cancer data. The displayed fit for each predictor corresponds to slices of the fitted model with the other predictors set to their medians. Panel (a): estimate of the effect of age on the relative probability of lung cancer. Panel (b): estimate of the effect of geography on the relative probability of lung cancer. Panel (c): estimate of the effect of age on the probability of ever being a smoker. Panel (d): estimate of the effect of geography on the probability of ever being a smoker.

splines as described in chapter 13 of Ruppert *et al.* (2003). Figure 24 provides a graphical description of this model. Our variational approximations for the missing smoking data node are similar in nature to those described in section 5.2 of Jaakkola & Jordan (2000).

The functional components of our variational approximation fit to the above model are shown in Figure 25. The upper panels of Figure 25 show the effects of age and geographical location on lung cancer occurrence (relative to other cancer types). The age curve is monotonic, as expected. The geographical fit suggests a 'hot spot' around $-70.4^\circ$ longitude and $41.65^\circ$ latitude. The lower panels are the fitted effects of age and geography on smoking status. Some geographical variability in smoking status is apparent. Furthermore, there is an interesting decline in the age curve after about 75 years. However, there is also a high degree of variability (not shown) in these function estimates for high ages.

## 8. Potential for new semiparametric regression applications

A final advantage of the graphical models viewpoint of semiparametric regression is that it brings the latter field closer to other areas of research that rely heavily on graphical model theory and methodology. Examples include social networks (e.g. Wasserman & Faust 1994),

causal inference (e.g. Cox & Wermuth 2001; van der Laan & Robins 2003), hidden Markov models (e.g. Cappé, Moulines & Ryden 2005) and phylogenetic trees (e.g. Jordan 2004).

Synergistic development of this type has been recently witnessed as a result of mixed model representations of semiparametric regression. Semiparametric regression methodology is now very much a part of longitudinal data analysis (e.g. Fitzmaurice *et al.* 2008), spatial statistics (e.g. Hennerfeind, Brezger & Fahrmeir 2006; Crainiceanu *et al.* 2008) and analysis of complex sample surveys (e.g. Breidt & Opsomer 2008). There is great potential for similar outcomes in the graphical model realm.

## 9. Concluding remarks

I have explained why I believe graphical models to be a useful structure for semiparametric regression analysis. Particular attention has been paid to non-standard situations in which there is more to gain from the graphical models viewpoint. As theory, methodology and software for graphical models continue to be developed, I envisage sophisticated semiparametric regression analyses becoming more routine and streamlined by taking advantage of graphical model representations.

## *References*

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**, 4702–4712.

BERRY, S.A., CARROLL, R.J. & RUPPERT, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *J. Amer. Statist. Assoc.* **97**, 160–169.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **36**, 192–236.

BISHOP, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

BISHOP, C.M., SPIEGELHALTER, D.J. & WINN, J. (2003). VIBES: A variational inference engine for Bayesian networks. In *Advances in Neural Information Processing Systems*, eds S. Becker, S. Thrun and K. Obermayer, pp. 793–800, Cambridge, MA: MIT Press.

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. *Aust. N. Z. J. Statist.* **49**, 287–301.

BREIDT, F.J. & OPSOMER, J.D. (2008). Nonparametric and semiparametric estimation in complex surveys. In *Sample Surveys: Theory, Methods and Inference, Handbook of Statistics*, eds C.R. Rao and D. Pfeffermann, Amsterdam: North Holland, in press.

BRESLOW, N.E. & CLAYTON, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9–25.

BÜHLMANN, P. & HOTHORN, T. (2007). Boosting algorithms: regularization, prediction and model fitting. *Statist. Sci.* **22**, 477–505.

BUJA, A., HASTIE, T. & TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17**, 453–510.

CAPPÉ, O., MOULINES, E. & RYDEN, T. (2005). *Inference in Hidden Markov Models*. New York: Springer.

CARROLL, R.J., RUPPERT, D., STEFANSKI, L.A. & CRAINICEANU, C.M. (2006). *Measurement Error in Nonlinear Models*, 2nd edn. Boca Raton, FL: Chapman & Hall/CRC.

CASELLA, G. & ROBERT, C.P. (2004). Introduction to the special issue: Bayes then and now. *Statist. Sci.* **19**, 1–2.

CASTILLO, E., GUTIÉRREZ, J.M. & HADI, A.S. (1997). *Expert Systems and Probabilistic Network Models*. New York: Springer.

CHEN, Q.X. & IBRAHIM, J.G. (2006). Semiparametric models for missing covariate and response data in regression models. *Biometrics* **62**, 177–184.

COWELL, R.G., DAWID, A.P., LAURITZEN, S.L. & SPIEGELHALTER, D.J. (1999). *Probabilistic Networks and Expert Systems*. New York: Springer.

COX, D.R. & WERMUTH, N. (2001). Causal inference and statistical fallacies. In *International Encyclopedia of the Social and Behavioral Sciences*, eds P.B. Bates & N.J. Smelser, pp. 1554–1661. Amsterdam: Elsevier.

CRAINICEANU, C.M., DIGGLE, P.J. & ROWLINGSON, B. (2008). Bivariate binomial spatial modelling Loa loa prevalence in tropical Africa (with discussion). *J. Amer. Statist. Assoc.* **103**, 21–43.

DUONG, T. (2008). ks: Kernel smoothing. R package version 1.5.5. [cited 28 January 2008] Available from URL: http://www.cran.r-project.org.

EUBANK, R.L. (1999). *Nonparametric Regression and Spline Smoothing*, 2nd edn. New York: Marcel Dekker.

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G. & MOLENBERGHS, G., eds (2008). *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. Boca Raton, FL: Chapman & Hall/CRC.

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. *Inform. Comput.*, **121**, 256–285.

FREUND, Y. & SCHAPIRE, R.E. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, ed. L. Saitta, pp. 148–156. San Francisco: Morgan Kauffman.

GEIGER, D., VERMA, T.S. & PEARL, J. (1990). Identifying independence in Bayesian networks. *Networks* **20**, 507–534.

GEMAN, S. & GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Pattern Anal. Mach. Intell.* **6**, 721–741.

GERACI, M. & BOTTAI, M. (2006). Use of auxiliary data in semi-parametric spatial regression with nonignorable missing responses. *Statist. Modelling* **6**, 321–336.

GIANOLA, D., FERNANDO, R.L. & STELLA, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* **173**, 1761–1776.

HALL, P., HUMPHREYS, K. & TITTERINGTON, D.M. (2002). On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **64**, 549–564.

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. & ZHU, J. (2006). Comment on paper by Moguerza & Muñoz. *Statist. Sci.* **21**, 352–357.

HENNERFEIND, A., BREZGER, A. & FAHRMEIR, L. (2006). Geoadditive survival models. *J. Amer. Statist. Assoc.* **101**, 1065–1075.

JAAKKOLA, T.S. & JORDAN, M.I. (2000). Bayesian parameter estimation via variational methods. *Statist. Comput.* **10**, 25–37.

JANK, W. & SHMUELI, G. (2007). Modelling concurrency of events in on-line auctions via spatiotemporal semiparametric models. *J. R. Statist. Soc. Ser. C* **56**, 1–27.

JENSEN, F.V. (1996). *An Introduction to Bayesian Networks*. London: UCL Press.

JORDAN, M.I. (1999). *Learning in Graphical Models*. Cambridge, MA: MIT Press.

JORDAN, M.I. (2004). Graphical models. *Statist. Sci.* **19**, 140–155.

JORDAN, M.I., GHAHRAMANI, Z., JAAKKOLA, T.S. & SAUL, L.K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37**, 183–233.

KAMMANN, E.E. & WAND, M.P. (2003). Geoadditive models. *J. R. Statist. Soc. Ser. C* **52**, 1–18.

LAURITZEN, S.L. (1996). *Graphical Models*. Oxford: Clarendon Press.

LIANG, F.M., TRUONG, Y.K. & WONG, W.H. (2001). Automatic Bayesian model averaging for linear regression and applications in Bayesian curve fitting. *Statist. Sinica* **4**, 1005–1029.

LUNN, D.J., THOMAS, A., BEST, N. & SPIEGELHALTER, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statist. Comput.* **10**, 325–337.

MINKA, T., WINN, J., GUIVER, G. & KANNAN, A. (2008). *Infer.Net*, Microsoft Research Cambridge, Cambridge, UK.

MURPHY, K. (2007). Software for graphical models: a review. *Int. Soc. Bayesian Anal. Bull.* **14**, 13–15.

NOTT, D. (2006). Semiparametric estimation of mean and variance functions for non-Gaussian data. *Comput. Statist.* **21**, 603–620.

PADOAN, S.A. & WAND, M.P. (2008). Mixed-model based additive models for sample extremes. *Statist. Probab. Lett.* **78**, 2850–2858.

PAGANO, M. & GAUVREAU, K. (1993). *Principles of Biostatistics*. Florence, KY: Duxbury.

PEARCE, N.D. & WAND, M.P. (2006). Penalized splines and reproducing kernel methods. *Amer. Statist.* **60**, 233–240.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

PINHEIRO, J., BATES, D., DEBROY, S., SARKAR, D. & the R Core team. (2008). nlme: linear and nonlinear mixed effects models. R package version 3.1–89. [cited 28 January 2008] Available from URL: http://www.cran.r-project.org.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.

ROBERT, C.P. & CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd edn. New York: Springer-Verlag.

RUPPERT, D., WAND, M. P. & CARROLL, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

RUPPERT, D., WAND, M.P. & CARROLL, R.J. (2009). Semiparametric regression during 2003–2007. *J. Amer. Statist. Assoc.* to appear.

SAS Institute, Inc. (2008). Cary, NC.

SCHAPIRE, R.E. (1990). The strength of weak learnability. *Machine Learning* **5**, 197–227.

SCHÖLKOPF, B. & SMOLA, A.J. (2002). *Learning with Kernels*. Cambridge, MA: MIT Press.

SPIEGELHALTER, D.J. & LAURITZEN, S.L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* **20**, 579–605.

STAUDENMAYER, J., LAKE, E.E. & WAND, M.P. (2009). Robustness for general design mixed models using the $t$-distribution. *Statist. Modelling*, in press.

TATIKONDA, S. & JORDAN, M.I. (2002). Loopy belief propagation and Gibbs measures. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, eds A. Darwiche & N. Friedman, pp. 493–500. San Mateo, CA: Morgan Kaufmann.

THURSTON, S., WAND, M.P. & WEINCKE, J. (2000). Negative binomial additive models. *Biometrics* **56**, 139–144.

TITTERINGTON, D.M. (2004). Bayesian methods for neural networks and related models. *Statist. Sci.* **19**, 128–139.

TUTZ, G. & BINDER, H. (2006). Generalized additive modelling with implicit variable selection by likelihood based boosting. *Biometrics* **62**, 961–971.

VAN DER LAAN, M.J. & ROBINS, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **40**, 364–372.

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

WAHBA, G. (2006). Comment on paper by Moguerza & Muñoz. *Statist. Sci.* **21**, 347–351.

WAND, M.P. (2003). Smoothing and mixed models. *Comput. Statist.* **18**, 223–249.

WAND, M.P. & ORMEROD, J.T. (2008). On O'Sullivan penalised splines and semiparametric regression. *Aust. N. Z. J. Statist.*, **50**, 179–198.

WAND, M.P. & RIPLEY, B.D. (2008). KernSmooth: functions for kernel smoothing for Wand & Jones 1995. R package version 2.22-22.

WANG, B. & TITTERINGTON, D.M. (2004). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Anal.* **1**, 625–650.

WANG, B. & TITTERINGTON, D.M. (2006). Lack of consistency of mean field and variational Bayes approximations for state space models. *Neur. Process. Lett.* **20**, 151–170.

WASSERMAN, L. (2004). *All of Statistics*. New York: Springer.

WASSERMAN, S. & FAUST, K. (1994). *Social Networks Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

WELHAM, S.J., CULLIS, B.R., KENWARD, M.G. & THOMPSON, R. (2007). A comparison of mixed model splines for curve fitting. *Aust. N. Z. J. Statist.* **49**, 1–23.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.

WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5**, 161–215.

WOOD, S.N. (2003). Thin plate regression splines. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **65**, 95–114.

WOOD, S.N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62**, 1025–1036.

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

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.

ZHAO, Y., STAUDENMAYER, J., COULL, B.A. & WAND, M.P. (2006). General design Bayesian generalized linear mixed models. *Statist. Sci.* **21**, 35–51.