1226507

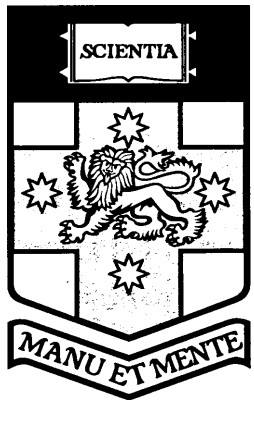
Date of completion : 13/08/2008.

ON SEMIPARAMETRIC REGRESSION AND DATA MINING

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
John T. Ormerod

Supervisor: Professor Matthew P. Wand



School of Mathematics and Statistics,
The University of New South Wales.

August 2008

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed
Date ...

"Deep within us all, emergent when the noise of other appetites is stilled, there is a drive to know, to understand, to see why, to discover the reason, to find the cause, to explain. Just what is wanted has many names. In what precisely it consists is a matter of dispute. But the fact of inquiry is beyond all doubt. It can absorb a man. It can keep him up for hours, day after day, year after year, in the narrow prison of his study or his laboratory. It can send him off on dangerous voyages of exploration. It can withdraw him from other interests, other pursuits, other pleasures, other achievements. It can fill his walking thoughts, hide him from the world of ordinary affairs, invade the very fabric of his dreams. It can demand endless sacrifices that are made without regret though there is only the hope, never a certain promise, of success."

Lonergan (1958)

Acknowledgements

This thesis has been the hardest undertaking I have ever made and would not have been possible without many many important people in my life.

The journey that has led to this thesis started long before enrolling for my doctorate. I thank my family, in particular my parents Neil and Thea Ormerod for raising me and giving me all the love and support I needed throughout these three years.

I lovingly acknowledge my fiancée Jane for her patience, particularly over the last few months of the writing of this thesis. I also thank her for keeping me sane and for otherwise bringing the balance to my life I have always needed.

A very special thanks to my supervisor Professor Matt Wand. Firstly for providing the funding for my research over the first year and secondly for his encouragement, advice, direction and timely comments.

Thank you to the people who read and edited my thesis including Matt Wand, Thea Ormerod, Nathan Pearce and Scott Scisson. This thesis has been substantially improved due your contributions.

For the staff at the School of Mathematics and Statistics at the University of New South Wales who have given help and advice, personal or academic, along the corridors of the Red Center including: Rob Womersly, Bruce Henry, Yanan Fan, Frances Kuo, Inge Koch and Peter Blennerhassett.

Finally, I would like to thank my friends, especially Nathan Pearce, Michael Roper, Robert Taggart, Patrick Costello and Ben Waterhouse. Your help and company throughout these years will be remembered.

Abstract

Semiparametric regression is playing an increasingly large role in the analysis of datasets exhibiting various complications (Ruppert, Wand & Carroll, 2003). In particular semiparametric regression a plays prominent role in the area of data mining where such complications are numerous (Hastie, Tibshirani & Friedman, 2001). In this thesis we develop fast, interpretable methods addressing many of the difficulties associated with data mining applications including: model selection, missing value analysis, outliers and heteroscedastic noise.

We focus on function estimation using penalised splines via mixed model methodology (Wahba 1990; Speed 1991; Ruppert *et al.* 2003). In dealing with the difficulties associated with data mining applications many of the models we consider deviate from typical normality assumptions. These models lead to likelihoods involving analytically intractable integrals. Thus, in keeping with the aim of speed, we seek analytic approximations to such integrals which are typically faster than numeric alternatives.

These analytic approximations not only include popular penalised quasi-likelihood (PQL) approximations (Breslow & Clayton, 1993) but variational approximations. Originating in physics, variational approximations are a relatively new class of approximations (to statistics) which are simple, fast, flexible and effective. They have recently been applied to statistical problems in machine learning where they are rapidly gaining popularity (Jordan, Ghahramani, Jaakkola & Saul 1999; Corduneanu & Bishop, 2001; Ueda & Ghahramani, 2002; Bishop & Winn, 2003; Winn & Bishop 2005).

We develop variational approximations to: generalized linear mixed models (GLMMs); Bayesian GLMMs; simple missing values models; and for outlier and heteroscedastic noise models, which are, to the best of our knowledge, *new*. These methods are quite effective and extremely fast, with fitting taking minutes if not seconds on a typical 2008 computer.

We also make a contribution to variational methods themselves. Variational approximations often underestimate the variance of posterior densities in Bayesian models (Humphreys & Titterington, 2000; Consonni & Marin, 2004; Wang & Titterington, 2005). We develop *grid-based variational posterior approximations*. These approximations combine a sequence of variational posterior approximations, can be extremely accurate and are reasonably fast.

Notation and Symbols

The following notation and symbols will be used unless otherwise stated.

0.1 Matrix Algebra

Vectors and matrices are in bold typeface. Vectors are denoted using lower case letters and matrices are denoted using upper case letters.

$\mathbb{R}^n$	The set of real vectors of dimension n.
$\mathbb{R}^{n\times m}$	The set of real matrices with n rows and m columns.
$\mathbb{R}^n_+$	The set of n dimensional positive real vectors.
T	Transpose. Vectors and matrices superscripted by T are transposed. All vectors are column vectors unless otherwise stated or transposed by T.
$(\cdot,\ldots,\cdot)$	Vertical concatenation. Lists of scalars, vectors and matrices within round brackets "( )" are concatenated **vertically**, e.g. $\mathbf{a} = (a_1,\ldots,a_n)$ is a column vector.
$[\cdot,\ldots,\cdot]$	Horizontal concatenation. Lists of scalars, vectors and matrices within square brackets "[ ]" are concatenated **horizontally**, e.g. $\mathbf{a} = [a_1,\ldots,a_n]$ is a row vector. Also $(\mathbf{a}_1,\ldots,\mathbf{a}_n) = [\mathbf{a}_1^T,\ldots,\mathbf{a}_n^T]^T$.
$\mathrm{diag}(\mathbf{a})$	Diagonal matrix. Let $\mathbf{a}\in\mathbb{R}^n$ then we denote the $n\times n$ matrix with diagonal entries a by $\mathrm{diag}(\mathbf{a}) = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_n \end{bmatrix}.$
$\mathrm{dg}(\mathbf{A})$	Diagonal vector. Let $\mathbf{A}\in\mathbb{R}^{n\times n}$ then we denote the vector of length n with entries equal to the diagonal elements of $\mathbf{A}$ by $\mathrm{dg}\left(\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}\right) = (A_{11}, A_{22},\ldots,A_{nn}).$
$\mathbf{1}$	An appropriately-sized vector or matrix of ones.
$\mathbf{0}$	An appropriately-sized vector or matrix of zeros.
$\mathbf{I}$	An appropriately-sized Identity matrix. A zero matrix with ones along the diagonal, i.e. $\mathbf{I} = \mathrm{diag}(\mathbf{1})$.

$\mathbf{e}_i$	An appropriately-sized vector of zeros, except the ith value which is 1.
$\mathbf{E}_{ij}$	An appropriately-sized matrix of zeros, except the (i,j)th entry which is 1.
$\mathrm{tr}(\mathbf{A})$	The trace of the matrix $\mathbf{A}$.
$\lvert\mathbf{A}\rvert$	The determinant of the matrix $\mathbf{A}$.
$\mathbf{A} \otimes \mathbf{B}$	Kronecker product. If $\mathbf{A} \in \mathbb{R}^{n\times m}$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$ then $\mathbf{A} \otimes \mathbf{B}$ is the $(np) \times (mq)$ matrix defined by $\mathbf{A} \otimes \mathbf{B} = [a_{ij}\mathbf{B}]_{1\leq i\leq n, 1\leq j\leq m}$.
$\mathbf{a} \odot \mathbf{b}$	Element-wise multiplication of vectors. If $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ then $\mathbf{a} \odot \mathbf{b} = (a_1 b_1, \ldots, a_n b_n)$.
$\mathbf{a}/\mathbf{b}$	Element-wise division of vectors. If $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ then $\mathbf{a}/\mathbf{b} = (a_1/b_1, \ldots, a_n/b_n)$.
$f(\mathbf{x})$	Univariate function of a vector. Let $f : \mathbb{R} \to \mathbb{R}$ be a function and $\mathbf{x} \in \mathbb{R}^n$ be a vector then we denote $f(\mathbf{x}) = (f(x_1), \ldots, f(x_n))$. For example, $\log(\mathbf{x}) = (\log(x_1), \ldots, \log(x_n))$.
$\lVert\mathbf{a}\rVert$	Vector 2-norm. If $\mathbf{a} \in \mathbb{R}^n$ then $\lVert\mathbf{a}\rVert = \left(\sum_{i=1}^n a_i^2\right)^{\frac{1}{2}}$.
$\mathbf{a} < \mathbf{b}$	Vector inequalities. If $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ then the inequality $\mathbf{a} > \mathbf{b}$ (similarly for $\mathbf{a} \geq \mathbf{b}$, $\mathbf{a} < \mathbf{b}$ and $\mathbf{a} \leq \mathbf{b}$) denotes $a_i > b_i$ for all $1 \leq i \leq n$.

0.2 Calculus

$\int f(\mathbf{x})d\mathbf{x}$	Integration. If integrals appear without integrals, i.e $\int f(\mathbf{x})d\mathbf{x}$ then the domain of integration is over all appropriate values of $\mathbf{x}$. For example, if $\mathbf{x} \in \mathbb{R}^n$ then $\int_{\mathbb{R}^n} f(\mathbf{x})d\mathbf{x} = \int f(\mathbf{x})d\mathbf{x}$ and if $\mathbf{x} \geq \mathbf{0}$ then $\int_{\mathbb{R}^n_+} f(\mathbf{x})d\mathbf{x} = \int f(\mathbf{x})d\mathbf{x}$.
$\mathsf{D}_\mathbf{x} f(\widehat{\mathbf{x}})$	Derivative vector. If $\mathbf{x} = (x_1, \ldots, x_n)$ then $\mathsf{D}_\mathbf{x} f(\widehat{\mathbf{x}}) = \left(\left.\frac{\partial f}{\partial x_i}\right\rvert_{\mathbf{x}=\widehat{\mathbf{x}}} \right)_{1\leq i\leq n}$
$\mathsf{H}_\mathbf{x} f(\widehat{\mathbf{x}})$	Hessian matrix. If $\mathbf{x} = (x_1, \ldots, x_n)$ then $\mathsf{H}_\mathbf{x} f(\widehat{\mathbf{x}}) = \mathsf{D}_\mathbf{x}\{\mathsf{D}_\mathbf{x} f(\widehat{\mathbf{x}})^T\} = \left[\left.\frac{\partial^2 f}{\partial x_i \partial x_j}\right\rvert_{\mathbf{x}=\widehat{\mathbf{x}}} \right]_{1\leq i,j\leq n}$

0.3 Probability

Let $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$ be a random vectors, i.e. vectors whose components are random variables.

$[\mathbf{x}]$	The probability density function of $\mathbf{x}$.
$[\mathbf{x}, \mathbf{y}]$	The joint density function of $\mathbf{x}$ and $\mathbf{y}$.
$[\mathbf{y}\lvert\mathbf{x}]$	The conditional distribution of $\mathbf{y}$ *given* $\mathbf{x}$, i.e. $[\mathbf{y}\lvert\mathbf{x}] = \frac{[\mathbf{x},\mathbf{y}]}{[\mathbf{x}]}$ for all $\mathbf{x}$ such that $[\mathbf{x}] > 0$.
$\ell(\boldsymbol{\theta})$	Log-likelihood of $\boldsymbol{\theta}$.
$\mathbb{P}$	Probability
$\mathbb{E}(\mathbf{x})$	Expectation of a random vector. The *mean* or *expected value* of $\mathbf{x}$, denoted $\mathbb{E}(\mathbf{x})$ contains the expected values of the components of $\mathbf{x}$, i.e. $\mathbb{E}(\mathbf{x}) = (\mathbb{E}(x_1), \ldots, \mathbb{E}(x_n))$. Furthermore, for some function f, let $\mathbb{E}_\mathbf{x}(f(\mathbf{x}, \mathbf{y}))$ denote the expecation of f with respect to $\mathbf{x}$.

$\text{Cov}(\mathbf{x})$ Covariance of a random vector. The covariance matrix is the $n \times n$ matrix, denoted $\text{Cov}(\mathbf{x})$, whose (i, j)th entry is the covariance between x_i and x_j and may be calculated via $\text{Cov}(\mathbf{x}) = \mathbb{E}\{[\mathbf{x} - \mathbb{E}(\mathbf{x})][\mathbf{x} - \mathbb{E}(\mathbf{x})]^T\}$.

$\text{Mgf}_{\mathbf{y}}(\mathbf{t})$ Moment Generating Function. $\mathbb{E}_{\mathbf{y}}\{\exp(\mathbf{y}^T\mathbf{t})\}$

$\mathcal{H}_{\mathbf{y}}$ (Shannon's) Entropy. Negative expectation of the log probability density function, i.e. $\mathcal{H}_{\mathbf{y}} = -\mathbb{E}_{\mathbf{y}}\{\log[\mathbf{y}]\}$.

$\mathcal{I}_{\boldsymbol{\theta}}$ Information matrix. $\mathcal{I}_{\boldsymbol{\theta}} = -\mathbb{E}\{\mathsf{H}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})\}$.

0.4 Miscellaneous

$\mathbb{I}_{\{x\}}$ Indicator variable (with condition x). Takes the value 1 if the condition x is true and 0 otherwise.

x_+^p Truncated power. If $x \in \mathbb{R}$ then $x_+^p = x^p \mathbb{I}_{\{x>0\}}$, i.e. $x_+^p = \begin{cases} 0 & \text{if } x \leq 0 \\ x^p & \text{if } x > 0. \end{cases}$

$\text{sign}(x)$ Sign function. If $x \in \mathbb{R}$ then $\text{sign}(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } x = 0 \\ 1 & \text{if } x > 0. \end{cases}$

$o(\cdot)$ Little "oh". We write $\eta_k = o(\nu_k)$ if the sequence of ratios approaches 0, i.e. $\lim_{k\to\infty} \eta_k/\nu_k = 0$.

$O(\cdot)$ Big "oh". We write $\eta_k = O(\nu_k)$ if for sufficiently large k the exists a constant C such that $|\eta_k| \leq C|\nu_k|$.

$a := b$ Assignment. We denote the assignment of the value for a to the value b by $a := b$.

Contents

[1]Sections 2.1-4 and 2.7 correspond to: Wand, M.P. & Ormerod, J.T. (2008). On Semiparametric Regression with O'Sullivan Penalised Splines. *Australian and New Zealand Journal of Statistics*, (in press), representing joint research between M.P Wand and J.T. Ormerod. Sections 2.5 and 2.6 contain additional material representing solo research by J.T. Ormerod.

[2] This chapter corresponds to: Kauermann, G., Ormerod, J.T. & Wand, M.P. (2008), Parsimonious Classification via Generalised Linear Mixed Models. (submitted), representing joint research between G. Kauermann, J.T. Ormerod and M.P Wand.

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 Introduction

In all of human history we have never witnessed the abundance of information with which we are confronted today. More than ever before, computers are used to record almost every measurable facet of life; from medical records, business transactions and sport statistics to travel documentation, grocery bills and weather reports. Such vast and varied information boggles the human capacity to understand its own environment. Two disciplines have emerged which deal with such data, namely Statistics and Computer Science. Each of these endeavours make sense of this data in such a way that we improve our understanding of the world around us and hence the decisions that we make.

Applications of such research touches on many facets of life. These are some examples:

- Commerce including fraud detection (Phua, Lee, Gayler & Smith, 2006), credit risk (Madeira, Oliveira & Conceição, 2003), economics (Hoover & Perez, 1999), finance (Kovalerchuk & Vityaev, 2000) and marketing (Büchner & Mulvenna, 1998; Berry & Linoff, 2004).
- Biology including ecology (Chau & Muttil, 2007), genetics (Perez-Iratxeta, Bork & Andrade, 2002) and bioinformatics (Frank, Hall, Trigg, Holmes & Witten, 2004).
- Medicine including medical diagnosis (Mangasarian, Street & Wolberg, 1995) and treatment evaluation (Lee, Mangasarian & Wolberg, 2000).
- Crime (Duffett & Vernik, 1997) and security (Lovell & Chen, 2005).
- Physics including meteorology (Luengo, Cofiño, & Gutiérrez, 2004) and astronomy (Karimabadi, Sipes, White, Marinucci, Dmitriev, Chao, Driscoll & Balac, 2007).
- Image recognition including handwriting (Vapnik, 1998, 2000), image and face recognition (Heisele, Ho & Poggio, 2003; Guo, Li & Chan, 2000).

These and other applications have led to some tremendous improvements in our lives.

Statistics and Computer Science have both played a role in the understanding of our world, albeit with different focuses. Statistics has traditionally concerned itself with design and analysis of experiments and the development of probabilistic models to describe collected data. Furthermore, the data used by statisticians has been relatively small in size with a focus on developing methods for accurate parameter estimation and accurate interpretation of these models.

Motivated by the recent availability of large data sets, the discipline of Computer Science has developed a huge array of algorithms to deal with large complex datasets independently. Many of the models developed contain no apparent underlying probabilistic structure and are focused on simple efficient algorithms with little regard to modeling statistical complexities exhibited in the data. Such algorithms usually have the advantage of being fast and adaptable to the problem at hand (Breiman, 2001). On the other hand, without probabilistic structure it is difficult to provide measures of confidence in the predictions made or draw theoretical conclusions from them.

Within these two disciplines there are a number of research areas with overlapping interests including, amongst others, semiparametric regression, statistical or machine learning theory, pattern recognition, computational learning theory and data mining. Sometimes the difference between calling one analysis, say "machine learning", and another, say "data mining", depends on context, for example on how the data was collected, the type of data to analysed and purpose of the analysis. There are other times when even these distinctions are blurred. For example, interpretability plays a far more important role in data mining than it does in machine learning, while in machine learning the accuracy of the results are of utmost importance.

In a recent book Hastie, Tibshirani & Friedman (2001) list the following characteristics typical of data mining applications. They:

- are large, both in number of observations and number of variables;
- contain mixed data types, i.e. quantitative, binary and categorical variables;
- contain many missing values;
- in which quantitative variables are often long-tailed and highly skewed and contain a substantial fraction of outliers;
- in which variables are measured on very different scales;
- contain many irrelevant predictors (i.e. parsimonious models are desirable)
- and for which accurate interpretation and prediction are important.

In addition data is not usually collected via experimental means. This potentially leads to additional complications including, but not limited to, measurement error, collection bias, correlation and change points.

The type of analysis required for such data also varies. This can be roughly categorised into supervised and unsupervised learning. Supervised learning is synonymous with predictive modeling in Statistics, while unsupervised learning is concerned with organising and summarising data. Each of these has a role to play within the data mining context. When used together, they can also lead to improved results. In this thesis we will be largely concerned with supervised learning, i.e. predictive modeling.

In another recent book Ruppert, Wand & Carroll (2003) showed how many of the complications listed above may be handled using semiparametric regression methodology. Indeed semiparametric regression models feature prominently in Hastie *et al.* (2001). As the name suggests, semiparametric regression combines parametric regression and nonparametric techniques to analyse data. It encompasses a large variety of regression

techniques, but can be roughly categorised into function estimation and longitudinal analysis. Within the context of semiparametric regression, function estimation is typically performed via penalised spline methodology which includes scatterplot smoothing, kriging and geoadditive models (Ruppert *et al.*, 2003). Longitudinal analysis, on the other hand, involves modeling of correlation in grouped data and leads to simple, hierarchical, crossed and nested random effect models (Verbeke & Molenberghs, 2000; McCulloch & Searle, 2001). Function estimation methodology is quite common within the context of data mining while longitudinal analysis is mostly ignored. Semiparametric regression techniques, as we will shortly see, are reasonably fast on large datasets, can naturally handle mixed data types, are highly interpretable and make good predictions. Furthermore, it can handle complications that arise in such analysis, for example missing value, variance function and measurement error models (Ruppert *et al.*, 2003) but such matters are subjects of ongoing research.

In the last few years we have seen the gap narrowing between statistical and Computer Science based approaches to data mining. Many researchers have taken advantage of the tremendous opportunities for cross-disciplinary research. This thesis offers a semiparametric regression approach to data mining which constitutes cross-disciplinary research of this type. Semiparametric regression methodology will be developed to handle the problems of missing data, robustness, model selection and interpretation associated with data mining. This research applies the variational methodology developed by computer scientists to statistical models where parameter estimation is extremely difficult. It thus represents a further narrowing of the gap between these two fields.

1.2 Semiparametric Regression

As previously stated, both responses and covariates in data mining applications can take a variety of forms. Data types can be either numeric or categorical in nature. Numeric data types include continuous, positive continuous and count data as subtypes, and categorical data types include binary, ordinal and nominal as subtypes. It can be vital to model these forms of data for a model to be fitted, interpreted and to make inferences from effectively. Traditionally data mining has been primarily concerned with the problems of classification (where the responses take distinct values called categories) and regression (where the responses are numeric).

The data mining problem of regression is routinely handled by semiparametric regression via penalised splines. Consider the following regression problem. Suppose we have been given the paired observations (y_i, x_i), $1 \leq i \leq n$ where y is the response variable (or target using data mining terminology), and x is the predictor variable. A penalised spline model for this data

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_m x^m + \sum_{j=1}^{K} u_j (x - \kappa_j)_+^m \tag{1.1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\mathbf{u} = (u_1, \ldots, u_K)$ is chosen to minimise

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \mathbf{u}^T \boldsymbol{\Omega} \mathbf{u} \tag{1.2}$$

with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ for a fixed smoothing parameter λ, penalty matrix $\boldsymbol{\Omega}$ and power m. Here $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_K)$ are called knots, $(x - \kappa_j)_+^m = (x - \kappa_j)^m \mathbb{I}_{\{x > \kappa_j\}}$ are called truncated power splines where $\mathbb{I}$ is an indicator variable which takes the value 1 when the condition in the subscript of $\mathbb{I}$ is true and 0 otherwise, and the set $\{(x - \kappa_j)_+^m\}_{1 \leq j \leq K}$ form a truncated power basis. This basis is often a first choice due to its conceptual simplicity. For fixed $\lambda > 0$, under certain light regularity conditions, $\boldsymbol{\beta}$ and $\mathbf{u}$ are uniquely defined and the quality of fit depends on the delicate matter of choosing λ.

There are a variety of alternatives for choosing the smoothing parameter λ including: Mallow's C_p criterion (Mallows, 1973); AIC by Akaike (1974) and similar AIC-like criteria by Hurvich, Simonoff, and Tsai (1998), Vaida & Blanchard (2005) and Wager, Vaida and Kauermann (2007) and generalised cross validation (GCV) by Craven and Wahba (1979). Alternatively, maximum likelihood via linear mixed models (LMM) or restricted maximum likelihood (Patterson & Thompson, 1971) can be used to fit nonparametric models (Wahba, 1990; Speed, 1991; Wand, 2003). The theory behind these linear smoothing parameter criteria are quite well understood.

For classification problems, i.e. where the y_is take categorical, usually binary, values there are far fewer criteria to choose from. In some commercial applications, subtleties over different response subtypes are often ignored in exchange for speed. For example, in the data mining application `MARS` (Salford Systems, 2000), based on the multivariate adaptive regression spline paper of Friedman (1991), binary responses (represented as 0/1 variables) are modelled using linear regression splines and classifications are made based on whether the regression function f exceeds $\frac{1}{2}$. A sounder, but slower, alternative is to use logistic linear mixed models which extend LMMs for smoothing parameter selection for binary responses. Finally, the generalisation bounds of Vapnik (1998, 2000) which can be used for hinge loss models for binary data are the basis for support vector machine (SVM) methods and their many variants (Schölkopf and Smola, 2002).

1.2.1 Generalised Linear Mixed Models

In contrast to typical data mining approaches, semiparametric regression via generalised linear mixed models (GLMM) offer a wider range of modelling alternatives to specifically take advantage of the structure of each different data type. For example, positive continuous responses can be modelled via gamma and inverse-Gaussian LMMs, and Poisson LMMs can be used to model count data. Different links can be used to model binary data for Bernoulli LMMs. Furthermore, unlike models based on hinge loss, the criteria to be optimised are smooth and so model fitting is much more straightforward to implement. In short, GLMMs are capable of handling a larger class of responses than are typically handled in data mining applications in a variety of different ways. The flexibility of GLMMs to handle both regression and classification problems, as well as a host of

other problems is the reason we will focus on GLMMs for smoothing parameter selection in this thesis.

Suppose we have been given the data $(y_i, \mathbf{x}_i)$, $1 \leq i \leq n$ and wish to predict the y_is based on the covariates $\mathbf{x}_i$ where each $\mathbf{x}_i$ is a row vector of dimension d with $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$. For the time being we will assume that none of the $\mathbf{x}_i$s or y_is contain missing values. Further suppose that the response vector $\mathbf{y}$ is modelled using the exponential family of distributions given by

$$\log[\mathbf{y}|\boldsymbol{\eta}] = \frac{\mathbf{y}^T\theta(\boldsymbol{\eta}) - \mathbf{1}^T b(\theta(\boldsymbol{\eta}))}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}, \phi) \tag{1.3}$$

where η_i is the predictor vector which depends on the $\mathbf{x}_i$s, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$, $\theta(\eta_i)$ is the canonical parameter, $\theta(\boldsymbol{\eta}) = (\theta(\eta_1), \ldots, \theta(\eta_n))$, $b(\theta_i)$ is the cumulant function (which is convex), $b(\theta(\boldsymbol{\eta})) = (b(\theta(\eta_1)), \ldots, b(\theta(\eta_n)))$, $c(y_i, \phi)$ is a normalising function, $c(\mathbf{y}, \phi) = (c(y_1, \phi), \ldots, c(y_n, \phi))$ and ϕ is a nuisance parameter. The mean and covariance is related to $\theta(\boldsymbol{\eta})$ via the equations

$$\mathbb{E}(\mathbf{y}|\boldsymbol{\eta}) = \boldsymbol{\mu} = b'(\theta(\boldsymbol{\eta}))$$

and

$$\mathrm{Cov}(\mathbf{y}|\boldsymbol{\eta}) = a(\phi)\mathrm{diag}(b''(\theta(\boldsymbol{\eta}))).$$

The link function $g(\cdot)$ determines the relationship between the mean μ_i and the predictor η_i via the equation

$$\eta_i = g(\mu_i)$$

The *canonical link* is the link function $g(\cdot)$ for which $\theta(\eta_i) = \eta_i$. Table 1.2.1 contains values for $g(\mu_i)$, $\mu(\theta_i)$, $\theta(\eta_i)$, $a(\phi)$, $b(\theta(\eta_i))$ and $c(y_i, \phi)$ for most of the models we will consider in this thesis. Note for gamma and inverse-Gaussian LMMs we are not using the canonical link. Instead the log link is used to ensure $\mathbb{E}(\mathbf{y}|\boldsymbol{\eta})$ is positive, a constraint required by these distributions.

We use the parameter η_i to model the dependence on the mean of $\mathbf{y}$ by the $\mathbf{x}_i$s via the equation $\eta_i = \eta(\mathbf{x}_i)$. We model $\eta(\mathbf{x}_i)$ as a linear combination of basis functions, described in detail in Sections 1.2.2–1.2.4 and Chapter 2, which we can write as

$$\eta(\mathbf{x}_i) = \eta_i = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i \tag{1.4}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are $n \times p$ and $n \times q$ matrices. We model $\mathbf{u}$ as a vector of random effects with distribution

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}_{\boldsymbol{\sigma}^2}) \tag{1.5}$$

where the $q \times q$ matrix $\mathbf{G}_{\boldsymbol{\sigma}^2}$ can be modelled using a variety of covariance structures (Zhao, Staudenmayer, Coull & Wand, 2006) and $\boldsymbol{\sigma}^2$ are called variance components which are used to parameterise $\mathbf{G}_{\boldsymbol{\sigma}^2}$. All of the examples considered in this thesis will use the

Model	$g(\mu_i)$	$\mu(\theta_i)$	$\theta(\eta_i)$	$a(\phi)$	$b(\theta(\eta_i))$	$c(y_i, \phi)$	$\frac{\partial a(\phi)}{\partial \phi}$	$\frac{\partial c(y_i, \phi)}{\partial \phi}$
Gaussian $y_i \sim N(\eta_i, \phi), \phi > 0$	μ_i	θ_i	η_i	ϕ	$\theta_i^2/2$	$-\frac{y_i^2}{2\phi} - \frac{\log(2\pi\phi)}{2}$	1	$\frac{y_i^2}{2\phi^2} - \frac{1}{2\phi}$
Logistic $y_i \sim$ Bernoulli $\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)$	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{e^{\theta_i}}{1+e^{\theta_i}}$	η_i	1	$\log(1+e^{\theta_i})$	0		
Poisson $y_i \sim$ Poisson (e^{η_i})	$\log(\mu_i)$	e^{θ_i}	η_i	1	e^{θ_i}	$-\log(y_i!)$		
Gamma $y_i \sim$ Gamma $\left(\phi^{-1}e^{\eta_i}, \phi\right), \phi > 0$	$\log(\mu_i)$	$-\frac{1}{\theta_i}$	$-e^{-\eta_i}$	ϕ^{-1}	$-\log(-\theta_i)$	$\phi\log(\phi y_i) - \log(y_i) - \log\Gamma(\phi)$	$-\phi^{-2}$	$\log(\phi y_i) + 1 - \psi(\phi)$
Inverse-Gaussian $y_i \sim IN\left(e^{\eta_i}, \phi\right), \phi > 0$	$\log(\mu_i)$	$\frac{1}{\sqrt{-2\theta_i}}$	$-\frac{e^{-2\eta_i}}{2}$	ϕ	$-\sqrt{-2\theta_i}$	$-\frac{1}{2y_i\phi} - \frac{1}{2}\log(2\pi\phi y_i^3)$	1	$\frac{1}{2y_i\phi^2} - \frac{1}{2\phi}$

Table 1.2.1: *A summary of model parameters for* $[\mathbf{y}|\boldsymbol{\eta}]$ *in equation (1.3). Here* $\theta_i = \theta(\eta_i)$.

covariance structure

$$\mathbf{G}_{\boldsymbol{\sigma}^2} = \sum_{i=1}^{v} \sigma_i^2 \mathbf{D}_i^{-1} \tag{1.6}$$

This mechanism effectively provides a quadratic penalty on the coefficient vector $\mathbf{u}$. Many software applications expect $\mathbf{G}_{\boldsymbol{\sigma}^2}$ to block diagonal multiples of the identity matrix. We will call this standard or canonical mixed model form. We will also denote the matrix $\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}$ by $\mathbf{D}_{\boldsymbol{\sigma}^2}$.

The marginal likelihood is obtained by "integrating out" the random effects vector $\mathbf{u}$. The marginal log-likelihood for this model can be written as

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \phi) &= \log \int [\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \phi][\mathbf{u}; \boldsymbol{\sigma}^2] d\mathbf{u} \\ &= \log \int \exp\left\{ \frac{\mathbf{y}^T \theta(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\theta(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}))}{a(\phi)} - \frac{1}{2}\mathbf{u}^T \mathbf{D}_{\boldsymbol{\sigma}^2} \mathbf{u} \right\} d\mathbf{u} \\ &\quad + \mathbf{1}^T c(\mathbf{y}, \phi) + \frac{1}{2} \log |\mathbf{D}_{\boldsymbol{\sigma}^2}| - \frac{q}{2} \log(2\pi) \end{aligned} \tag{1.7}$$

where $\lambda_i = a(\phi)\sigma_i^{-2}$ are the smoothing parameters. In general there is no closed form expression for (1.7) except in the case where $\mathbf{y}|\mathbf{u}$ is Gaussian where equations (1.3–1.5) describe a LMM. In this case statistical theory is extremely mature (Searle, Casella, & McCulloch, 1992; and Verbeke & Molenberghs, 2000) mainly due to the fact that in this special case the integral (1.7) is analytically tractable.

Suppose that $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\phi})$ is the maximum likelihood estimator of $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \phi)$. For a given $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \phi)$ the *best predictor* of $\mathbf{u}$ is

$$\widetilde{\mathbf{u}} = \mathbb{E}(\mathbf{u}|\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \phi) \tag{1.8}$$

which suggests the predictor

$$\widehat{\mathbf{u}} = \mathbb{E}(\mathbf{u}|\mathbf{y}; \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\phi}). \tag{1.9}$$

Using these we make predictions using

$$\mathbb{E}(\mathbf{y}|\mathbf{u}) = b'(\theta(\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{u}})). \tag{1.10}$$

Alternatively, when logistic regression is used for the purposes of classification then the predicted class of y_{new} based on $\mathbf{x}_{\text{new}}$ is given by

$$\text{sign}\left\{\widehat{\eta}(\mathbf{x}_{\text{new}})\right\}. \tag{1.11}$$

The quality of predictions can be measured in a number of ways. These include, amongst others: deviance, classification error and Akaike information criteria. Suppose we reparameterize (1.3) in terms of the mean value parameter vector $\boldsymbol{\mu}$ instead of the canonical parameter vector $\boldsymbol{\theta}$ and denote the conditional log-likelihood $\log[\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \phi]$ as $\ell(\boldsymbol{\mu}, \phi; \mathbf{y})$. Then the deviance for a given $\widehat{\boldsymbol{\mu}}$ (and fixed ϕ) is defined by

$$\mathcal{D}(\mathbf{y}, \widehat{\boldsymbol{\mu}}) \quad = -2\left\{\ell(\widehat{\boldsymbol{\mu}}, \phi; \mathbf{y}) - \ell(\mathbf{y}, \phi; \mathbf{y})\right\}. \tag{1.12}$$

The deviances for particular distributions are given in Table 1.2.2 (McCullagh & Nelder, 1989).

Model	Deviance
Normal	$\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2$
Logistic	$2\sum_{i=1}^{n} y_i \log(y_i/\widehat{\mu}_i) + (1-y_i)\log((1-y_i)/(1-\widehat{\mu}_i))$
Poisson	$2\sum_{i=1}^{n} y_i \log(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)$
Gamma	$2\sum_{i=1}^{n} -\log(y_i/\widehat{\mu}_i) + (y_i - \widehat{\mu}_i)/\widehat{\mu}_i$
Inverse-Gaussian	$\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2/(\widehat{\mu}_i^2 y_i)$

Table 1.2.2: *A summary of deviances for each of the models in Table 1.2.1.*

The deviance discrepancy measure can be also useful in simulation settings where the true mean $\boldsymbol{\mu}^*$ is known. In this case we can measure the quality of the fit using $\mathcal{D}(\boldsymbol{\mu}^*, \widehat{\boldsymbol{\mu}})$.

Bayesian GLMMs are also of considerable interest. Within this context we will only consider the simplest case of using inverse-gamma conjugate priors on the variance components $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_v^2)$ and on the nuisance parameter ϕ, i.e.

$$\begin{aligned} \sigma_i^2 &\sim IG(A_{\sigma^2,i}, B_{\sigma^2,i}) \\ \phi &\sim IG(A_\phi, B_\phi) \end{aligned} \tag{1.13}$$

where $(A_{\sigma^2,i}, B_{\sigma^2,i})$, $1 \leq i \leq v$ and (A_ϕ, B_ϕ) are chosen sufficiently small characterising little prior knowledge about the parameters σ_i^2 and ϕ. In this case the priors are called diffuse, vague or noninformative. We note that other priors, especially for σ_i^2 are also used (see for example, Gelman, 2006), however for simplicity we will focus on these priors. Furthermore, since the prior hyperparameters are fixed, $\mathbf{x}$ should be scaled to have zero mean and unit variance to improve scale invariance. This has the additional benefit of improving numerical stability when fitting this model.

1.2.2 Additive Models

The modelling of covariates of mixed types for the purposes of interpretability can be handled by imposing a particular structure on $\eta(\mathbf{x}_i)$. Interpretability of results depends highly on the user's ability to visualise the surface $\eta(\mathbf{x}_i)$. Several alternatives to aid with this exist. These include additive models (Hastie & Tibshirani, 1990), analysis of variance decomposition models (ANOVA, e.g. Gu, 2002) and tree structures (Breiman, Friedman, Olshen & Stone, 1984). For simplicity, throughout this thesis we will focus on additive models. Additive models estimate functions using a sum of lower dimensional functions. For example,

$$\eta(\mathbf{x}) = \beta_0 + \eta_1(x_1) + \eta_2(x_2) + \eta_3(x_3) \tag{1.14}$$

or

$$\eta(\mathbf{x}) = \beta_0 + \eta_1(x_1) + \eta_{2,3}(x_2, x_3) \tag{1.15}$$

where η_j is a function of x_j for all j and $\eta_{2,3}$ and is a function of both x_2 and x_3. In general let $\mathcal{I} = \{I_1, \ldots, I_{|\mathcal{I}|}\}$ be a partition of a subset of the indices $\{1, \ldots, d\}$. For example if $d = 3$ then $\mathcal{I} = \{1, 2, 3\}$ corresponds to the case (1.14) while $\mathcal{I} = \{1, \{2, 3\}\}$ corresponds to the case (1.15). We can now write

$$\eta(\mathbf{x}) = \beta_0 + \sum_{i=1}^{|\mathcal{I}|} \eta_i(\mathbf{x}_{I_i}).$$

We model each of the $\eta_{I_i}(\mathbf{x}_{I_i})$ depending on the variables type(s) of $\mathbf{x}_{I_i}$, the dimension of I_i and other prior information specific to the problem at hand.

Suppose that $\mathbf{x}_{I_i}$ is one dimensional. If $\mathbf{x}_{I_i}$ is binary where $\mathbf{x}_{I_i}$ is encoded as 0 or 1, i.e. $\mathbf{x}_{I_i} \in \{0, 1\}$ then

$$\eta_i(\mathbf{x}_{I_i}) = \mathbb{I}_{\{\mathbf{x}_{I_i}=1\}}\beta_{i1}$$

and let $\mathbf{X}_i = (\mathbb{I}_{\{\mathbf{x}_{jI_i}=1\}})$, $1 \leq j \leq n$, $\mathbf{Z}_i = \emptyset$, $\mathbf{u}_i = \emptyset$ and $\mathbf{\Omega}_i = \emptyset$.

If $\mathbf{x}_{I_i}$ is a nominal categorical variable where $\mathbf{x}_{I_i} \in \{1, \ldots, C\}$ then

$$\eta_i(\mathbf{x}_{I_i}) = \sum_{j=2}^{C} \mathbb{I}_{\{\mathbf{x}_{I_i}=j\}}\beta_{ij}$$

with $\boldsymbol{\beta}_i = (\beta_{i2}, \ldots, \beta_{iC})$, $\mathbf{X}_i = [\mathbb{I}_{\{\mathbf{x}_{jI_i}=k\}}]_{jk}$, $1 \leq j \leq n$, $2 \leq k \leq C$, $\mathbf{Z}_i = \emptyset$, $\mathbf{u}_i = \emptyset$ and $\mathbf{\Omega}_i = \emptyset$.

If $\mathbf{x}_{I_i}$ is an ordinal categorical variable where $\mathbf{x}_{I_i} \in \{1, \ldots, C\}$ then we could use the same form for $\eta_i(\mathbf{x}_{I_i})$ as for nominal categorical variables or alternatively we could use

$$\eta_i(\mathbf{x}_{I_i}) = \mathbf{x}_{I_i}\beta_{i1}$$

with $\boldsymbol{\beta}_i = (\beta_{i1})$, $\mathbf{X}_i = [\mathbf{x}_{jI_i}]_j$, $1 \leq j \leq n$, $\mathbf{Z}_i = \emptyset$, $\mathbf{u}_i = \emptyset$ and $\mathbf{\Omega}_i = \emptyset$.

If $\mathbf{x}_{I_i}$ is a continuous variable then we normally construct $\eta_i(\mathbf{x}_{I_i})$ as a linear combination of spline functions

$$\eta_i(\mathbf{x}_{I_i}) = \sum_{j=1}^{p_i} \beta_{ij} X_{ij}(\mathbf{x}_{I_i}) + \sum_{j=1}^{q_i} u_{ij} Z_{ij}(\mathbf{x}_{I_i})$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip_i})$, $\mathbf{u}_i = (u_{i1}, \ldots, u_{iq_i})$ and let

$$\begin{aligned} \mathbf{X}_i &= [X_{ik}(\mathbf{x}_{jI_i})]_{ij} \text{ for } 1 \leq j \leq n, 1 \leq k \leq p_i \quad \text{and} \\ \mathbf{Z}_i &= [Z_{ik}(\mathbf{x}_{jI_i})]_{ij} \text{ for } 1 \leq j \leq n, 1 \leq k \leq q_i. \end{aligned}$$

Associated with the spline functions $\{Z_{ik}(\mathbf{x}_{I_i})\}_{k=1}^{q_i}$ is a penalty matrix $\mathbf{\Omega}_i$. The basis functions $\{X_{ik}(\mathbf{x}_{I_i})\}_{k=1}^{p_i}$, $\{Z_{ik}(\mathbf{x}_{I_i})\}_{k=1}^{q_i}$ and penalty matrix $\mathbf{\Omega}_i$ can be modelled in a variety of ways which will explore in Sections 1.2.3–1.2.4 and in Chapter 2.

There are a number of ways to specify the $\eta_i(\mathbf{x}_{I_i})$ for multidimensional $\mathbf{x}_{I_i}$. A simple approach is to construct $\eta_i(\mathbf{x}_{I_i})$ via a tensor products of univariate functions of each $\iota \in I$. In this case we could use

$$\eta_i(\mathbf{x}_{I_i}) = \otimes_{\iota \in I_i} \eta_i(x_\iota).$$

although it is not always clear how to define the $\mathbf{\Omega}_i$ matrix is such cases. In fact, benefits can be obtained from using multiple penalties for tensor products of splines (Wood, 2006). Unfortunately these cannot be put into canonical mixed model form. Alternatively, multidimensional splines can be used, the prime example of which is the class of splines which are called radial splines (Wahba, 1990; Ruppert *et al.*, 2003; Fasshauer, 2007). These include thin plate splines (Wood, 2003) which are also very popular and are described in Section 1.2.4.

Although these ideas generalise to high dimensions, due to interpretability and curse of dimensionality issues, it is rare to have more than 2 or 3 variables handled together.

Finally, we can construct the $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{D}_i$ matrices using

$$\begin{aligned} \mathbf{X} &\equiv [\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_{|\mathcal{I}|}] \\ \mathbf{Z} &\equiv [\mathbf{Z}_1, \ldots, \mathbf{Z}_{|\mathcal{I}|}] \\ \mathbf{D}_i &\equiv \underset{1 \le j \le |\mathcal{I}|}{\text{blockdiag}} \left(\mathbf{\Omega}_j \mathbb{I}_{\{j=i\}} \right) \end{aligned}$$

and $\boldsymbol{\beta} \equiv (\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{|\mathcal{I}|})$ and $\mathbf{u} \equiv (\mathbf{u}_1, \ldots, \mathbf{u}_{|\mathcal{I}|})$ with $v = |\mathcal{I}|$ being the number of variance components.

1.2.3 Univariate Splines

There are numerous spline functions which are used for function approximation. Some common choices are *truncated power splines*, *B-splines* and *thin plate splines*. For univariate x each of these are bases for the set of polynomial splines on the interval $[a, b]$ defined by

$$\begin{aligned} \mathcal{S}_m(\boldsymbol{\kappa}) &= \{ f : [a,b] \to \mathbb{R} \mid f \in \mathcal{P}^m \text{ on } x \in (\kappa_i, \kappa_{i+1}) \text{ for } 0 \le i \le k \\ &\qquad \text{and } f(\kappa_i) \in \mathcal{C}^{m-1} \text{ for } 1 \le i \le k \} \end{aligned}$$

where $\mathcal{P}^m$ is the set of polynomials of degree m or less, $\mathcal{C}^i$ denotes the set of ith times continuously differentiable functions, and $\boldsymbol{\kappa} = (\kappa_0, \ldots, \kappa_{K+1})$ is a sequence of knots satisfying $a = \kappa_0 < \kappa_1 < \kappa_2 < \ldots < \kappa_K < \kappa_{K+1} = b$.

The truncated power basis is a less commonly used basis due to numerical instability when using the basis in practice. On the other hand the truncated power basis is intuitively simple and has the advantage that each basis function only depends on one knot. The truncated power spline basis is written (for example, Ruppert *et al.* 2003) as $\{1, \ldots, x^m, (x-\kappa_1)_+^m, \ldots, (x-\kappa_K)_+^m\}$ where the appropriate $\mathbf{X}$ and $\mathbf{Z}$ matrices are

$$\mathbf{X} = [1, x_i \ldots, x_i^m]_{1 \le i \le n} \quad \text{and} \quad \mathbf{Z} = \underset{1 \le i \le n, 1 \le j \le K}{\left[(\mathbf{x}_i - \boldsymbol{\kappa}_j)_+^m \right]}$$

and it is common to use $\mathbf{\Omega} = \mathbf{I}_K$ for this basis.

The B-spline basis is well known for its numerical robustness compared to other bases for $\mathcal{S}_m(\boldsymbol{\kappa})$ and its compact support (de Boor, 1972; Lyche & Schumaker, 1973). The basis is defined by $\{B_{m,i}(x;\boldsymbol{\kappa})\}_{1\leq i\leq m}$ where

$$\begin{aligned} B_{m,i}(x;\boldsymbol{\kappa}) &\equiv (\kappa_{i+m+1}-\kappa_i)[\kappa_i,\ldots,\kappa_{i+m+1}](\cdot - x)_+^m \\ &\equiv \frac{x-\kappa_i}{\kappa_{i+m}-\kappa_i}Q_{m-1,i}(x;\boldsymbol{\kappa}) + \frac{\kappa_{i+m+1}-x}{\kappa_{i+m+1}-\kappa_{i+1}}Q_{m-1,i+1}(x;\boldsymbol{\kappa}) \end{aligned} \tag{1.16}$$

where "[ ]" denotes the divided difference which has the following properties

$$\begin{aligned} [\kappa_i,\kappa_j]g(\cdot) &\equiv (g(\kappa_j)-g(\kappa_i))/(\kappa_j-\kappa_i) \\ [\kappa_1,\ldots,\kappa_n]g(\cdot) &\equiv ([\kappa_2,\ldots,\kappa_n]g(\cdot)-[\kappa_1,\ldots,\kappa_{n-1}]g(\cdot))/(\kappa_n-\kappa_1) \\ [\kappa_1,\ldots,\kappa_n]g(\cdot) &\equiv g^{(n-1)}(t) \text{ if } \kappa_1=\ldots=\kappa_n=t \end{aligned}$$

and

$$Q_{m,i}(x;\boldsymbol{\kappa}) \equiv \begin{cases} B_{m,i}(x;\boldsymbol{\kappa}) & \kappa_{i+m}>\kappa_i \\ 0 & \text{otherwise.} \end{cases} \qquad B_{0,i}(x;\boldsymbol{\kappa}) \equiv \begin{cases} 1 & x\in[\kappa_i,\kappa_{i+1}) \\ 0 & \text{otherwise.} \end{cases}$$

For B-splines $\boldsymbol{\kappa}$ is the extended knot vector sequence

$$a=\kappa_1=\kappa_2=\kappa_3=\kappa_4<\kappa_5<\ldots<\kappa_{K+4}<\kappa_{K+5}=\kappa_{K+6}=\kappa_{K+7}=\kappa_{K+8}=b.$$

We will consider B-splines including computational aspects of an appropriate matrix $\boldsymbol{\Omega}$ in Chapter 2. Finally, thin plate splines generalise to higher dimensions and are considered in the next section.

1.2.4 Multivariate Splines

As previously stated, multidimensional splines can be constructed using a tensor product of univariate splines and a mesh of knots. The downside of such an approach is that the number of knots increases exponentially with the dimension of the data. Radial basis splines represent a class of meshfree splines which avoid this problem by allowing knots to be specified independently in x space and are described in Ruppert *et al.* (2003, Chapter 13) and in much greater detail in Fasshauer (2007).

Thin plate splines are perhaps the most popular type of radial basis spline (Ruppert *et al.*, 2003; Wood, 2003). For *thin plate splines* the matrix $\mathbf{X}$ has columns spanning the space of all $d=\dim(\mathbf{x})$ dimensional polynomials in the components of x with degree less than some integer m satisfying $2m-d>0$ (except the intercept),

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

where $\mathbf{\Omega} = [C(\|\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}\|)]^{-1/2}_{1 \leq k,k' \leq K}$ has the singular value decomposition $\mathbf{\Omega} = \mathbf{U}\text{diag}(\mathbf{d})\mathbf{V}^T$ so that $\mathbf{\Omega}^{-1/2} = \mathbf{V}\text{diag}(\mathbf{d}^{-1/2})\mathbf{U}^T$,

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

and $\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_K$ are knots of dimension d (see Wood, 2003 or Ruppert *et al.*, 2003, Chapter 13 for details).

These knots may be selected in a number of ways. The simplest of these is to use an equally spaced grid of points (of a specified size) but again the number of knots increase exponentially with the dimension of $\mathbf{x}$. Computationally much better alternatives are to use quasi-random sequences, for example Halton sequences (see for example, Fasshauer, 2007), or to use space filling designs via clustering (Nychka & Sultzman, 1998) and implemented in the `R` package `FUNFITS` (Nychka *et al.*, 1998). The appropriate penalty matrix is $\mathbf{\Omega} = \mathbf{I}_K$. One property that distinguishes these bases, which we will exploit in Chapter 6, is that the number of columns in $\mathbf{Z}$ and $\mathbf{\Omega}$, the number of penalised basis functions, are equal to the number knots.

1.3 Semiparametric Regression for Data Mining

It is clear from the above discussion that semiparametric regression can be used naturally to handle mixed data types and can be constructed so as to produce interpretable models. The other complications associated with data mining listed in the introduction are matters of ongoing research. We will now discuss some of the progress made in these areas.

1.3.1 Computational Scalability

One of the key problems with semiparametric regression via GLMMs when applied to data mining problems is computational scalability. The key problem with GLMMs is the intractability of the marginal likelihood. This intractability has been the driving force behind much of the research into GLMMs for the past few decades. The aim of obtaining accurate estimators computationally efficiently remains elusive. The computational scalability of finding estimators to GLMMs largely depends on the method of approximation chosen. This in turn depends on the relative trade-off between accuracy and computational efficiency. These approximations can be roughly categorised as analytic or numerical in their approach.

Analytic approximations for fitting GLMMs include Laplace's method (Wolfinger, 1993); penalised quasi-likelihood (Breslow & Clayton, 1993) and Solomon-Cox approximations (Solomon & Cox, 1992). These approximations have the advantage of being computationally fast but are comparatively crude approximations of the marginal likelihood. Furthermore such estimators which can have significant bias (Breslow & Lin, 1995; Lin & Breslow, 1996; Sutradhara & Rao, 2001). Nevertheless analytic approximations can be useful in a number of contexts; they can be used

- as a starting point for other more accurate approximations,
- as the basis for a model selection procedure (for example in Chapter 3), and
- when criteria other than accuracy of approximating the marginal likelihood or biases are of utmost importance, for example residual deviance or classification error (Kauermann, Ormerod & Wand, 2008).

Analytic approximations are typically based on using information on the integrand's derivatives. For example Laplace's approximation of the integral $\int e^{-tg(\mathbf{x})}d\mathbf{x}$ is (assuming g is twice continuously differentiable)

$$\int e^{-tg(\mathbf{x})}d\mathbf{x} = \sqrt{\frac{(2\pi/t)^{\dim(\mathbf{x})}}{\mathsf{H}_{\mathbf{x}}g(\widehat{\mathbf{x}})}}e^{-tg(\widehat{\mathbf{x}})} + O(t^{-1}) \tag{1.17}$$

where $\widehat{\mathbf{x}}$ maximises g, i.e.

$$\mathsf{D}_{\mathbf{x}}g(\widehat{\mathbf{x}}) = \mathbf{0}. \tag{1.18}$$

The right hand side of (1.17) becomes more accurate as $t \to \infty$ (e.g. Barndorff-Nielsen & Cox, 1989; Tierney, Kass & Kadane, 1989; Raudenbush, Yang & Yosef, 2000; Young & Smith, 2005). Fortunately, even though for most models we consider $t = 1$, Laplace's method may be reasonably accurate, in particular when the integrand (1.17) is proportionally similar to Gaussian in shape. Such is the case for GLMMs where the posterior distribution $\mathbf{u}|\mathbf{y}$ is nearly Gaussian in shape.

Some effort has been made to improve the accuracy of these methods using higher order Laplace approximations (Raudenbush *et al.*, 2000). Unfortunately the computational cost of such methods increases exponentially in the order of the Laplace approximation. In practice this means that, for large datasets or complicated models, only moderate improvements in accuracy are possible. Alternatively, higher order approximations can be obtained when considering ratios of integrals of the form (1.17), e.g. Tierney *et al.* (1989). These occur in cases when calculating iterates of an expectation maximisation (EM) algorithm and its variants, or the best predictor for the random effects $\mathbf{u}$ in equation (1.9, see also Section 10.8 of Ruppert *et al.*, 2003).

Numeric approximations tend to be far more accurate than analytic approximations but are usually much slower. The practical use of some numeric approximations can be restricted when the dimension of the integrals to be evaluated is high or where the dataset is sufficiently large or complex. The two most commonly used approximations are Gauss-Hermite quadrature (GHQ, see e.g. Naylor & Smith, 1982; Lesaffre & Spiessens, 2001) and Monte Carlo methods (McColloch, 1994, 1997; Gelman, Carlin, Stern & Rubin, 1995; Clayton, 1996; Gilks, Richardson & Spiegelhalter, 1996; Robert & Casella, 1999).

The Gauss-Hermite quadrature approach has proven to be very effective in analyses for generalised longitudinal models where the integrals to be evaluated are typically low dimensional. However the use of Gauss-Hermite quadrature for more general GLMMs is in practice restricted, since the number of quadrature points increases exponentially in

the dimension of the integral to be calculated. In such cases Monte Carlo type methods are often preferred.

The most accurate method for fitting GLMMs is via Monte Carlo methods. One such variety, Markov chain Monte Carlo (MCMC) has been the driving algorithm behind Bayesian Statistics for the last two decades. Several spin-offs of MCMC methodology include Monte Carlo Expectation Maximisation (ECEM) and Monte Carlo Expectation Conditional Maximization (MCECM) methods (Wei & Tanner, 1990; Lui & Rubin, 1995; McCulloch, 1997; Booth & Hobert, 1998), Monte Carlo relative likelihood (Geyer, 1992) and Monte Carlo Newton-Raphson (Kuk & Cheng, 1997). The basic intention of these algorithms is to develop methods for sampling from posterior densities. Different sampling strategies include the traditional Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller & Teller, 1953; Hastings, 1970), adaptive rejection sampling (Gilks & Wild, 1992; Robert & Casella, 1999) and importance sampling (Rubinstein, 1981; Booth & Hobert, 1998).

Monte Carlo methods suffer from at least two major drawbacks. Firstly, major difficulties are associated in assessing the convergence of MCMC methods. While some progress has been made in this respect (e.g. Rosenthal, 1995; Cowles & Carlin, 1996; Cowles & Rosenthal, 1998), the applications of this theory have remained limited to special cases and therefore caution needs to be applied when taking such an approach. Secondly, although there has been extensive research in designing efficient MCMC samplers, such methods can still be painfully slow when the dataset to be analysed is suitably large or the model to be fitted is sufficiently complex.

Some attempts to remedy this problem include sequential Monte Carlo (SMC) and quasi-Monte Carlo methods. SMC generalises importance sampling by producing a weighted sample from the stationary distribution while retaining some of the benefits of MCMC (Del Moral, Doucet & Jasra, 2006; Fan, Leslie & Wand, 2007). Quasi-Monte Carlo methods (Hickernell, Lemieux & Owen, 2005; Kuo, Dunsmuir, Sloan, Wand & Womersley, 2008) offer yet another alternative that by-passes the problems associated with random sampling by choosing points deterministically. Convergence for this approach is provably faster than Monte Carlo methods under certain circumstances.

While numerical approximations have an assured place in statistical analysis, their application to data mining problems is highly questionable due to the problem of computational scalability. Furthermore, along the philosophy of Tukey (1954, 1962) and Box (1979), all models are approximations. Thus, approximate solutions to "more realistic" models are better than fitting "less realistic" models exactly. Much of this thesis is dedicated instead towards developing analytic approximations to models.

1.3.2 *Missing Values*

Missing data is a common complication in many statistical analyses. It occurs in many fields including the social sciences when dealing with surveys, clinical trials when patients are dropped from a study, in engineering because of equipment malfunction, in data mining for example when new data becomes available or data entry when a field in a form is overlooked. It can also occur by design, for example for confidentiality reasons. Dealing with missing values is often an ignored problem in many statistical analyses. Ignoring this problem can have disastrous consequences including underestimated variances, less efficient and biased estimators and ultimately incorrect inferences.

There have been many approaches to missing value problems in general (Schafer, 1997; Little & Rubin, 2002). Some of the most successful of these methods are likelihood based models (Ibrahim, 1990; Ibrahim *et al.*, 2001; Little & Rubin, 2002). In Chapter 4 we will examine some simple missing value models, so it is necessary to introduce some nomenclature.

Let $\mathbf{X}$ be an $n \times d$ matrix of observations (covariates) where some of the observations are missing. Rubin (1976) formalised nomenclature for the missing data mechanism by introducing the indicator matrix $\mathbf{M}$, with entries $M_{ij} = 0$ if X_{ij} is observed and $M_{ij} = 1$ if X_{ij} is missing. A parametric model then specifies the joint distribution of $\mathbf{X}$ and $\mathbf{M}$. There are two main ways if specifying the joint distribution of $\mathbf{X}$ and $\mathbf{M}$. *Selection models* specify

$$[\mathbf{X}, \mathbf{M}|\boldsymbol{\theta}, \boldsymbol{\vartheta}] = [\mathbf{X}|\boldsymbol{\theta}][\mathbf{M}|\mathbf{X}, \boldsymbol{\vartheta}], \tag{1.19}$$

where $[\mathbf{X}|\boldsymbol{\theta}]$ represents the complete model for $\mathbf{X}$, $[\mathbf{M}|\mathbf{X}, \boldsymbol{\vartheta}]$ represents the model for the missing data mechanism, and $(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ are unknown parameters. The second main approach to specifying the joint distribution of $\mathbf{X}$ and $\mathbf{M}$ are called pattern mixture models which specify

$$[\mathbf{X}, \mathbf{M}|\boldsymbol{\psi}, \boldsymbol{\varphi}] = [\mathbf{X}|\mathbf{M}, \boldsymbol{\psi}][\mathbf{M}|\boldsymbol{\varphi}], \tag{1.20}$$

where the distribution of $\mathbf{X}$ depends on the missing data pattern $\mathbf{M}$ and $(\boldsymbol{\psi}, \boldsymbol{\varphi})$ are unknown parameters, possibly different from $(\boldsymbol{\theta}, \boldsymbol{\vartheta})$.

Equations (1.19) and (1.20) represent two different ways of factoring $[\mathbf{X}, \mathbf{M}|\boldsymbol{\phi}, \boldsymbol{\varphi}]$. Rubin calls the data, where missing, *missing at completely at random* (MCAR) if $\mathbf{M}$ is independent of $\mathbf{X}$ and in this case the two model specifications (1.19) and (1.20) are equivalent if $\boldsymbol{\theta} = \boldsymbol{\phi}$ and $\boldsymbol{\vartheta} = \boldsymbol{\varphi}$.

Many maximum likelihood missing data models are based on *ignorable* selection models where $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ are distinct and the data where missing at random which implies

$$[\mathbf{M}|\mathbf{X}, \boldsymbol{\vartheta}] = [\mathbf{M}|\mathbf{X}_{obs}, \boldsymbol{\vartheta}]$$

where $\mathbf{X}_{obs}$ denotes the set of observed components of $\mathbf{X}$. Rubin (1976) showed that maximum likelihood inference for $\boldsymbol{\theta}$ under such models does not depend on, and hence ignore, $[\mathbf{M}|\mathbf{X}, \boldsymbol{\vartheta}]$ and can be based solely on the likelihood obtained by integrating out

the missing values of $\mathbf{X}$ from the density $[\mathbf{X}|\boldsymbol{\theta}]$. Note that the MAR condition is less restrictive than MCAR.

Example 1.1 [Little, 1993]: Suppose that we have two covariates X_1 and X_2 where X_1 is fully observed while X_2 is sometimes missing. Thus there would be two cases for the rows of M, $(M_{i,1}, M_{i,2}) = (0,0)$ and $(M_{i,1}, M_{i,2}) = (0,1)$. A selection model might use

$$[M_{i,2} = 1|X_{i,1}, X_{i,2}, \boldsymbol{\vartheta}] = g(\vartheta_0 + \vartheta_1 X_{i,1} + \vartheta_2 X_{i,2})$$

where $\boldsymbol{\vartheta} = (\vartheta_0, \vartheta_1, \vartheta_2)$ and $g(\cdot)$ is some function which takes values on $[0,1]$. The data is MCAR if $\vartheta_1 = \vartheta_2 = 0$ and is MAR if $\vartheta_2 = 0$ because the missingness of X_2 depends only on the values of X_1 which are always observed. Finally if $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ are distinct and $\vartheta_2 = 0$ then the selection model is ignorable.

When selection models are used to handle missing data either one of these types of missingness are typically assumed. In addition to these are a variety of additional assumptions about the nature of the missing data mechanism, some of which are outlined and treated in Little (1992, 1993), Horton & Laird (1999), Ibrahim *et al.* (2001), Thijs, Molenberghs, Michiels, Verbeke & Curran (2002) and Horton & Kleinman, (2007).

1.3.3 Robustness

Data can contain many deficiencies which may hinder or otherwise ruin analysis. These deficiencies are particularly prevalent in data mining applications. In these applications, as previously stated, data may contain a substantial number of outliers and distributions of numeric predictor and response variables are often long-tailed and highly skewed (Hastie *et al.* 2001). In additional, amongst other difficulties, covariates may be subject to measurement error (Carroll, Ruppert, Stefanski & Crainiceanu, 2006) for function estimation, if the mean function contains jumps or cusps (i.e. change points), or if the function is spatially inhomogeneous (i.e. different levels are smoothing are required in different regions of the function). In these cases typical function estimation procedures can deliver poor results.

The topic of robustness in Statistics has been subject to an enormous amount of research over the past few decades (e.g, Hampel, Ronchetti, Rousseeuw & Stahel, 1986; Rousseeuw & Lerow, 1987; Staudte & Sheather, 1990; Wilcox, 1997). Roughly speaking when modelling our data we typically use a number of working assumptions, for example about the distribution of the data and relationships between observations. The aim for robust models is to perform not much worse within a range of alternatives to these assumptions (Garthwaite, Jolliffe & Jones, 2002).

1.3.4 Parsimony

Finally, parsimonious modelling involves looking for models which are as simple as possible, but no simpler than necessary. Thus parsimony might relate to a number of facets of a particular model. For example, in semiparametric regression, linear models are simpler than nonlinear models, models which incorporate less covariates are simpler than

models with more covariates, models with homogeneous noise are simpler than models with heterogeneous noise. For tree-type models, smaller trees with fewer variables are simpler than larger trees which use more variables.

The question of which covariates to include in a model is quite important. If there are many irrelevant covariates, as is common to data mining applications, using them can have several detrimental consequences including accuracy of fit, interpretability and additional computational costs. Penalization approaches (for example, Hastie *et al.* 2001, Chapter 3), related to the idea of shrinkage, do to some extent elevate some of the problems of dealing with irrelevant predictors, but such an approach is not perfect. A parsimonious solution containing only relevant predictors is more desirable.

The classical approach to model selection in Statistics is via hypothesis testing which dates back to the founding fathers of Statistics, Neyman & Pearson (1933) and Fisher [1935](1956). The classical hypothesis testing methods, the likelihood ratio test (the origin of which is discussed in Giri, 1964a, 1964b), the Rao score test (Rao, 1973) and the Wald test (Wald, 1943) are based on asymptotics. Recently, for models subject to constraints, more powerful versions of these tests have been developed (Self & Liang, 1987; Silvapulle & Sen, 2005).

New approaches to model selection are continually being developed. Some of these include criteria based (Akaike, 1974), optimisation based and Bayesian (Yau, Kohn & Wood, 2003) approaches. In Computing Science the topics of variable selection and feature selection (e.g. Guyon & Elisseeff, 2003) have similar aims. Finally, within the context of semiparametric regression, hypothesis testing is a matter of ongoing research. Recent accounts can be found in Hastie *et al.* (2001) and Ruppert *et al.* (2003).

1.4 Thesis Outline

As we have discussed there has been, in some respects, substantial progress in many areas of semiparametric regression related to data mining. However, most of the research areas associated with the individual difficulties in data mining (as listed in the introduction) are far from a settled state. In this thesis we make a number of contributions to these areas.

Firstly, while penalised spline methodology is in a highly mature state, at the practical, if not the theoretical level, we will examine computational aspects of a class of reduced knot smoothing splines we will call O'Sullivan splines (or O-splines for short) in Chapter 2. This work improves upon the related penalised spline methodology of Eilers & Marx (1996) which uses a finite difference approximations of the "roughness" penalty matrix. We refer to the approach which specifically uses the basis/penalty of Eilers & Marx (1996) as P-splines. We improve upon this work by deriving an exact method for calculating the roughness penalty matrix which is easy to implement, efficient to calculate and allows greater flexibility in the selection of knots. We also show that O-splines have numerical advantages over P-splines and can be seamlessly integrated into Bayesian GLMM methodology.

Secondly, in Chapter 3 we will examine the problem of model selection. We develop an algorithm which is similar in vain to the regression spline methodology of Stone, Hanson, Kooperberg & Truong (1997). This algorithm is based on PQL approximations of the GLMM (Breslow & Clayton, 1993) and approximate score Statistics (Rao, 1973, Lin 1997). In furtherance to this end we also develop an efficient method for fitting Logistic LMMs. We show that the algorithm is reasonably accurate and has better computational scalability to similar methods.

Laplace-like approximations are reasonably fast and may be used on medium scale data mining problems. When the integrand to be approximated is not Gaussian in shape their application is questionable. Such is the case where the GLMM in Section 1.2 is modified to handle the various complications arising in data mining applications. For this reason we have pursued variational approximations.

Variational methods are simple, fast and flexible class of approximations with origins in machine learning literature (Jordan, Ghahramani, Jaakkola, & Saul 1999; Corduneanu & Bishop, 2001; Ueda & Ghahramani (2002), Bishop & Winn, 2003; MacKay, 2003; Titterington, 2004; Winn & Bishop 2005). The aim of variational methods is to transform problems into optimisation problems (Jaakkola, 2001). Within the context of Statistics these methods transform integral problems into optimisation problems, usually in an approximate way. This is done by constructing a lower bound on the marginal likelihood which is tightened using an iterative scheme related to the EM algorithm of Dempster, Laird & Rubin (1977). These approximations are based on the observation that for any distribution $\delta(\psi;\boldsymbol{\xi})$ we have

$$\ell(\boldsymbol{\theta}) = \log \int [\mathbf{y}, \psi; \boldsymbol{\theta}] d\psi \geq \ell_L(\boldsymbol{\theta}; \boldsymbol{\xi}) = \mathbb{E}_\delta \log[\mathbf{y}, \psi; \boldsymbol{\theta}] + \mathcal{H}_\delta \tag{1.21}$$

where ψ is a vector of parameters which we want to integrate out, $\delta(\psi;\boldsymbol{\xi})$ is a density which approximates the conditional distribution $\psi|\mathbf{y}$, $\mathbb{E}_\delta$ denotes expectation with respect to δ and $\mathcal{H}_\delta = -\mathbb{E}_q \log(\delta(\psi;\boldsymbol{\xi}))$ is the entropy of δ. Using similar terminology to that of Jaakkola & Jordan (2000) we call (1.21) the density transform of the likelihood.

In Chapter 4 we will review variational approximations to integrals arising in both frequentist and Bayesian statistical models. We discuss methods for alternative optimisation techniques fitting these approximations, and a new method for approximating posterior densities. We then apply these approximations to some simple missing value models, and compare their speed and accuracy to MCMC approximations. Variational approximations have been successfully applied to missing problems (Saul, Jaakkola & Jordan, 1996; Jaakkola & Jordan, 2000; Williams, Liao, Xue & Carin, 2005; and Consonni & Marin, 2007) and represent an exciting alternative in the area.

As previously discussed, the key problem to using GLMMs in semiparametric regression is the intractability of the high dimensional integral in the likelihood. Thus in Chapter 5 we develop variational approximations for GLMMs and Bayesian GLMMs. In particular, we will develop Gaussian approximations which are more accurate than the

Laplace approximation in the Kullback-Leibler divergence sense. We discuss several algorithms for fitting these approximations and compare them numerically with existing methods.

Finally, in Chapter 6 we will consider some robust semiparametric models. We will consider several types of robustness, namely robustness to outlier models, variance function estimation and spatially adaptive variance components models. For continuous responses Student's t regression is often a good starting point for outlier models (Lange, Little & Taylor 1989). Instead of considering robust modelling for skewed noise, we consider heterogeneous variance models (Davidian & Carroll 1987; Carroll & Ruppert, 1988; Ruppert *et al.* 2003; Crainiceanu *et al.*, 2006). This type of model does provide some robustness to the assumption of homogeneous noise. Highly adaptive smoothing might also be viewed as a type of robustness. There are hundreds of highly adaptive methods the most accurate being the Bayesian regression spline methods of DiMatteo, Genovese & Kass (2001) and Denison, Holmes, Mallick & Smith (2002) or the genetic algorithm for regression splines of Pittman (2002). While not quite as accurate as these methods, the adaptive variance component ideas of Baladandayuthapani, Mallick & Carroll (2005), Crainiceanu, Ruppert, Carroll, Adarsh & Goodner (2007) and Krivobokova, Crainiceanu & Kauermann (2007) are almost as accurate. However, in all of these methods, except Krivobokova *et al.* (2007), models are fit via MCMC methods and would thus be inappropriate for data mining applications where speed is important. In Chapter 6 we will develop an extremely fast algorithm combining all of the above types of robustness which performs quite well in practice.

CHAPTER 2

On Semiparametric Regression with O'Sullivan Penalised Splines[1]

2.1 Introduction

Splines continue to play a central role in semiparametric regression modelling. Recent synopses include Eubank (1999), Gu (2002), Ruppert, Wand & Carroll (2003) and Denison, Holmes, Mallick & Smith (2002). In all but the last reference, smooth functional relationships are fitted using a large basis of spline functions subject to penalization. Up until the mid-1990s most literature on spline-based nonparametric regression was concerned with *smoothing splines*, and their multivariate extension *thin plate splines*, where the penalty takes a particular form and the number of basis functions roughly equals the sample size (e.g. Wahba, 1990; Green & Silverman, 1994). However, in recent years, there has been a great deal of research on more general spline/penalty strategies, most of which use considerably fewer basis functions. Driving forces include:

- more complicated models, often with several smooth functions;
- larger data sets, where smoothing and thin-plate splines become computationally intractable,
- mixed model and Bayesian representations of smoothers that lend themselves to the use of established software, such as `BUGS` (Spiegelhalter, Thomas & Best, 2000), `lme()` in `R` (R Development Core Team, 2007) and `PROC MIXED` in `SAS` (SAS Institute, Inc., 2007); provided the number of basis functions is relatively low.

Ruppert, Wand & Carroll (2003) summarise and provide access to many of these developments. The term *penalised splines* has emerged as a descriptor for general spline fitting subject to penalties. Other descriptors used in the literature include P-splines (Eilers & Marx, 1996), pseudosplines (Hastie, 1996), reduced knot splines (White, Thompson & Brotherstone, 1999) and low-rank spline smoothers (Wood, 2003).

O'Sullivan (1986, Section 3) introduced a class of penalised splines based on B-spline basis functions. O'Sullivan penalised splines are a direct generalisation of smoothing splines in that the latter arises when the maximal number of B-spline basis functions are included. Like smoothing splines, O'Sullivan penalised splines possess the attractive feature of natural boundary conditions (e.g. Green & Silverman, 1994, p.12). They have also become the most widely used class of penalised splines in statistical analyses due to their

[1]Sections 2.1-4 and 2.7 correspond to: Wand, M.P. & Ormerod, J.T. (2008). On Semiparametric Regression with O'Sullivan Penalised Splines. *Australian and New Zealand Journal of Statistics*, (in press), representing joint research between M.P Wand and J.T. Ormerod. Sections 2.5 and 2.6 contain additional material representing solo research by J.T. Ormerod.

implementation in the popular R and S-PLUS (Insightful Corporation, 2007) function `smooth.spline()` and associated generalised additive model software (e.g. the `gam` library in R; Hastie, 2006). Despite the omnipresence of O'Sullivan penalised splines, their use in semiparametric regression contexts, particularly those involving mixed model and Bayesian representations, is not very common. Recently, Welham, Cullis, Kenward & Thompson (2007) showed how most of the commonly used penalised splines can be treated within a single mixed model framework, although they did not work explicitly with the form given in O'Sullivan (1986).

Our contributions in this chapter are:

1. Provide an exact matrix expression for the penalty of O'Sullivan splines that allows implementation in a few lines of a matrix-based computing language.
2. Compare O'Sullivan splines with their closest penalised spline relative, P-splines (Eilers & Marx, 1996), which reveal some noticeable differences near the boundaries.
3. Demonstrate explicitly, including with R code, how O'Sullivan splines can be simply added to the mixed model-based regression armoury.
4. Investigate their efficacy in Bayesian semiparametric regression using Markov chain Monte Carlo (MCMC) software such as BUGS and its variants.
5. Explore several extensions of O'Sullivan splines including: general degree O'Sullivan splines and their mixed model formulation, derivative estimation and bivariate tensor product O'Sullivan splines and their mixed model formulation.

We conclude that the several attractive features of O'Sullivan penalised splines – smoothness, numerical stability, natural boundary properties, direct generalisation of smoothing splines – makes them a very good choice of basis in semiparametric regression.

2.2 O'Sullivan Penalised Splines

O'Sullivan penalised splines have already been described several times in the literature. A recent reference is the Chapter 5 Appendix of Hastie, Tibshirani & Friedman (2001). A brief sketch is given here for convenience.

Consider the simplest nonparametric regression setting

$$y_i = f(x_i) + \varepsilon_i, \; 1 \leq i \leq n \tag{2.1}$$

where $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$. Suppose that the estimate of f is required over $[a, b]$, an interval containing the x_is. For an integer $K \leq n$ let $\kappa_1, \ldots, \kappa_{K+8}$ be a knot sequence such that

$$a = \kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 < \kappa_5 < \ldots < \kappa_{K+4} < \kappa_{K+5} = \kappa_{K+6} = \kappa_{K+7} = \kappa_{K+8} = b$$

and let $B_1(\cdot), \ldots, B_{K+4}(\cdot)$ be the cubic B-spline basis functions defined by these knots (see e.g. pp.160-161 of Hastie *et al.*, 2001). Set up the $n \times (K+4)$ design matrix $\mathbf{B}$ with

(i,k)th entry $B_{ik} = B_k(x_i)$ and $\mathbf{\Omega}$ the $(K+4)\times(K+4)$ penalty matrix with (k,k')th entry

$$\mathbf{\Omega}_{kk'} = \int_a^b B_k''(x)B_{k'}''(x)dx.$$

Then an estimate of f at location $x \in \mathbb{R}$ can be obtained as

$$\widehat{f}_O(x;\lambda) \equiv \mathbf{B}_x\widehat{\boldsymbol{\nu}}_O \text{ where } \widehat{\boldsymbol{\nu}}_O \equiv (\mathbf{B}^T\mathbf{B} + \lambda\mathbf{\Omega})^{-1}\mathbf{B}^T\mathbf{y} \tag{2.2}$$

where $\mathbf{B}_x = [B_1(x), \ldots, B_{K+4}(x)]$ and $\lambda > 0$ is a smoothing parameter.

Note that the cubic smoothing spline arises in the special case $K = n$ and $\kappa_{k+4} = x_k$ for $1 \leq k \leq n$, provided the x_is are distinct (e.g. Green & Silverman, 1994, Section 3.6). Apart from giving a smooth (twice continuously differentiable) scatterplot smooth, $\widehat{f}_O(\cdot;\lambda)$ has good numerical properties. The basis functions are bounded and so not prone to overflow problems. Moreover, $\mathbf{B}^T\mathbf{B}$ is 4-banded, which leads to $O(n)$ algorithms when K is close to n (e.g. Hastie, *et al.*, 2001). In addition, $\widehat{f}_O(\cdot;\lambda)$ satisfies so-called natural boundary conditions, meaning that

$$\widehat{f}_O''(a;\lambda) = \widehat{f}_O'''(a;\lambda) = \widehat{f}_O''(b;\lambda) = \widehat{f}_O'''(a;\lambda) = 0$$

and implying that $\widehat{f}_O(\cdot;\lambda)$ is approximately linear over $[a,\kappa_5]$ and $[\kappa_{K+4},b]$ (linearity is exact if $\kappa_5 = \min(x_i)$ and $\kappa_{K+4} = \max(x_i)$). Figure 2.1 illustrates these natural boundary properties of $\widehat{f}_O(\cdot;\lambda)$ for data on ratios of strontium isotopes found in fossil shells and their age; see Chaudhuri & Marron (1999) for details. Also, $\widehat{f}_O(\cdot;\lambda)$ approximates the least squares line as $\lambda \to \infty$. The implication for mixed model smoothing is that the induced fixed effects component corresponds to straight line basis functions. Details are given in Section 2.4.

Computation of the design matrix $\mathbf{B}$ is usually quite easy. For example, B-splines are readily available in the `Matlab` (The Mathworks, Inc., 2007), `R` and `S-PLUS` computing environments. Otherwise recurrence formulae (e.g. de Boor, 1978; Eilers & Marx, 1996) can be called upon. However, computation of $\mathbf{\Omega}$ requires some additional effort. In Section 2.6.1, while treating general degree O'Sullivan penalised splines, we derive an exact matrix algebraic expression for the corresponding penalty matrices. In the cubic case our theorem reduces to the expression:

$$\mathbf{\Omega} = (\overline{\mathbf{B}}'')^T\text{diag}(\mathbf{w})\overline{\mathbf{B}}'' \tag{2.3}$$

where $\overline{\mathbf{B}}''$ is the $3(K+7)\times(K+4)$ matrix with (i,j)th entry $B_j''(\overline{x}_i)$, $\overline{x}_i$ is the ith entry of the vector

$$\overline{\mathbf{x}} = \left(\kappa_1, \frac{\kappa_1+\kappa_2}{2}, \kappa_2, \kappa_2, \frac{\kappa_2+\kappa_3}{2}, \kappa_3, \ldots, \kappa_{K+7}, \frac{\kappa_{K+7}+\kappa_{K+8}}{2}, \kappa_{K+8}\right).$$

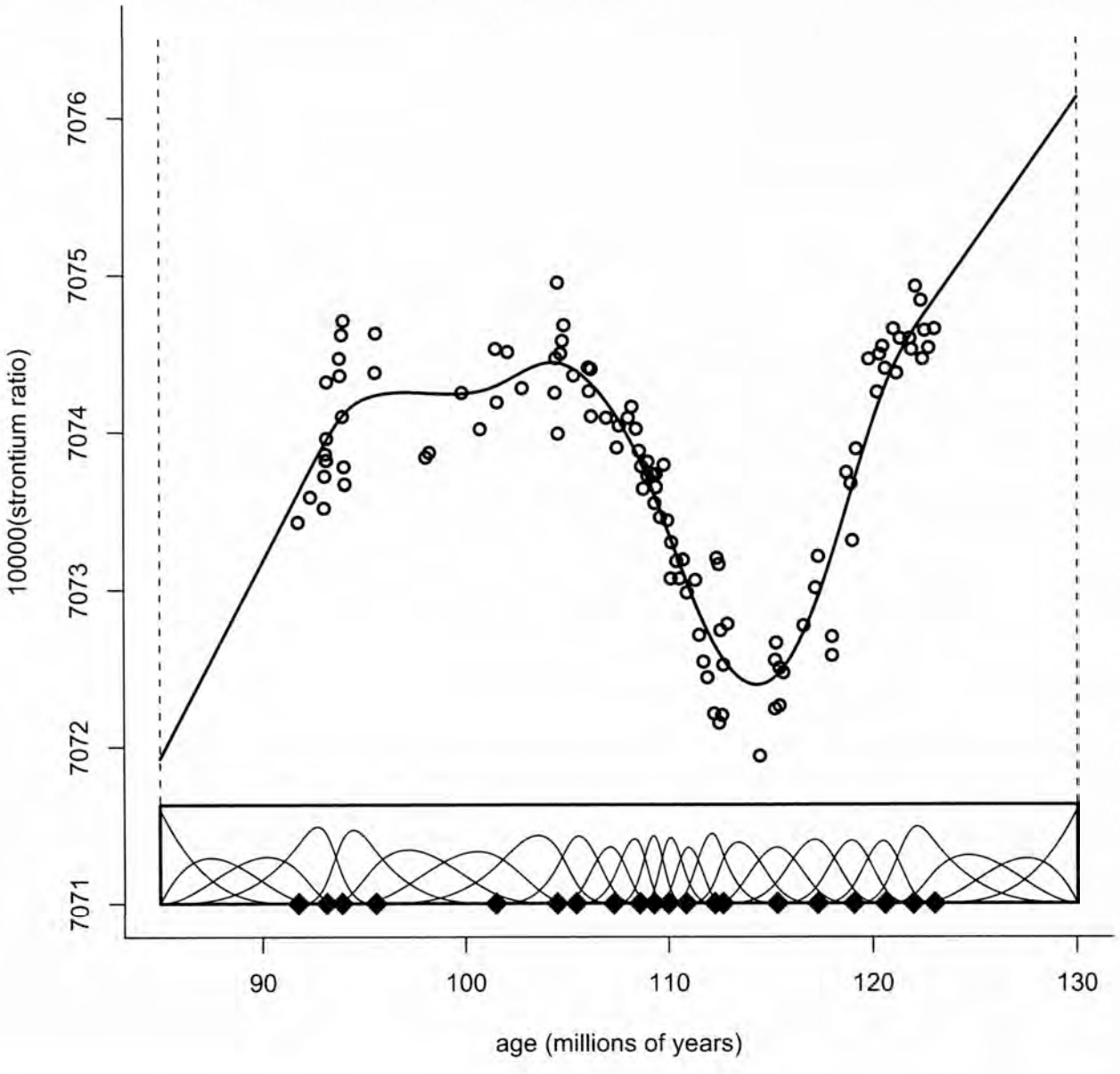


Figure 2.1: *Illustration of natural boundary properties of a 20-interior knot O'Sullivan penalised spline fit to the fossil data over the interval* $[85, 130]$ *millions of years. The interior knots are shown as solid diamonds (♦). Inset: The 24 B-spline basis functions.*

and $\mathbf{w}$ is the $3(K+7) \times 1$ vector given by

$$\mathbf{w} = \left(\frac{1}{6}(\Delta\boldsymbol{\kappa})_1, \frac{4}{6}(\Delta\boldsymbol{\kappa})_1, \frac{1}{6}(\Delta\boldsymbol{\kappa})_1, \frac{1}{6}(\Delta\boldsymbol{\kappa})_2, \frac{4}{6}(\Delta\boldsymbol{\kappa})_2, \frac{1}{6}(\Delta\boldsymbol{\kappa})_2, \ldots, \frac{1}{6}(\Delta\boldsymbol{\kappa})_{K+7}, \frac{4}{6}(\Delta\boldsymbol{\kappa})_{K+7}, \frac{1}{6}(\Delta\boldsymbol{\kappa})_{K+7}\right),$$

where $(\Delta\boldsymbol{\kappa})_k \equiv \kappa_{k+1} - \kappa_k$, $1 \leq k \leq K+7$. Result (2.3) is none other than Simpson's rule applied over each of the inter-knot intervals. This is because each $B_i'' B_j''$ function is piecewise quadratic. For commonly used values of K, (2.3) allows straightforward computation of $\boldsymbol{\Omega}$ in matrix-based languages such as `Matlab`, `R` and `S-PLUS`. In the Appendix of this chapter we demonstrate computation of $\boldsymbol{\Omega}$ in 4 lines of `R` code.

2.2.1 *Knot Selection*

Lastly, we mention knot choice. The `R` and `S-PLUS` function `smooth.spline()` uses

$$\kappa_k \simeq \left(\frac{k}{K+1}\right) \text{th sample quantile of the } x_i\text{'s.}$$

where

$$K = \begin{cases} n & n < 50 \\ 100 & n = 200 \\ 200 & n = 800 \\ 200 + (n - 3200)^{\frac{1}{5}} & n > 3200 \end{cases}$$

Other values of n between 50 and 3200 are handled via a logarithmic interpolation. For many functional relationships, fewer knots are sufficient. Figure 2.1 is one example, where only $K = 20$ interior knots are used without compromising the quality of the fit. A common default in the penalised spline literature is $K = \min(n_U/4, 35)$, where n_U is the number of unique x_i's (e.g. Ruppert *et al.*, 2003). Ruppert (2002) discusses a 'hi-tech' choice of K. The distribution of the knots, for a given K, may have some effect on the results. As mentioned above, `smooth.spline()` uses quantile-based knots while e.g. Eilers & Marx (1996) recommend equally-spaced knots. In most situations this effect will be minor. However, for either strategy, it is possible to construct regression functions and predictor variable distributions for which problems arise. More sophisticated knot placement strategies may help. For example, Luo & Wahba (1997) propose more sophisticated basis function reduction methods that could be adapted to the current context.

2.3 Comparison with P-Splines

The closest relatives of O'Sullivan penalised splines are the P-splines of Eilers & Marx (1996). If the interior knots $\kappa_5, \ldots, \kappa_{K+4}$ are taken to be equally-spaced then the family of cubic P-splines is given by (2.2) with the $\boldsymbol{\Omega}$ replaced by $\mathbf{D}_k^T\mathbf{D}_k$, where $\mathbf{D}_k$ is the kth-order differencing matrix. This differencing penalty corresponds to a discrete approximation to the integrated square of the kth derivative of the B-spline smoother. The choice $k = 2$ leads to the cubic P-spline estimate

$$\widehat{f}_P(x;\lambda) \equiv \mathbf{B}_x\widehat{\boldsymbol{\nu}}_P \text{ where } \widehat{\boldsymbol{\nu}}_P \equiv (\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}_k^T\mathbf{D}_k)^{-1}\mathbf{B}\mathbf{y} \tag{2.4}$$

having the property that $\widehat{f}_P(\cdot;\lambda)$ approaches the least squares line as $\lambda \to \infty$. In this sense, (2.4) is the closest relative of $\widehat{f}_P(\cdot;\lambda)$. If the interior knots are equally-spaced then the bands in the interior rows are, up to multiplicative factors, as follows:

O'Sullivan penalised splines (2.2):	3,	0,	−27,	48,	−27,	0,	3
Cubic P-splines; 2nd order diff. (2.4):	0,	8,	−32,	48,	−32,	8,	0

Figure 2.2 facilitates visual comparison of the two. It is seen that the differences are relatively small, although not negligible.

What are the relative advantages of smoothers based on cubic P-splines and O'Sullivan penalised splines, or O-splines for short? A theoretical comparison between P-splines and O-splines in terms of estimation performance, perhaps in the spirit of Hall & Opsomer (2005), would be ideal – although this is beyond the scope of the current chapter.

Eilers & Marx (1996) partially justify use of P-splines rather than O'Sullivan splines based on simplicity of the P-spline penalty matrix. However, as seen from (2.3), the penalty matrix needed for O-splines can be obtained straightforwardly. Furthermore the discrete approximation of P-splines requires equally-spaced knots which, depending on f, may not be desirable.

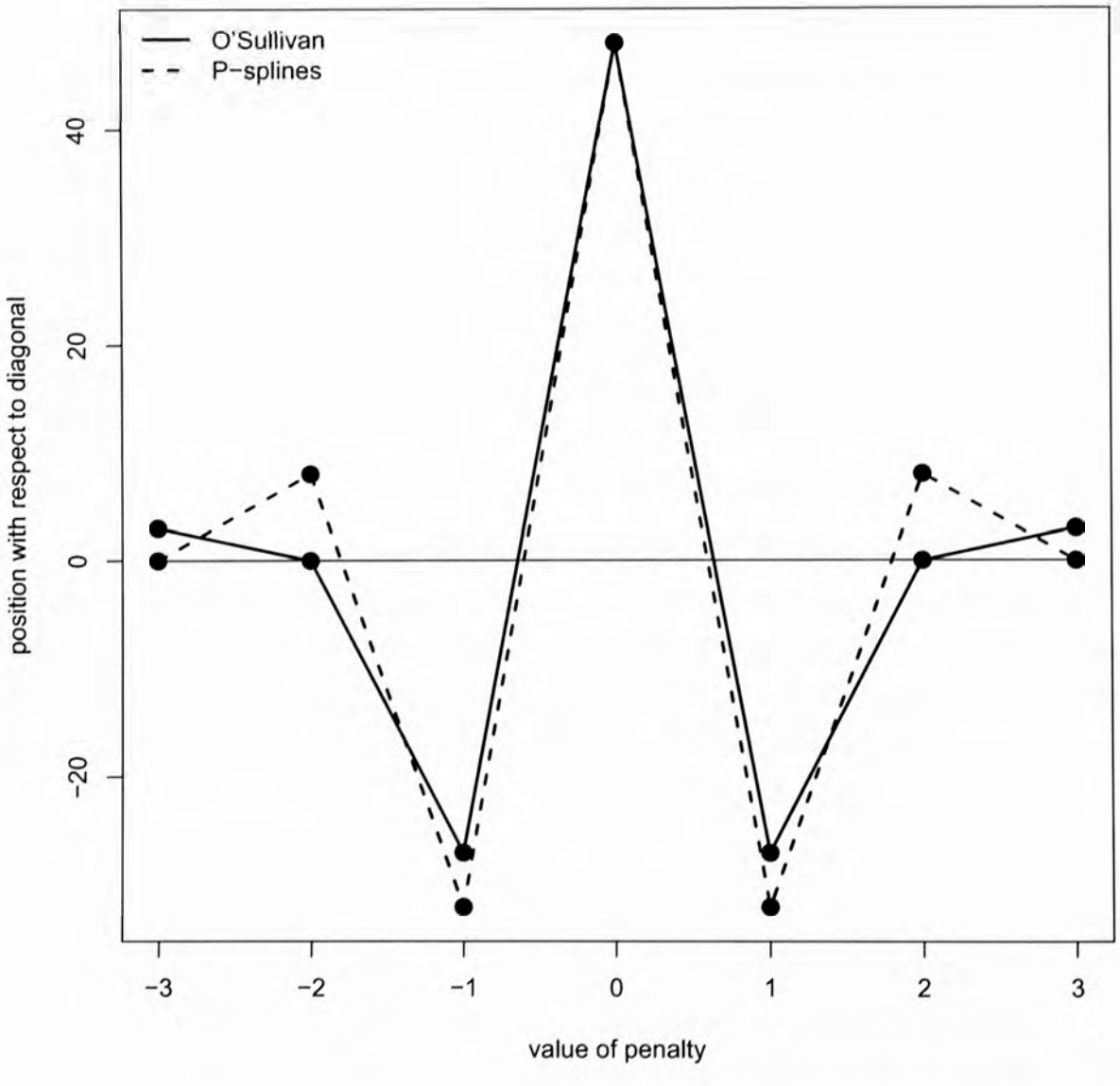


Figure 2.2: *Comparison of near-diagonal entries of the penalty matrices for O'Sullivan penalised splines and cubic P-splines with $k = 2$ and equally-spaced interior knots.*

A possible advantage of P-splines is the option of higher-order penalties, although the resulting smoothers can have erratic extrapolation behaviour. A possible advantage of O-splines is their direct relationship with time-honoured smoothing splines, and their attractive theoretical properties (e.g. Nussbaum, 1985). From the results described in Section 2.2 is clear that O-splines approach smoothing splines as $K \to n$. But how close are O-splines to smoothing splines for common (smaller) choices of K, and are they closer than P-splines with the same value of K and interior knots? To address these questions we conducted an empirical study based on the eighteen homoscedastic nonparametric regression settings in Wand (2000). For O-splines we used $K = 100$ equally spaced interior knots with 4 repeated knots at each boundary as described in Section 2.2. However, for P-splines we used the knot sequence described in the Appendix of Eilers & Marx (1996) which involves extending the knots beyond the boundary rather than repeating them. For each setting 200 samples were generated and smoothing spline estimates $\widehat{f}_S(\cdot\,;\lambda)$, with smoothing parameter chosen via generalised cross-validation, were obtained. We then computed $\widehat{f}_O(\cdot\,;\lambda)$ and $\widehat{f}_P(\cdot\,;\lambda)$ to have the same effective degrees of freedom as $\widehat{f}_S(\cdot\,;\lambda)$ and recorded closeness measures $d(\widehat{f}_O(\cdot\,;\lambda), \widehat{f}_S(\cdot\,;\lambda); A)$ and $d(\widehat{f}_P(\cdot\,;\lambda), \widehat{f}_S(\cdot\,;\lambda); A)$ where

$$d(f, g; A) \equiv \int_A (f - g)^2 dx$$

We took A corresponding to the intervals (a, κ_5) (left boundary), (κ_5, κ_{K+5}) (interior), (κ_{K+5}, b) (right boundary) and (a, b) (total region) where the κ_k denote the knots used for the O-spline fits. The Wand (2000) settings all involve predictor data within the unit interval. We took $(a, b) = (-0.1, 1.1)$ to assess behaviour beyond the range of the data. Wilcoxon tests on the 200 differences $d(\widehat{f}_O(\cdot\,;\lambda), \widehat{f}_S(\cdot\,;\lambda); A) - d(\widehat{f}_P(\cdot\,;\lambda), \widehat{f}_S(\cdot\,;\lambda); A)$ were carried out for each setting and choice of A. Apart from being distribution-free, Wilcoxon tests have the advantage of being invariant to normalisation and whether differences or ratios are used. In all 72 cases O-splines were closer to smoothing splines than P-splines in the sense that the Wilcoxon p-value < 0.01.

To appreciate the practical significance of these results we plotted the data and estimates at the 90th percentiles of each of the $d(\widehat{f}_O(\cdot\,;\lambda), \widehat{f}_S(\cdot\,;\lambda); A)$ and $d(\widehat{f}_P(\cdot\,;\lambda), \widehat{f}_S(\cdot\,;\lambda); A)$ samples, corresponding to relatively high discrepancies. Some examples are shown in Figure 2.3.

In each panel of Figure 2.3 all curve estimates in the interior are almost indistinguishable to the naked eye. However, big differences occur at the boundary. P-splines have a tendency to deviate from the natural boundary behaviour of smoothing splines. We also observed this phenomenon in the other 16 settings. Further study into this differing extrapolation behaviour would be worthwhile. We speculate that it comes from differences between the exact integral penalty and its discrete approximation near the boundary.

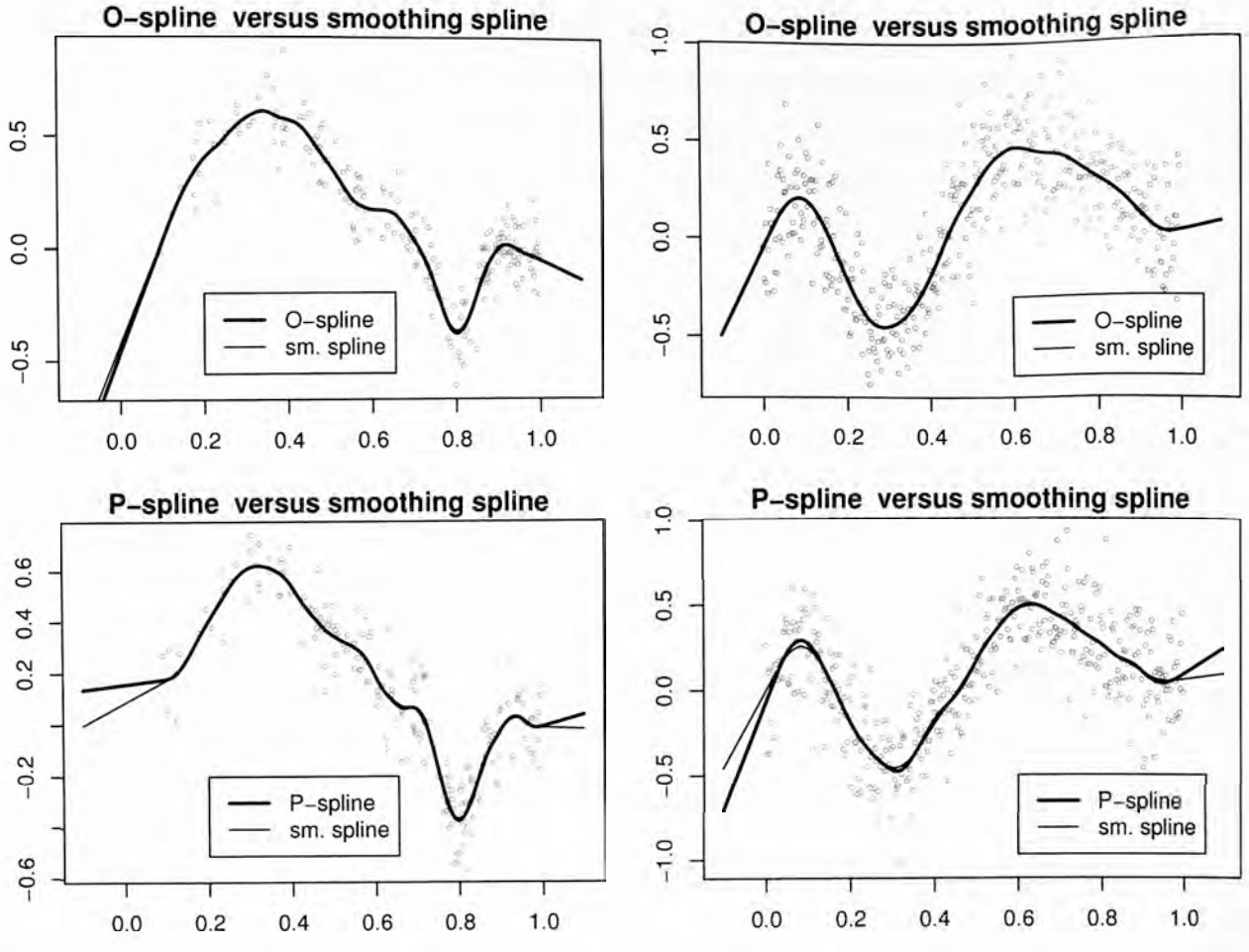


Figure 2.3: *O-spline and P-spline fits compared with smoothing spline fits corresponding to the 90th percentiles of the* $d(\widehat{f}_O, \widehat{f}_S; A)$ *and* $d(\widehat{f}_P, \widehat{f}_S; A)$ *samples; for two of the homoscedastic settings of Wand (2000).*

2.4 Mixed Model Formulation

There are several ways by which $\widehat{\boldsymbol{\nu}}_O$ in (2.2) can be expressed as a best linear unbiased predictor (BLUP) in a mixed model (e.g. Speed, 1990; Verbyla, 1994). However, from a software standpoint, the most convenient form is $\widehat{\boldsymbol{\nu}}_O = (\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}})$ where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ are (empirical) BLUPs of $\boldsymbol{\beta}$ and $\mathbf{u}$ in the mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \quad \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_\varepsilon^2 \mathbf{I} \end{bmatrix} \right) \tag{2.5}$$

for some design matrices $\mathbf{X}$ and $\mathbf{Z}$. An explicit expression for the BLUP in (2.5) (e.g. Ruppert *et al.*, 2003; Section 4.5.3) is

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{bmatrix} = \widehat{\boldsymbol{\nu}}_O = \left(\mathbf{C}^T\mathbf{C} + \lambda \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \right)^{-1} \mathbf{C}^T\mathbf{y}, \quad \lambda = \sigma_\varepsilon^2/\sigma_u^2 \tag{2.6}$$

where $\mathbf{C} = [\mathbf{X}, \mathbf{Z}]$, $\mathbf{I}$ is the identity matrix with the same number of columns as $\mathbf{Z}$. This "canonical" or standard mixed model form can be achieved if a $(K+4) \times (K+4)$ linear

transformation matrix $\mathbf{L}$ can be found such that $\mathbf{C} = \mathbf{BL}$ and

$$\mathbf{L}^T \boldsymbol{\Omega} \mathbf{L} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{2.7}$$

The usual method for obtaining $\mathbf{L}$ is spectral decomposition (e.g. Nychka & Cummins, 1996; Cantoni & Hastie, 2002; Welham *et al.*, 2007). It follows from results in the smoothing spline literature (e.g. Speed, 1991, Section 6) that

$$\text{rank}(\boldsymbol{\Omega}) = K + 2.$$

Hence, the spectral decomposition of $\boldsymbol{\Omega}$ is of the form $\boldsymbol{\Omega} = \mathbf{U}\text{diag}(\mathbf{d})\mathbf{U}^T$ where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{d}$ is a $(K+4) \times 1$ vector with exactly 2 zero entries and $K+2$ positive entries. Let $\mathbf{d}_Z$ be the $(K+2) \times 1$ sub-vector of $\mathbf{d}$ containing these positive entries, and let $\mathbf{U}_Z$ be the $(K+4) \times (K+2)$ sub-matrix of $\mathbf{U}$ with columns corresponding to positive entries of $\mathbf{d}$. Then an appropriate linear transformation is $\mathbf{L} = [\mathbf{U}_X, \mathbf{U}_Z\text{diag}(\mathbf{d}_Z^{-1/2})]$. This leads to the fixed and random effects design matrices:

$$\mathbf{X} = \mathbf{B}\mathbf{U}_X \text{ and } \mathbf{Z} = \mathbf{B}\mathbf{U}_Z\text{diag}(\mathbf{d}_Z^{-1/2}). \tag{2.8}$$

However, following again from the aforementioned smoothing spline literature (e.g. Speed, 1991, Section 6), $\mathbf{B}\mathbf{U}_X$ is a basis for the space of straight lines so the simpler specification $\mathbf{X} = [1, x_i]_{1 \le i \le n}$ may be used instead without affecting the fit. Figure 2.4 allows comparison of the original B-spline basis, corresponding to $\mathbf{B}$, and the basis corresponding to $\mathbf{Z}$. Notice the damping of the $\mathbf{Z}$ basis functions with increasing oscillation. This compensates for the fact that the penalty is a multiple of the identity matrix. In the Appendix it is shown how the `R` linear mixed model function `lme()` can be used to obtain $\widehat{f}_O(\cdot; \lambda)$ based on (2.5), with $\mathbf{Z}$ given by (2.8). For simple scatterplot smoothing there is little difference between this approach and direct use of `smooth.spline()`, and the answers are equivalent if the knot sequence and λ values are equal. The default choice of λ differs: `lme()` uses restricted maximum likelihood (REML) to choose λ, while `smooth.spline()` uses generalised cross-validation (GCV). The main advantage of the mixed model formulation of penalised splines is the incorporation into more complex models. Several examples are given in, for example, Ruppert *et al.* (2003). We will briefly describe one of them here.

2.4.1 Longitudinal Data

Figure 2.5 displays a longitudinal data set on bone mineral acquisition in young females (source: Bachrach, Hastie, Wang, Narasimhan & Marcus, 1999). The data consists of spinal bone mineral density (SBMD) measurements on each of 230 female subjects aged between 8 and 27. Each subject is measured between one and four times. Let n_i denote the number of measurements for subject i. The subjects have been divided into four ethnic groups: Asian, Black, Hispanic and White.

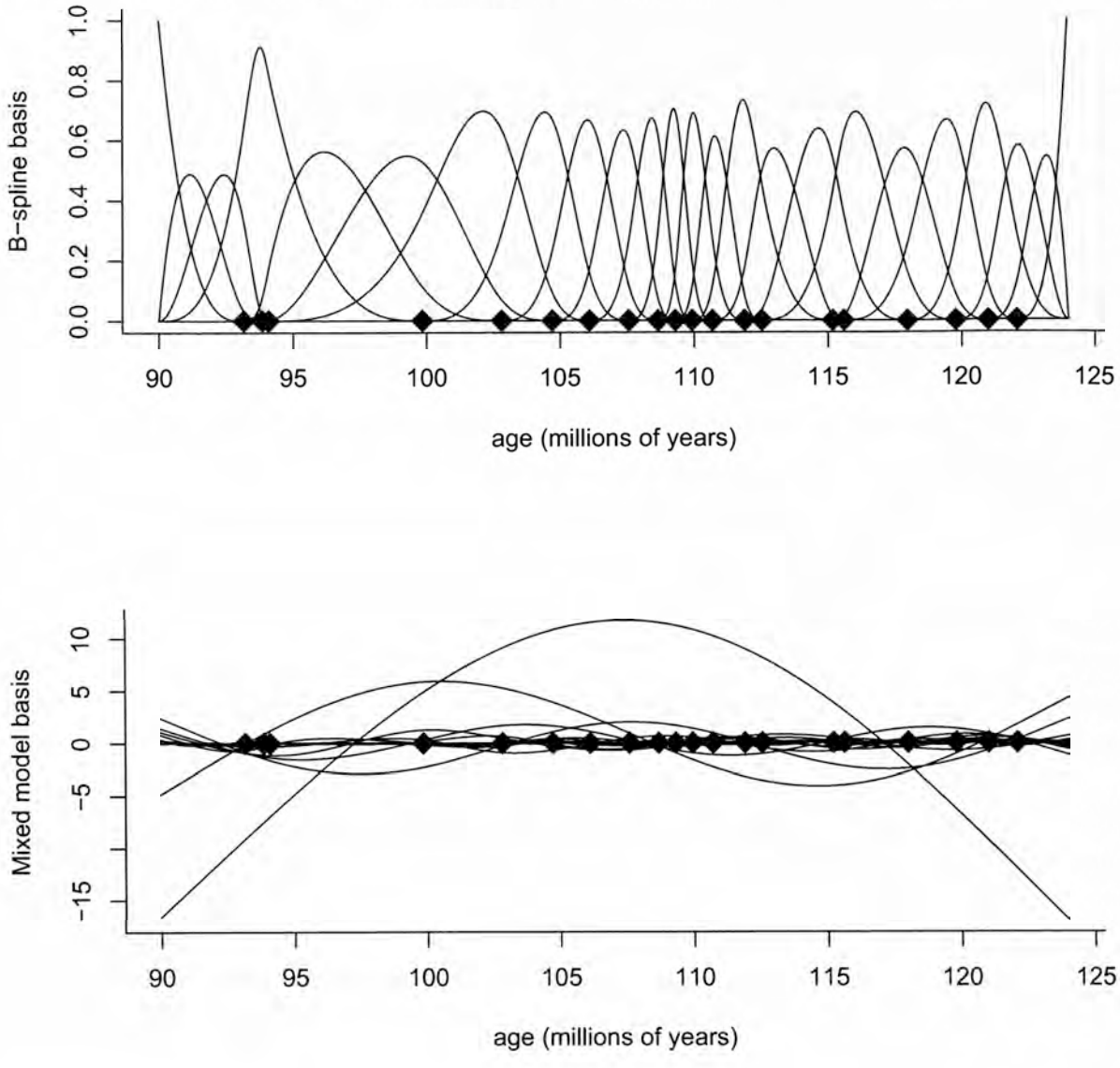


Figure 2.4: *Comparison of B-spline basis and* $\mathbf{Z}$ *basis for the fossil data example of Figure 2.2. The interior knots are shown as solid diamonds (♦).*

A useful additive mixed model for these data is:

$$\texttt{SBMD}_{ij} = U_i + f(\texttt{age}_{ij}) + \beta_1 \texttt{Black}_i + \beta_2 \texttt{Hispanic}_i + \beta_3 \texttt{White}_i + \varepsilon_{ij},\ 1 \leq i \leq 230, 1 \leq j \leq n_i$$

where are U_i i.i.d. $N(0, \sigma_u^2)$ random intercepts for each subject, $\texttt{Black}_i$, $\texttt{Hispanic}_i$ and $\texttt{White}_i$ are ethnicity indicators and ε_{ij} i.i.d. $N(0, \sigma_\varepsilon^2)$ are random errors. More sophisticated models that account for, say, serial correlation could be entertained. O'Sullivan penalised splines can be used to fit (2.5) with the design matrices set up as follows. Based on the $\texttt{age}_i$ values and appropriate knots, set up

$$\mathbf{Z}_{\text{spline}} = \mathbf{B}\mathbf{U}_Z \text{diag}(\mathbf{d}_Z^{-1/2})$$

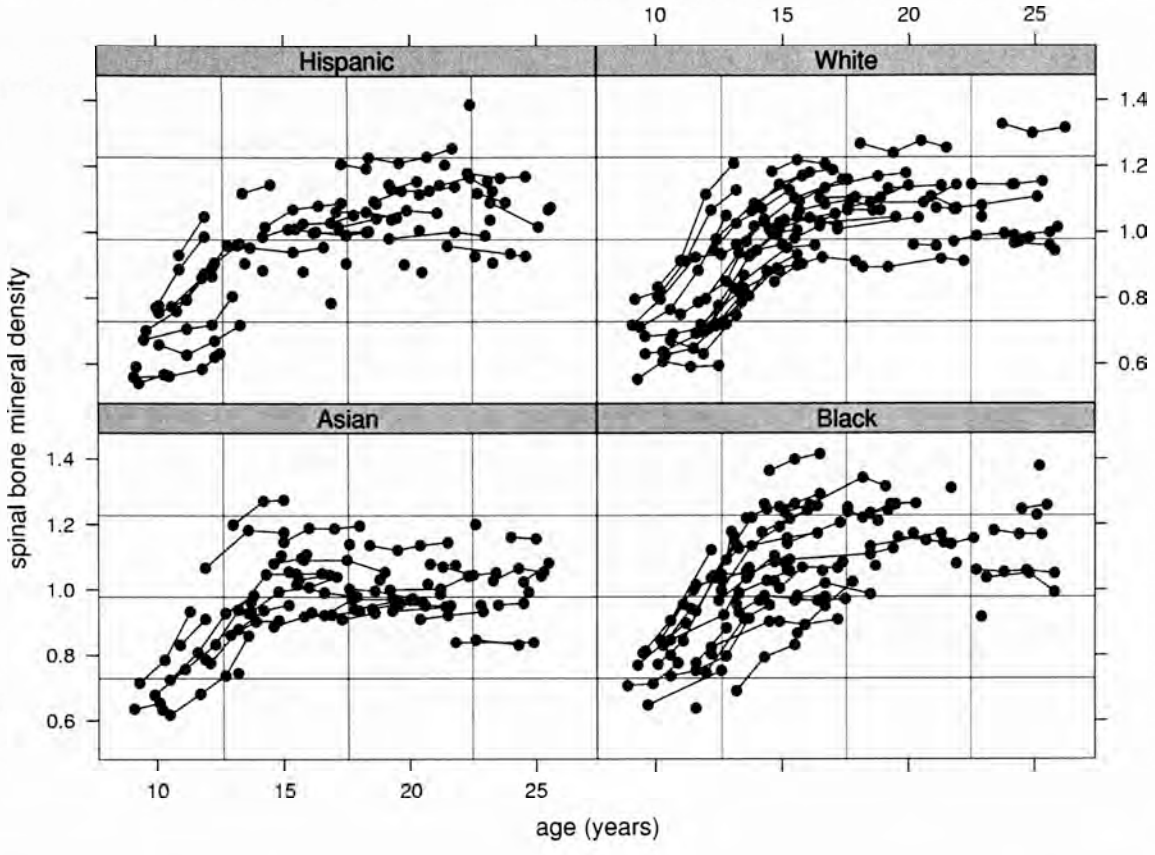


Figure 2.5: *The spinal bone mineral data. Lines connect measurements taken on the same subject.*

analogous to the $\mathbf{Z}$ matrix of (2.8) for simple scatterplot smoothing. In the Appendix of this chapter, when fitting data of this type, we use 15 interior knots corresponding to quantiles of the unique age values. Form

$$\mathbf{X} = \begin{bmatrix} 1 & \texttt{age}_{1,1} & \texttt{Black}_{1,1} & \texttt{Hispanic}_{1,1} & \texttt{White}_{1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \texttt{age}_{1,n_1} & \texttt{Black}_{1,n_1} & \texttt{Hispanic}_{1,n_1} & \texttt{White}_{1,n_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \texttt{age}_{230,1} & \texttt{Black}_{230,1} & \texttt{Hispanic}_{230,1} & \texttt{White}_{230,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \texttt{age}_{230,n_1} & \texttt{Black}_{230,n_1} & \texttt{Hispanic}_{230,n_1} & \texttt{White}_{230,n_1} \end{bmatrix}$$

and

$$\mathbf{Z}_{\text{subj}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Concatenate $\mathbf{Z}_{\text{subj}}$ and $\mathbf{Z}_{\text{spline}}$ to form

$$\mathbf{Z} = [\mathbf{Z}_{\text{subj}}, \mathbf{Z}_{\text{spline}}].$$

The appropriate mixed model is then

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

The Appendix of this chapter contains the `R` code for fitting this model. Note, in particular, that it circumvents explicit specification of $\mathbf{Z}_{\text{subj}}$. This is important for large longitudinal datasets.

2.5 Bayesian Analysis and Markov Chain Monte Carlo

A particularly attractive advantage of penalised splines, compared with smoothing splines, is the ease with which they can be fed into Markov Chain Monte Carlo (MCMC) schemes for fitting Bayesian semiparametric regression models – due to the reduction in the number of basis functions. For simple scatterplot smoothing this involves the Bayesian version of (2.5), namely

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2 \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_\varepsilon^2\mathbf{I}), \ \mathbf{u}|\sigma_u^2 \sim N(\mathbf{0}, \sigma_u^2\mathbf{I})$$

and suitable (usually diffuse) prior distributions for $\boldsymbol{\beta}$, σ_ε^2 and σ_u^2. However, the big advantages of a Bayesian/MCMC approach are realised when handling complications such as measurement error (e.g. Carroll, Ruppert, Stefanski & Crainiceanu, 2006) and generalised responses (e.g. Zhao, Staudenmayer, Coull & Wand, 2006), which are hindered by analytically intractable integrals in the likelihood.

Crainiceanu, Ruppert & Wand (2005) focus on use of the MCMC package `WinBUGS` (Windows version of `BUGS`, Spiegelhalter *et al.*, 2000) for Bayesian penalised spline models. They reported that the choice of basis functions can have a substantial impact on the convergence of the chain. We decided to conduct some convergence checks for MCMC fitting of the regression model

$$\text{logit}\left\{\mathbb{P}\left(\texttt{union}_i = 1|\texttt{wage}_i\right)\right\} = f(\texttt{wage}_i) \tag{2.9}$$

with f estimated via O'Sullivan penalised splines. Here $(\texttt{wage}_i, \texttt{union}_i)$, $1 \leq i \leq 534$, are pairs of wage amounts (dollars per hour) and trade union membership indicators for a sample of U.S. workers (source: Berndt, 1991). We expressed (2.9) as the Bayesian logistic mixed model:

$$\text{logit}\left\{\mathbb{P}\left(\texttt{union}_i = 1|\texttt{wage}_i\right)\right\} = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i, \ 1 \leq i \leq 534$$

where $\mathbf{X} = [1, \texttt{wage}_i]_{1 \leq i \leq 534}$ and $\mathbf{Z} = \mathbf{B}\mathbf{U}_Z \text{diag}(\mathbf{d}_Z^{-1/2})$, using the notation of Section 2.4. Given that there are 238 unique values for $\texttt{wage}_i$ a thorough analysis would vary the number of knots used. However, for reasons of computational efficiency, we only used 15 interior knots with quantile spacing.

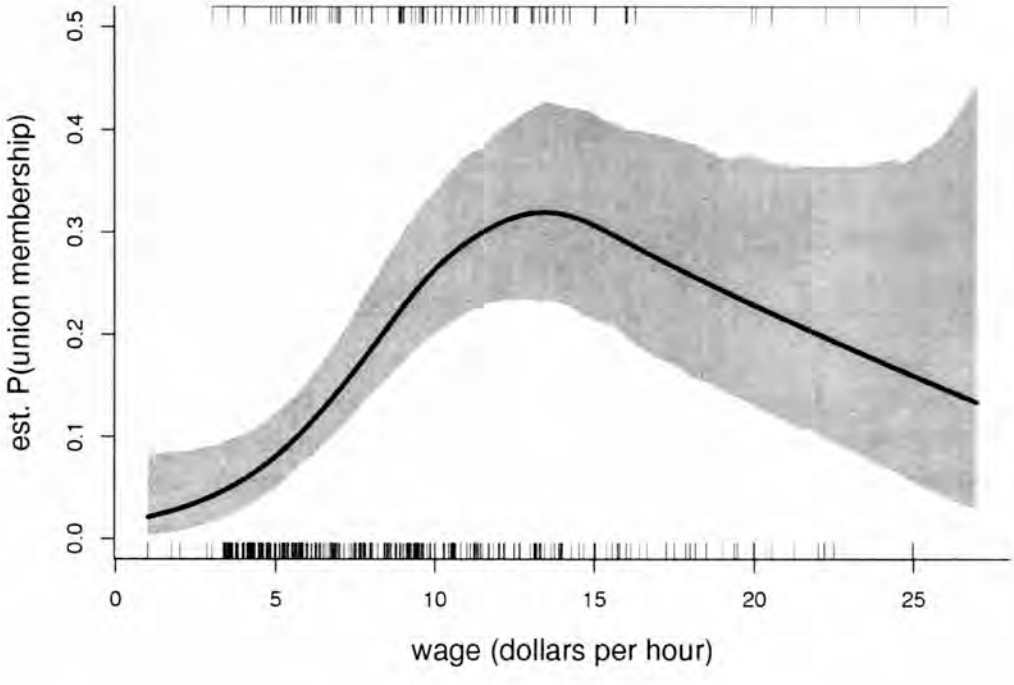


Figure 2.6: *Fit of (2.9) using O'Sullivan penalised splines. The values for each of the* $\texttt{wage}_i$*s have been jittered along the lines* $y = 1$ *and* $y = 0$ *corresponding to the value of* y_i.

Following the advice of Zhao *et al.* (2006) we used `WinBUGS` to generate chains of length 5,000 after a burn-in of 5,000 and applied a thinning factor of 5, resulting in posterior samples of size 1,000. Also in keeping with the recommendations of Zhao *et al.* (2006) we placed diffuse priors on the fixed effect parameters and variance component: β_0, β_1 independent $N(0, 10^8)$ and the prior density of σ_u^2 proportional to $(\sigma_u^2)^{-1.01} e^{-1/(100\sigma_u^2)}$, the inverse-gamma distribution with shape and rate parameter both 0.01, after scaling the predictor to have unit variance. Zhao *et al.* (2006) found that the results can be sensitive to the choice of the inverse-gamma hyperparameter.

The pointwise posterior mean effect of wage on the probability of trade union membership, together with 95% pointwise credible sets, is shown in Figure 2.6. Figure 2.7 allows assessment of convergence of the MCMC at each quartile of the wage sample and is seen to be excellent in each case.

We also compared the quality of mixing using the following logistic additive models involving 6 predictors and 3 smooth functions

$$\begin{aligned} \text{logit}\left\{\mathbb{P}\left(\texttt{union}_i = 1 | \texttt{wage}_i\right)\right\} &= \beta_0 + \beta_1 \texttt{female}_i + \beta_2 \texttt{Race}_i + \beta_3 \texttt{south}_i \\ &\quad + f_1(\texttt{wage}_i) + f_2(\texttt{age}_i) + f_3(\texttt{education}_i) \end{aligned} \tag{2.10}$$

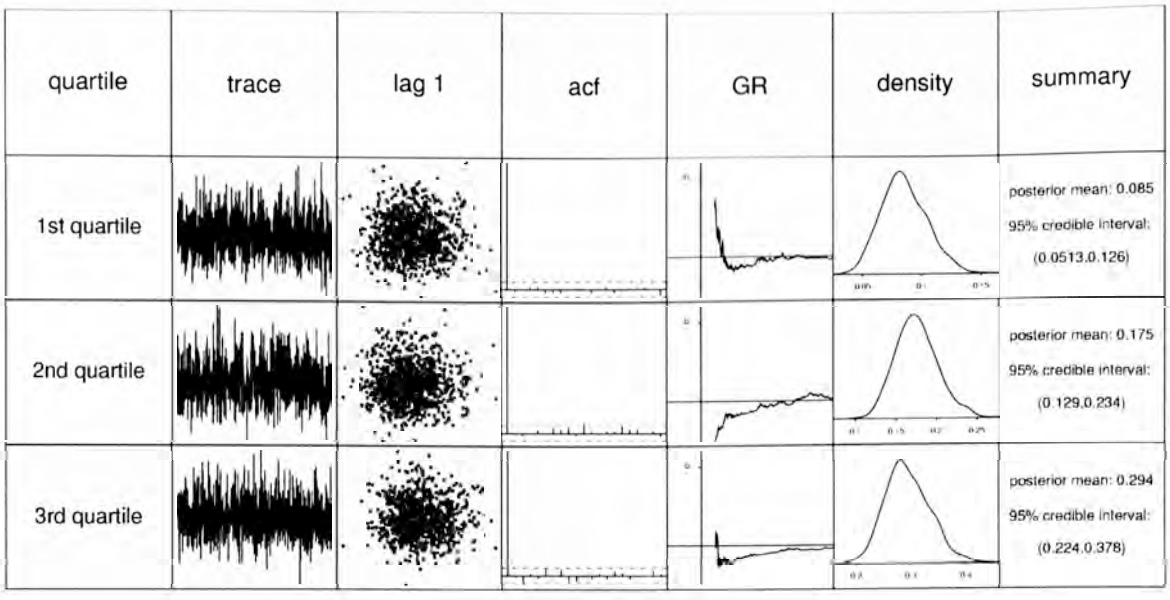


Figure 2.7: *Assessment of MCMC convergence for O'Sullivan penalised spline estimation of (2.9) at each quartile of wage. The columns are: quartile of wage, trace plot of sample of corresponding coefficient, plot of sample against 1-lagged sample, sample autocorrelation function, Gelman-Rubin $\sqrt{\widehat{R}}$ diagnostic, kernel estimates posterior density and basic numerical summaries.*

where f_1, f_2 and f_3 are modelled using truncated power splines, radial basis functions and O-splines respectively. The pointwise posterior mean effect of wage, age and education on the probability of trade union membership, together with 95% pointwise credible sets using O-splines, is shown in Figure 2.8. Figure 2.9 allows assessment of convergence for the median of wage for each basis type. From Figure 2.9 we see that mixing was good for O-splines and for the radial power basis but was not as good for truncated power splines.

Several examples of semiparametric regression with `WinBUGS`, including code, are given in Crainiceanu *et al.* (2005) and Zhao *et al.* (2006).

2.6 Extensions

In Section 2.2 an efficient method was described for calculation of the roughness penalty. We will now extend these results in several ways including to general degree O'Sullivan splines, derivative estimation and bivariate roughness penalties.

2.6.1 General Degree

Cubic O'Sullivan penalised splines have a natural extension to general odd degree splines. Higher degree splines have a role to play when smoother curve estimates are required. This arises, for example, in feature significance methodology (e.g. Chaudhuri & Marron, 1999; Hannig & Marron, 2006) where first and second derivatives of the fit are required. Return to the simple nonparametric regression setting (2.1) and let m be a general positive integer. Form the knot sequence

$$a = \kappa_1 = \ldots = \kappa_{2m} < \kappa_{2m+1} < \kappa_{2m+K} < \kappa_{2m+K+1} = \ldots = \kappa_{4m+K} = b$$

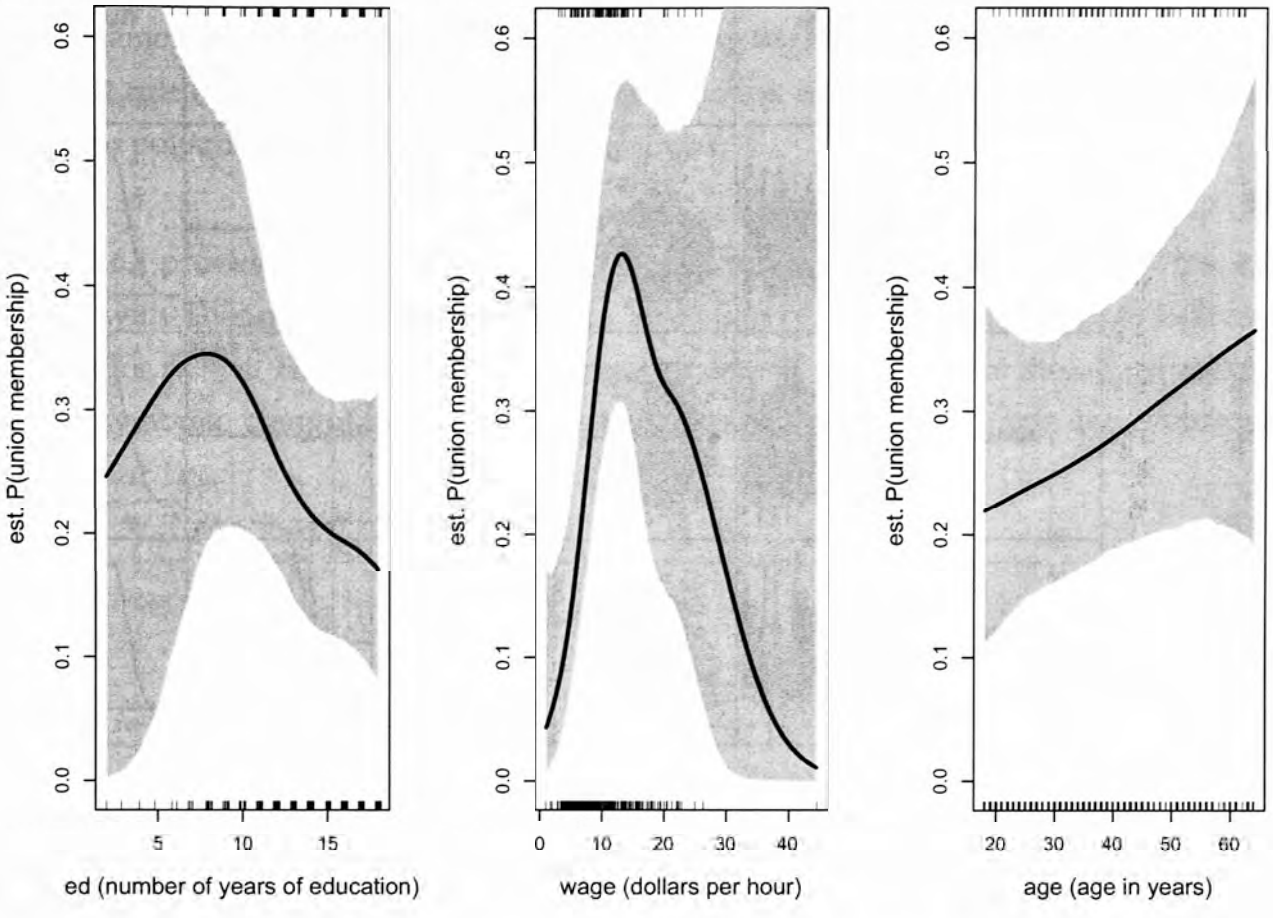


Figure 2.8: *Fit of (2.10) using O'Sullivan penalised splines.*

and let $B_{2m-1,1}, \ldots, B_{2m-1,K+2m}$ be the degree $(2m-1)$ B-spline basis defined by these knots. Order m O'Sullivan penalised splines then take the general form

$$\widehat{f}_O(x; m, \lambda) \equiv \mathbf{B}_{2m-1,x}\widehat{\boldsymbol{\nu}}_O \text{ where } \widehat{\boldsymbol{\nu}}_O \equiv \left(\mathbf{B}_{2m-1,x}^T\mathbf{B}_{2m-1,x} + \lambda\boldsymbol{\Omega}^{(m)}\right)^{-1}\mathbf{B}_{2m-1,x}^T\mathbf{y}.$$

Here $\mathbf{B}_{2m-1}$ is the $n \times (K+4m)$ design matrix with (i,k)th entry $B_{2m-1,k}(x_i)$, $\mathbf{B}_{2m-1,x} = [B_{2m-1,1}(x), \ldots, B_{2m-1,K+2m}(x)]$ and $\boldsymbol{\Omega}^{(m)}$ is the $(K+2m) \times (K+2m)$ penalty matrix with (k,k')th entry

$$\boldsymbol{\Omega}^{(m)}_{kk'} = \int_a^b B^{(m)}_{2m-1,k}(x)B^{(m)}_{2m-1,k'}(x)dx.$$

In the special case where the interior knots coincide with the x_is (assumed distinct), $\widehat{f}_O(\cdot; m, \lambda)$ corresponds to the order m smoothing spline; i.e. the minimiser of

$$\sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda\int_a^b f^{(m)}(x)^2dx$$

(e.g. Schoenberg, 1964).

We are now ready to state our result for exact computation of O'Sullivan spline penalty matrices.

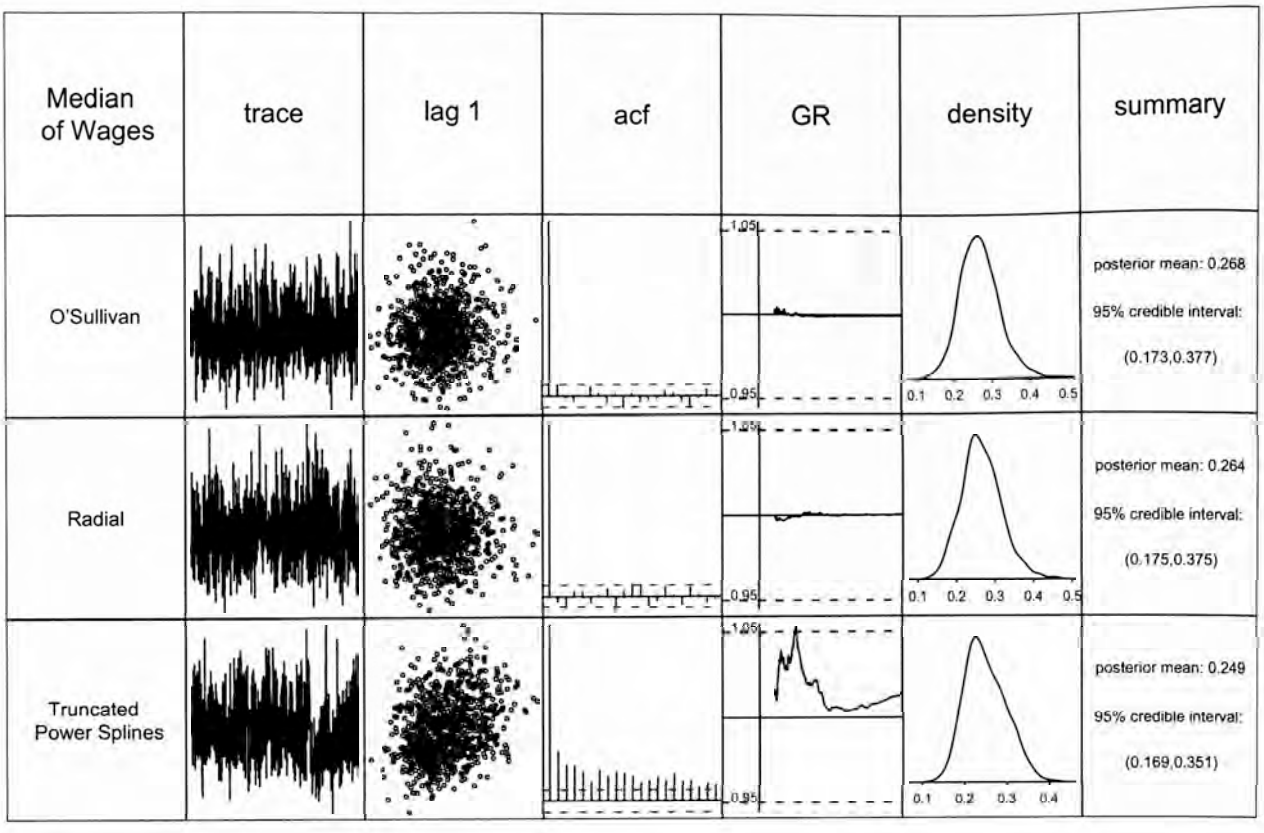


Figure 2.9: *Assessment of MCMC convergence for O'Sullivan, radial and truncated power spline fits of (2.10) for the median value of Wages. The columns are: predictor, trace plot of sample of corresponding coefficient, plot of sample against 1-lagged sample, sample autocorrelation function, Gelman-Rubin $\sqrt{\widehat{R}}$ diagnostic, kernel estimates posterior density and basic numerical summaries.*

Theorem 2.1: *The penalty matrix $\mathbf{\Omega}^{(m)}$ admits the exact explicit expression*

$$\mathbf{\Omega}^{(m)} = (\widetilde{\mathbf{B}}^{(m)})^T \mathrm{diag}(\mathbf{w}) \widetilde{\mathbf{B}}^{(m)}$$

where $\widetilde{\mathbf{B}}^{(m)}$ is the $(2m-1)(K+4m-1) \times (K+2m)$ matrix with (i,j)th entry $B^{(m)}_{2m-1,j}(\widetilde{x}_i)$ and $\mathbf{w}$ is a $(2m-1)(K+4m-1) \times 1$ vector with ith entry w_i. The $\widetilde{x}_i$ and w_i values are obtained according to

$$\widetilde{x}_{(2m-1)(\ell-1)+\ell'+1} = \kappa_\ell + \ell' h_{m,\ell},$$
$$w_{(2m-1)(\ell-1)+\ell'+1} = h_{m,\ell}\omega_{m,\ell'}$$

for $1 \le \ell \le K+4m-1$, $0 \le \ell' \le 2m-2$. Here, for $1 \le k \le K+2m$, $h_{1,k} = \kappa_{k+1} - \kappa_k$ and, for $m \ge 2$, $h_{m,k} = (\kappa_{k+1} - \kappa_k)/(2m-2)$. Lastly, for all $m \ge 1$,

$$\omega_{m,k} = \frac{(-1)^k}{k!(2m-2-k)!} \int_0^{2m-1} \frac{t(t-1)\ldots(t-2m+1)}{t-k} dt, \; k = 0, \ldots, 2m-2.$$

Proof. The (k,k')th entry of $\mathbf{\Omega}^{(m)}$ is

$$\mathbf{\Omega}^{(m)}_{kk'} = \int_a^b B^{(m)}_{2m-1,k}(x) B^{(m)}_{2m-1,k'}(x) dx = \sum_{i=1}^{K+4m-1} \int_{\kappa_i}^{\kappa_{i+1}} B^{(m)}_{2m-1,k}(x) B^{(m)}_{2m-1,k'}(x) dx. \quad (2.11)$$

Since $B^{(m)}_{2m-1,k}(x)B^{(m)}_{2m-1,k'}(x)$ are degree $m-1$ polynomials on each interval $x \in (\kappa_i, \kappa_{i+1})$ for $1 \leq i \leq K+4m-1$ the function $B^{(m)}_{2m-1,k}(x)B^{(m)}_{2m-1,k'}(x)$ is a degree $2(m-1)$ polynomial on the same interval. The result follows by applying the Newton-Cotes integration $(2m-1)$-point rule (e.g. Whittaker & Robinson, 1967) to the right hand side of (2.11) which is exact for polynomials of degree $2(m-1)$ or lower.

■

Table 2.6.1 provides values of $\omega_{m,k}$ for O'Sullivan polynomials up to degree 7. This, together with Theorem 2.1, allows direct computation of penalty matrices of O'Sullivan splines for $m \leq 4$. Higher values of m require a one-off calculation of the $\omega_{m,k}$ through, say, a symbolic computation package such as `Maple` (Waterloo Maple Inc., 2007) or `Mathematica`.

m/k	0	1	2	3	4	5	6
1	1						
2	1/3	4/3	1/3				
3	14/45	64/45	8/15	64/45	14/45		
4	41/140	54/35	27/140	68/35	27/140	54/35	41/140

Table 2.6.1: *Table of $\omega_{m,k}$ values for $m \leq 4$.*

Recall from Section 2.2 that, in the case of cubic O'Sullivan splines, Newton-Cotes integration reduces to Simpson's rule and a simpler, more revealing, expression results in the shape of (2.3).

A mixed model formulation for general degree penalties can be obtained using similar methods at those in Section 2.4. We seek a mixed model of the form (2.6) and again wish to find a $(K+2m) \times (K+2m)$ linear transformation matrix $\mathbf{L}$ can be found such that $\mathbf{C} = \mathbf{BL}$. To this end we first note that

$$\text{rank}(\boldsymbol{\Omega}^{(m)}) = K + m.$$

The spectral decomposition of $\boldsymbol{\Omega}^{(m)}$ is of the form $\boldsymbol{\Omega}^{(m)} = \mathbf{U}\text{diag}(\mathbf{d})\mathbf{U}^T$ where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{d}$ is a $(K+2m) \times 1$ vector with exactly m zero entries and $K+2m$ positive entries. Let $\mathbf{d}_Z$ be the $(K+2m) \times 1$ sub-vector of $\mathbf{d}$ containing these positive entries, and let $\mathbf{U}_Z$ be the $(K+2m) \times (K+m)$ sub-matrix of $\mathbf{U}$ with columns corresponding to positive entries of $\mathbf{d}$. Then an appropriate linear transformation is $\mathbf{L} = [\mathbf{U}_X, \mathbf{U}_Z\text{diag}(\mathbf{d}_Z^{-1/2})]$. This leads to the fixed and random effects design matrices with the same form as (2.8). However, we instead note that $\mathbf{BU}_X$ is a basis for the space of degree $m-1$ polynomials so the simpler specification uses

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

which may be used instead without affecting the fit.

2.6.2 *Derivative Plots*

A simple but perhaps underutilized tool in semiparametric regression are derivative plots. These are simple to use and can aid in the understanding of the fitted model. Suppose for a fixed λ the vector $\widehat{\boldsymbol{\nu}}_O$ was obtained from (2.2). Then an estimate of the derivatives of $\widehat{f}_O(x;\lambda)$ with respect to x are given by

$$\widehat{f}'_O(x;\lambda) \equiv \mathbf{B}'_x \widehat{\boldsymbol{\nu}}_O$$

where $\mathbf{B}'_x = \left[B'_1(x), \ldots, B'_{K+4}(x)\right]$ and the derivatives of B-splines with respect to x can be calculated recursively (de Boor, 1972)

$$B'_{p,k}(x) = \frac{p}{\kappa_{k+p} - \kappa_k} B_{p-1,k}(x) - \frac{p}{\kappa_{k+p+1} - \kappa_{k+1}} B_{p-1,k+1}(x).$$

If $\widehat{\boldsymbol{\nu}}_O$ was obtained using the mixed model formulation of O'Sullivan splines as in Section 2.4 or for general degree penalty using Section 2.6.1 then

$$\widehat{f}'_O(x;\lambda) \equiv \mathbf{C}'_x \widehat{\boldsymbol{\nu}}_O$$

where $\mathbf{C}'_x = [\mathbf{X}'_x, \mathbf{Z}'_x]$ and

$$\begin{aligned} \mathbf{X}'_x &= [0, 1, 2x, \ldots, (m-1)x^{m-2}] \\ \mathbf{Z}'_x &= \mathbf{B}'_x \mathbf{U}_Z \text{diag}(\mathbf{d}_Z^{-1/2}). \end{aligned}$$

This follows due to the fact that $\mathbf{U}_Z \text{diag}(\mathbf{d}_Z^{-1/2})$ is independent of x. Figure 2.10 illustrates the derivatives for the fit in Figure 2.1.

2.6.3 *Alternative Mixed Model Formulation*

The mixed model formulation of O'Sullivan splines as in Section 2.4 is not unique. For simplicity we will restrict our analysis to the cubic B-spline ($m = 2$) case. Using the properties of B-spline formula (1.16) we can write $B_1(x)$ and $B_2(x)$ as

$$\begin{aligned} B_1(x) &= c_{11} + c_{12}x + c_{13}x^2 + c_{14}x^3 + c_{15}(\kappa_5 - x)_+^3 \\ B_2(x) &= c_{21}x + c_{22}x^2 + c_{23}x^3 + c_{24}(\kappa_5 - x)_+^3 + c_{15}(\kappa_6 - x)_+^3 \end{aligned}$$

for some constants c_{ij}. If we remove these from the B-spline basis $\{B_k(x)\}_{k=1}^{K+4}$ then the space of straight lines is not included in the span of $\{B_k(x)\}_{k=3}^{K+4}$. Thus we let $\kappa_1, \ldots, \kappa_{K+6}$ be a knot sequence such that

$$a = \kappa_1 = \kappa_2 < \kappa_3 < \ldots < \kappa_{K+2} < \kappa_{K+3} = \kappa_{K+4} = \kappa_{K+5} = \kappa_{K+6} = b$$

and let $\{\widetilde{B}_k(\cdot)\}_{k=1}^{K+2}$ be the cubic B-spline basis functions defined by these knots. Adding the space of lines the basis

$$\{1, x\} \cup \{\widetilde{B}_k(x)\}_{k=1}^{K+2}$$

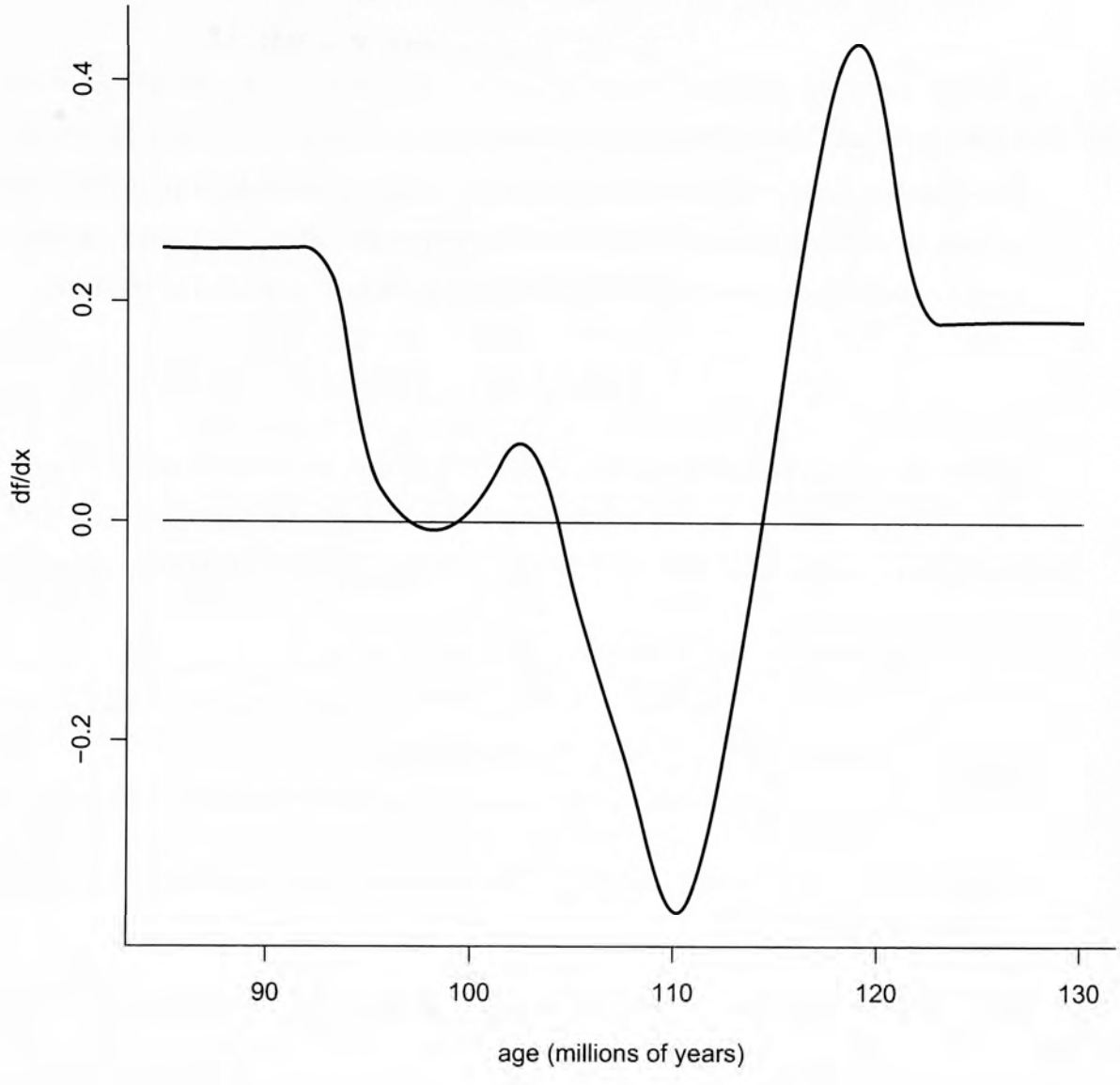


Figure 2.10: *Illustration of derivatives of O'Sullivan penalised spline fit of fossil data over the interval* $[85, 130]$ *millions of years.*

spans the same space of functions as the B-spline basis $\{B_k(x)\}_{k=1}^{K+4}$. We now set up the $n \times (K+2)$ design matrix $\mathbf{B}$ with (i,k)th entry $B_{ik} = \widetilde{B}_k(x_i)$ and $\boldsymbol{\Omega}$ is given by

$$\boldsymbol{\Omega} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{\Omega}} \end{bmatrix}$$

where $\widetilde{\boldsymbol{\Omega}}$ is a $(K+2) \times (K+2)$ penalty matrix for the reduced basis $\{\widetilde{B}_k(\cdot)\}_{k=1}^{K+2}$. We can calculate $\widetilde{\boldsymbol{\Omega}}$ using (2.3) with the reduced knot sequence. We now note that because the new basis $\{\widetilde{B}_k(\cdot)\}_{k=1}^{K+2}$ does not span the space of straight lines the corresponding penalty matrix $\widetilde{\boldsymbol{\Omega}}$ is full rank. Let

$$\widetilde{\boldsymbol{\Omega}} = \mathbf{R}^T\mathbf{R}$$

be the Cholesky factorization of $\widetilde{\mathbf{\Omega}}$ where $\mathbf{R}$ is an upper triangular matrix of the same dimensions as $\widetilde{\mathbf{\Omega}}$. Then an alternative mixed model formulation uses

$$\mathbf{X} = [1, x_i]_{1 \leq i \leq n} \text{ and } \mathbf{Z} = \mathbf{B}\mathbf{R}^{-1}.$$

2.6.4 *Bivariate Tensor Product O-Splines*

Thus far we have only considered at most additive models of univariate O-splines. Bivariate smoothing is also of considerable interest and can be performed by considering tensor products of univariate O-splines. We seek to fit a model of the form

$$y_i = f(x_{i1}, x_{i2}) + \varepsilon_i, \ 1 \leq i \leq n$$

where $\mathbf{x}_i = (x_{i1}, x_{i2})$ and $(y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^2$. Suppose that the estimate of f is required over $[a_1, b_1] \times [a_2, b_2]$ region containing the $\mathbf{x}_i$s. For an integer $K_1 \leq n$ and $K_2 \leq n$ let $\boldsymbol{\kappa}_1 = (\kappa_{1,1}, \ldots, \kappa_{K_1+8,1})$ and $\boldsymbol{\kappa}_2 = (\kappa_{1,2}, \ldots, \kappa_{K_2+8,2})$ be knot sequences such that

$$\begin{aligned}
a_1 = \kappa_{1,1} = \kappa_{2,1} = \kappa_{3,1} = \kappa_{4,1} < \kappa_{5,1} < \ldots \\
< \kappa_{K_1+4,1} < \kappa_{K_1+5,1} = \kappa_{K_1+6,1} = \kappa_{K_1+7,1} = \kappa_{K_1+8,1} = b_1 \\
a_2 = \kappa_{1,2} = \kappa_{2,2} = \kappa_{3,2} = \kappa_{4,2} < \kappa_{5,2} < \ldots \\
< \kappa_{K_2+4,2} < \kappa_{K_2+5,2} = \kappa_{K_2+6,2} = \kappa_{K_2+7,2} = \kappa_{K_2+8,2} = b_2
\end{aligned}$$

and let $\{B_{i1}(\cdot)\}_{i=1}^{K_1+4}$ and $\{B_{i2}(\cdot)\}_{i=1}^{K_2+4}$ be the cubic B-spline basis functions using $\boldsymbol{\kappa}_1$ and $\boldsymbol{\kappa}_2$ respectively. Let

$$f(x_1, x_2) = \sum_{i=1}^{K_1+4} \sum_{j=1}^{K_2+4} B_{i1}(x_1) B_{j2}(x_2) \nu_{ij}$$

Consider the problem of seeking for fixed λ the $\boldsymbol{\nu} = (\nu_{1,1}, \nu_{1,2}, \ldots, \nu_{K_1+4,K_2+4})$ which minimises

$$\sum_{i=1}^{n} (y_i - f(x_{i1}, x_{i2}))^2 + \lambda J_2(f)$$

where

$$\begin{aligned}
J_2(f) &= \int_{a_2}^{b_2} \int_{a_1}^{b_1} \left(\frac{\partial^2 f}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2}\right)^2 dx_1 dx_2 \\
&= \boldsymbol{\nu}^T \left(\mathbf{\Omega}^{(2,0)} + 2\mathbf{\Omega}^{(1,1)} + \mathbf{\Omega}^{(0,2)}\right) \boldsymbol{\nu} \\
&= \boldsymbol{\nu}^T \mathbf{\Omega} \boldsymbol{\nu}
\end{aligned}$$

with

$$\int_{a_2}^{b_2} \int_{a_1}^{b_1} \left(\frac{\partial^2 f}{\partial x_1^s \partial x_2^t}\right)^2 dx_1 dx_2 = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \sum_{i'=1}^{K_1} \sum_{j'=1}^{K_2} \nu_{ij} \nu_{i'j'} \mathbf{\Omega}^{(s,t)}_{ij,i'j'} = \boldsymbol{\nu}^T \mathbf{\Omega}^{(s,t)} \boldsymbol{\nu}. \quad (2.12)$$

and the pair (s, t) may take the values $(2, 0)$, $(1, 1)$ or $(0, 2)$. Also note that $J_2(f)$ is the same penalty determining bivariate thin-plate splines (e.g. Wahba, 1990; Wood 2003). We can use the following theorem to calculate $\mathbf{\Omega}^{(2,0)}$, $\mathbf{\Omega}^{(1,1)}$ and $\mathbf{\Omega}^{(0,2)}$ and hence $\mathbf{\Omega}$.

Theorem 2.2: *The penalty matrix* $\mathbf{\Omega}^{(s,t)}$ *admits the exact explicit expression*

$$\mathbf{\Omega}^{(s,t)} = \left[(\widetilde{\mathbf{B}}_1^{(s)})^T \mathrm{diag}(\mathbf{w}_1)\widetilde{\mathbf{B}}_1^{(s)}\right] \otimes \left[(\widetilde{\mathbf{B}}_2^{(t)})^T \mathrm{diag}(\mathbf{w}_2)\widetilde{\mathbf{B}}_2^{(t)}\right] \tag{2.13}$$

where $\widetilde{\mathbf{B}}_k^{(s)}$ *is the* $(6-2s)(K+7) \times (K+4)$ *matrix with* (i,j)*th entry* $B_{jk}^{(s)}(\widetilde{x}_{i,k})$ *and* $\mathbf{w}_k$ *is a vector of length* $(6-2s)(K+7)$ *with* i*th entry* w_{ik}*. The* $\widetilde{x}_{i,k}$ *and* $w_{i,k}$ *values are obtained according to*

$$\begin{aligned} \widetilde{x}_{(7-2s)(\ell-1)+\ell'+1,k} &= \kappa_\ell + \ell' h_{s,\ell,k}, \\ w_{(7-2s)(\ell-1)+\ell'+1,k} &= h_{s,\ell,k}\omega_{s,\ell'} \\ h_{s,\ell,k} &= (\kappa_{\ell+1,k} - \kappa_{\ell,k})/(6-2s) \end{aligned}$$

for $1 \leq \ell \leq K+7, 0 \leq \ell' \leq 6-2s, k = 1,2$ *and*

$$\omega_{s,\ell'} = \frac{(-1)^{\ell'}}{(\ell')!(6-2s-\ell')!}\int_0^{7-2s} \frac{t(t-1)\ldots(t+2s-7)}{t-\ell'}dt, \ \ell' = 0,\ldots,6-2s.$$

Proof. From equation (2.12) we can deduce

$$\begin{aligned} \mathbf{\Omega}_{ij,i'j'}^{(s,t)} &= \int_{a_2}^{b_2}\int_{a_1}^{b_1} B_{i1}^{(s)}(x_1)B_{j2}^{(t)}(x_2)B_{i'1}^{(s)}(x_1)B_{j'2}^{(t)}(x_2)dx_1dx_2 \\ &= \left(\int_{a_1}^{b_1} B_{i1}^{(s)}(x_1)B_{i'1}^{(s)}(x_1)dx_1\right)\left(\int_{a_2}^{b_2} B_{j2}^{(t)}(x_2)B_{j'2}^{(t)}(x_2)dx_2\right). \end{aligned}$$

Now

$$\int_{a_1}^{b_1} B_{i1}^{(s)}(x_1)B_{i'1}^{(s)}(x_1)dx_1 = \sum_{\ell=1}^{K+7}\int_{\kappa_{\ell,1}}^{\kappa_{\ell+1,1}} B_{i1}^{(s)}(x_1)B_{i'1}^{(s)}(x_1)dx_1. \tag{2.14}$$

Since $B_{i1}(x_1)$ and $B_{i'1}(x_1)$ are cubic B-splines for all (i,i'), the $B_{i1}(x_1)B_{i'1}(x_1)$ are degree 6 polynomials on each interval $x \in (\kappa_{\ell,k}, \kappa_{\ell+1,k})$ for $1 \leq \ell \leq K+7$ the function $B_{i1}^{(s)}(x_1)B_{i'1}^{(s)}(x_1)$ is a degree $6-2s$ polynomial on the same interval. The result follows by applying the Newton-Cotes integration $(7-2s)$-point rule (e.g. Whittaker & Robinson, 1967) to the right hand side of (2.14) which is exact for polynomials of degree $6-2s$ or lower. Similar arguments can be made for $B_{j2}^{(t)}(x_2)B_{j'2}^{(t)}(x_2)$. Using these we have

$$\mathbf{\Omega}_{ij,i'j'}^{(s,t)} = \left[(\widetilde{\mathbf{B}}_1^{(s)})^T \mathrm{diag}(\mathbf{w}_1)\widetilde{\mathbf{B}}_1^{(s)}\right]_{i,i'} \otimes \left[(\widetilde{\mathbf{B}}_2^{(t)})^T \mathrm{diag}(\mathbf{w}_2)\widetilde{\mathbf{B}}_2^{(t)}\right]_{j,j'}$$

which we can use to deduce the result.

■

The expression (2.13) bears a resemblance to the penalties used for the scale-invariant tensor product splines of Wood (2006) and could be adapted to this purpose with slight modifications.

A mixed model formulation satisfying (2.5) and (2.6) can be obtained by first noticing

$$\left\{f(x_1,x_2), f:\mathbb{R}^2 \to \mathbb{R} \text{ such that } J_2(f) = 0\right\} = \mathrm{span}\left\{1, x_1, x_2\right\}, \tag{2.15}$$

the space of bivariate planes. This implies that

$$\text{rank}(\boldsymbol{\Omega}) = (K+4)^2 - 3.$$

Hence, the spectral decomposition of $\boldsymbol{\Omega}$ is of the form $\boldsymbol{\Omega} = \mathbf{U}\text{diag}(\mathbf{d})\mathbf{U}^T$ where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{d}$ is a $(K+4)^2 \times 1$ vector with exactly 3 zero entries and $(K+4)^2 - 3$ positive entries. Let $\mathbf{d}_Z$ be the $((K+4)^2-3)\times 1$ sub-vector of $\mathbf{d}$ containing these positive entries, and let $\mathbf{U}_Z$ be the $(K+4)^2 \times ((K+4)^2 - 3)$ sub-matrix of $\mathbf{U}$ with columns corresponding to positive entries of $\mathbf{d}$. Then an appropriate linear transformation is $\mathbf{L} = [\mathbf{U}_X|\mathbf{U}_Z\text{diag}(\mathbf{d}_Z^{-1/2})]$. This leads to the fixed and random effects design matrices:

$$\mathbf{X} = \mathbf{B}\mathbf{U}_X \text{ and } \mathbf{Z} = \mathbf{B}\mathbf{U}_Z\text{diag}(\mathbf{d}_Z^{-1/2}).$$

However, $\mathbf{B}\mathbf{U}_X$ is a basis for the space of bivariate planes so the simpler specification $\mathbf{X} = [1, x_{i1}, x_{i2}]_{1\le i\le n}$ may be used instead without affecting the fit.

Figures 2.11–2.12 illustrates a fit and error for using bivariate tensor product O'Sullivan spline for $(y_i, \mathbf{x}_i)$, $1 \le i \le 400$, where $x_{i1} \sim$ Unif(0,1), $x_{i2} \sim$ Unif(0,1), $y_i \sim N(f(x_{i1}, x_{i2}), 0.1^2)$ and

$$\begin{aligned} f(x_1, x_2) \;=\; & \frac{0.75}{\pi\sigma_{x1}\sigma_{x2}} \exp\left\{-\frac{(x_1-0.2)^2}{\sigma_{x1}^2} - \frac{(x_2-0.2)^2}{\sigma_{x2}^2}\right\} \\ & + \frac{0.45}{\pi\sigma_{x1}\sigma_{x2}} \exp\left\{-\frac{(x-0.7)^2}{\sigma_{x1}^2} - \frac{(x_2-0.8)^2}{\sigma_{x2}^2}\right\} \end{aligned}$$

which is used in Wood (2003). Here we use $\sigma_{x1} = 0.3$ and $\sigma_{x2} = 0.4$. We fit the model using the `R` function `lme()` to fit the linear mixed model based on (2.5).

2.7 Closing Remarks

Smoothing splines have a special place in semiparametric regression. They are based on simple and intuitive principles, have an attractive theory (e.g. Nussbaum, 1985; Wahba, 1990; Eubank, 1994; Solo, 2000) and possess good practical properties such as natural boundary behaviour. Penalised splines, including P-splines, have gained popularity for reasons stated in the introduction. However, proponents of penalised splines have been viewed by some, especially in the smoothing spline community, as ignoring the benefits that have been established for smoothing splines over the past few decades. O'Sullivan penalised splines, being a direct generalisation and closer approximation of smoothing splines, provide an attractive link between the two streams of semiparametric regression research and allow analysts to enjoy the best of both worlds.

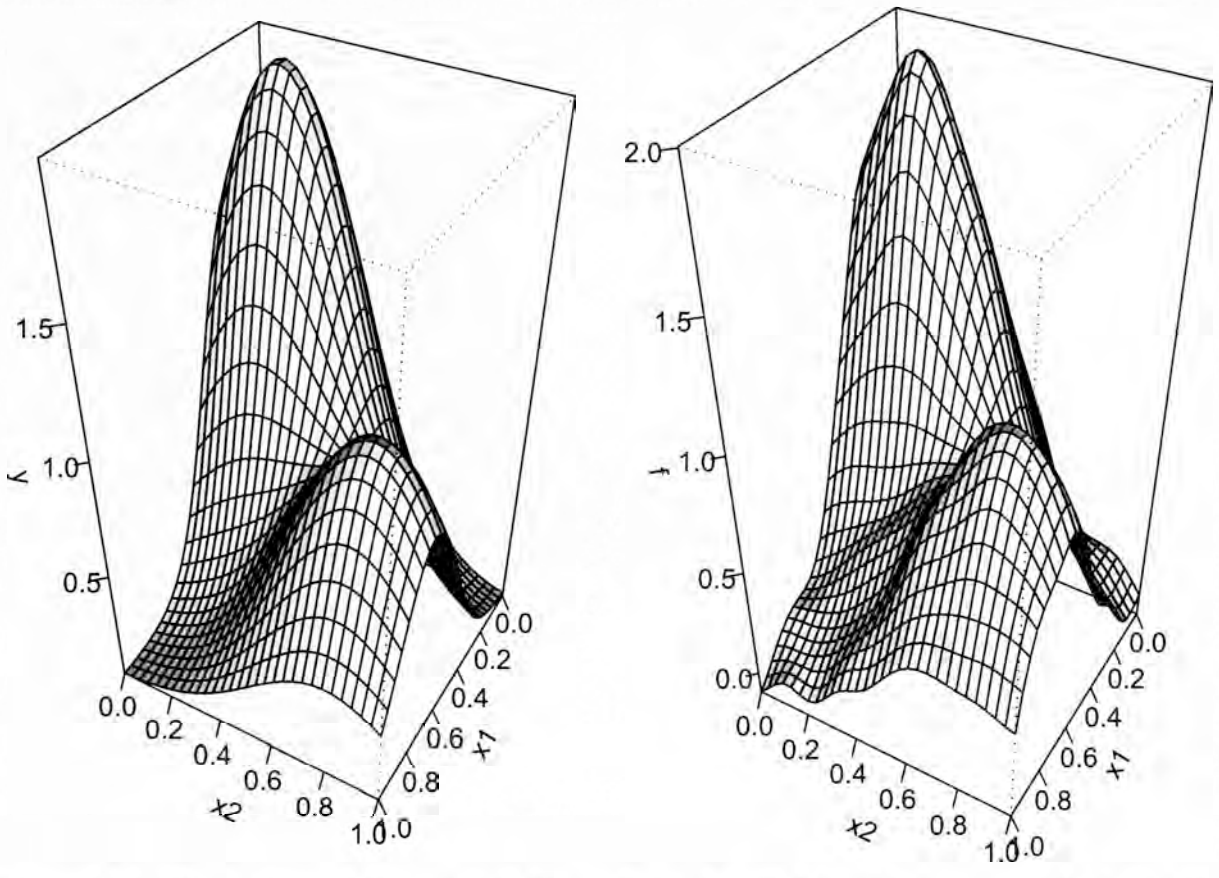


Figure 2.11: *Illustration of $f(x_1, x_2)$ (left panel) used to fit bivariate tensor product O-splines (right panel) for $(y_i, \mathbf{x}_i)$, $1 \leq i \leq 400$ where $x_{i1} \sim$ Unif(0,1), $x_{i2} \sim$ Unif(0,1) and $y_i \sim N(f(x_1, x_2), 0.1^2)$.*

Appendix: Code

In this Appendix we provide R code for use of O'Sullivan penalised splines in the simplest semiparametric regression setting: scatterplot smoothing. The extensions to more complex models, such as those described by Ngo & Wand (2004) and Crainiceanu, Ruppert & Wand (2005), is straightforward. We illustrate one of these extensions: additive mixed models.

Direct scatterplot smoothing with user choice of smoothing parameter

Obtain scatterplot data corresponding to environmental data from the R package `lattice`. Set up plotting grid, knots and smoothing parameter:

```
library(lattice) ; attach(environmental)
x <- radiation ; y <- ozone^(1/3)
a <- 0 ; b <- 350 ; xg <- seq(a,b,length=101)
numIntKnots <- 20 ; lambda <-  1000
```

Set up the design matrix and related quantities:

```
library(splines)
intKnots <- quantile(unique(x),seq(0,1,length=
        (numIntKnots+2))[-c(1,(numIntKnots+2))])
names(intKnots) <- NULL
```

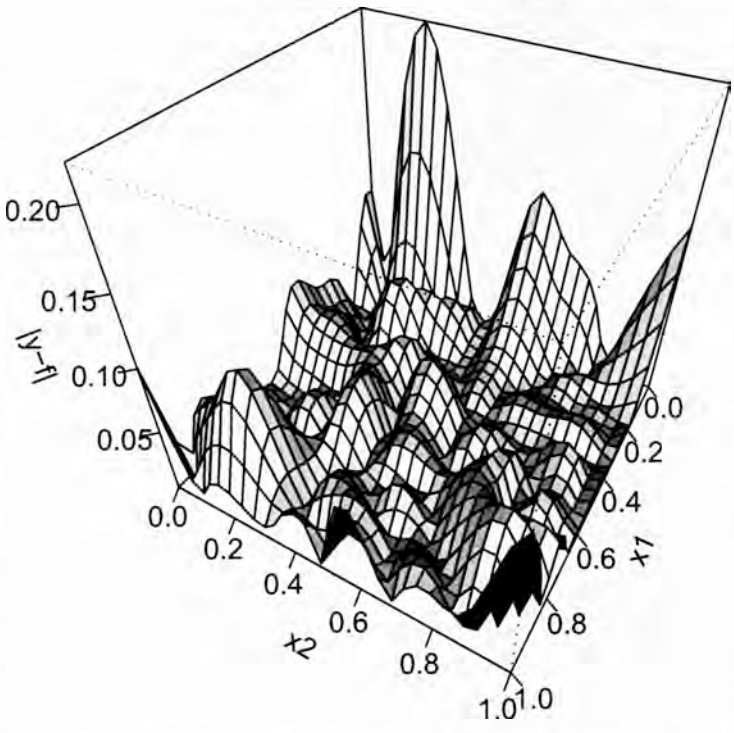


Figure 2.12: *Absolute error between $f(x_1, x_2)$ and bivariate tensor product O-splines fit.*

```
B <- bs(x,knots=intKnots,degree=3,
      Boundary.knots=c(a,b),intercept=TRUE)
BTB <- crossprod(B) ; BTy <- crossprod(B,y)
```

Create the $\boldsymbol{\Omega}$ matrix.

```
formOmega <- function(a,b,intKnots)
{
   allKnots <- c(rep(a,4),intKnots,rep(b,4))
   K <- length(intKnots) ; L <- 3*(K+8)
   xtilde <- (rep(allKnots,each=3)[-c(1,(L-1),L)]+
              rep(allKnots,each=3)[-c(1,2,L)])/2
   wts <- rep(diff(allKnots),each=3)*rep(c(1,4,1)/6,K+7)
   Bdd <- spline.des(allKnots,xtilde,derivs=rep(2,length(xtilde)),
                  outer.ok=TRUE)$design
   Omega     <- t(Bdd*wts)%*%Bdd
   return(Omega)
}
Omega <- formOmega(a,b,intKnots)
```

Obtain the coefficients:

```
nuHat <- solve(BTB+lambda*Omega,BTy)
```

For large K the following alternative Cholesky-based approach can be considerably faster ($O(K)$, because $\mathbf{B}^T\mathbf{B} + \lambda\boldsymbol{\Omega}$ is banded diagonal):

```
cholFac <- chol(BTB+lambda*Omega)
nuHat <- backsolve(cholFac,forwardsolve(t(cholFac),BTy))
```

Further improvements would be possible if the R functions `chol()`, `backsolve()` and `forwardsolve()` had a bandwidth argument which would exploit the banded diagonal structure of the various matrices above.

Display the fit:

```
Bg <- bs(xg,knots=intKnots,degree=3,
            Boundary.knots=c(a,b),intercept=TRUE)
fhatg <- Bg%*%nuHat
plot(x,y,xlim=range(xg),bty="l",type="n",xlab="radiation",
     ylab="cuberoot of ozone",main="(a) direct fit; user
     choice of smooth. par.")
lines(xg,fhatg,lwd=2)
points(x,y,lwd=2)
```

Mixed model scatterplot smoothing with REML choice of smoothing parameter

Obtain the spectral decomposition of Ω:

```
eigOmega <- eigen(Omega)
```

Obtain the matrix for linear transformation of $\mathbf{B}$ to $\mathbf{Z}$:

```
indsZ <- 1:(numIntKnots+2)
UZ <- eigOmega$vectors[,indsZ]
LZ <- t(t(UZ)/sqrt(eigOmega$values[indsZ]))
```

Perform stability check

```
indsX <- (numIntKnots+3):(numIntKnots+4)
UX <- eigOmega$vectors[,indsX]
L <- cbind( UX, LZ )
stabCheck <- t(crossprod(L,t(crossprod(L,Omega))))
if (sum(stabCheck^2) > 1.0001*(numIntKnots+2))
    print("WARNING: NUMERICAL INSTABILITY ARISING FROM
                    SPECTRAL DECOMPOSITION")
```

Form the $\mathbf{X}$ and $\mathbf{Z}$ matrices:

```
X <- cbind(rep(1,length(x)),x)
Z <- B%*%LZ
```

Fit using `lme()` with REML choice of smoothing parameter:

```
library(nlme)
group <- rep(1,length(x))
gpData <- groupedData(y~x|group,data=data.frame(x,y))
fit <- lme(y~-1+X,random=pdIdent(~-1+Z),data=gpData)
```

Extract coefficients and plot scatterplot smooth over a grid:

```
betaHat <- fit$coef$fixed
uHat <- unlist(fit$coef$random)
Zg <- Bg%*%LZ
fhatgREML <- betaHat[1] + betaHat[2]*xg + Zg%*%uHat
plot(x,y,xlim=range(xg),bty="l",type="n",xlab="radiation",
      ylab="cuberoot of ozone",main="(b) mixed model fit;
      REML choice of smooth. par.")
lines(xg,fhatgREML,lwd=2)
```

```
points(x,y,lwd=2)
```

Execution of the above code leads to Figure 2.13.

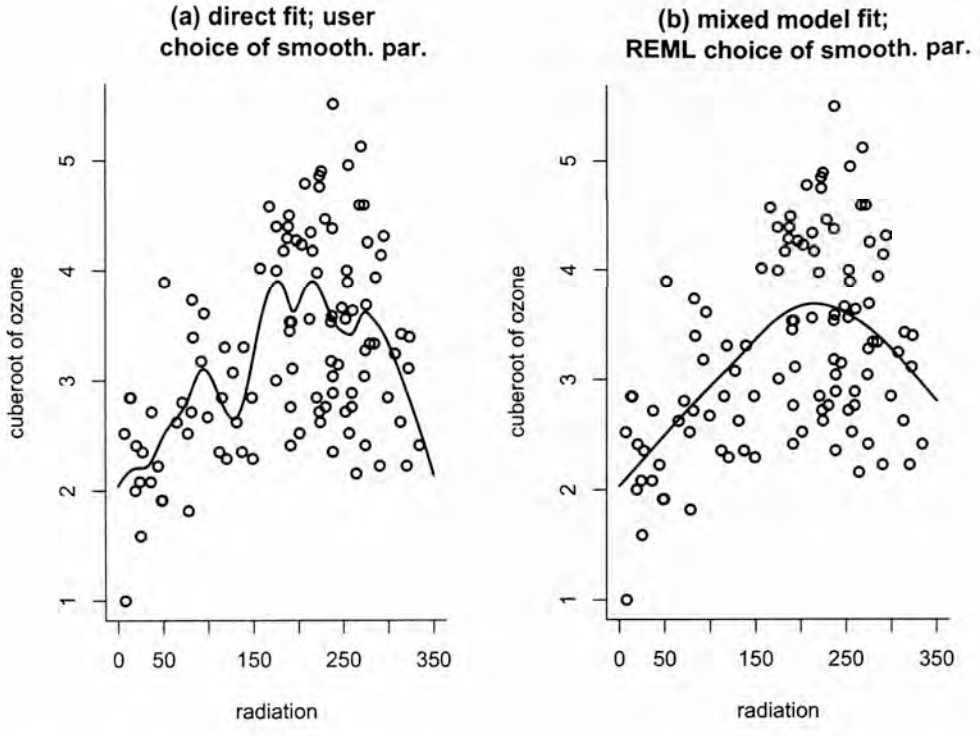


Figure 2.13: *Plots obtained from execution of the first two chunks of code in this Appendix.*

Fitting an additive mixed model

The spinal bone mineral density data of Bachrach *et al.* (1999) are not publicly available. Therefore we will illustrate fitting of additive mixed models using simulated data. For simplicity we will use two ethnicity categories rather than four.

Generate data and set up basic variables for the spline component:

```
set.seed(394600) ; m <- 230 ; nVals <- sample(1:4,m,replace=TRUE)
betaVal <- 0.1 ; sigU <- 0.25 ; sigEps <- 0.05
f <- function(x) { return(1 + pnorm((2*x-36)/5)/2) }
U <- rnorm(m,0,sigU)
age <- NULL ; ethnicity <- NULL
Uvals <- NULL ; idNum <- NULL
for (i in 1:m) {
   idNum <- c(idNum,rep(i,nVals[i]))
   stt <- runif(1,8,28-(nVals[i]-1))
   age <- c(age,seq(stt,by=1,length=nVals[i]))
   xCurr <- sample(c(0,1),1)
   ethnicity <- c(ethnicity,rep(xCurr,nVals[i]))
   Uvals <- c(Uvals,rep(U[i],nVals[i]))
}
epsVals <- rnorm(sum(nVals),0,sigEps)
SBMD <- f(age) + betaVal*ethnicity + Uvals + epsVals
```

Set up basic variables for the spline component.

```
a <- 8 ; b <- 28; numIntKnots <- 15
intKnots <- quantile(unique(age),seq(0,1,length=
                (numIntKnots+2))[-c(1,(numIntKnots+2))])
```

Obtain the spline component of the **Z** matrix.

```
B <- bs(age,knots=intKnots,degree=3,
        Boundary.knots=c(a,b),intercept=TRUE)
Omega <- formOmega(a,b,intKnots)
eigOmega <- eigen(Omega)
indsZ <- 1:(numIntKnots+2)
UZ <- eigOmega$vectors[,indsZ]
LZ <- t(t(UZ)/sqrt(eigOmega$values[indsZ]))
ZSpline <- B%*%LZ
```

Obtain the **X** matrix:

```
X <- cbind(rep(1,length(SBMD)),age,ethnicity)
```

Set up variables required for fitting via `lme()`. Note that the random intercept is taken care of via the tree identification numbers variable `idNum`, and that explicit formation of the random effect contribution to the **Z** matrix is not required.

```
groupVec <- factor(rep(1,length(SBMD)))
ZBlock <- list(list(groupVec=pdIdent(~ZSpline-1)),
                    list(idNum=pdIdent(~1)))
ZBlock <- unlist(ZBlock,recursive=FALSE)
dataFr <- groupedData(SBMD~ethnicity|groupVec,
                      data=data.frame(SBMD,X,ZSpline,idNum))
fit <- lme(SBMD~-1+X,data=dataFr,random=ZBlock)
betaHat <- fit$coef$fixed
uHat <- unlist(fit$coef$random)
uSplineHat <- uHat[1:ncol(ZSpline)]
```

Plot the data and fitted curve estimates together.

```
ng <- 101 ; ageg <- seq(a,b,length=ng)
Bg <- bs(ageg,knots=intKnots,degree=3,
         Boundary.knots=c(a,b),intercept=TRUE)
ZgSpline <- Bg%*%LZ
plotMatrix0 <- cbind(rep(1,ng),ageg,rep(0,ng),ZgSpline)
fhatgREML <- plotMatrix0 %*% c(betaHat, uSplineHat)
xLabs <- paste("ethnicity =",as.character(ethnicity))
pobj <- xyplot(SBMD~age|xLabs,groups=idNum,xlab="age (years)",
        ylab="spinal bone mineral density",subscripts=TRUE,
        panel=function(x,y,subscripts,groups)
        {
           panel.grid()  ; panel.superpose(x,y,subscripts,groups,
                                  type="b",col="grey60",pch=16)
           panelInd <- any(ethnicity[subscripts]==1)
           panel.xyplot(ageg,fhatgREML+panelInd*betaHat[3],
                        lwd=3,type="l",col="black")
        })
```

```
print(pobj)
```

Print approximate 95% confidence intervals for key parameters.

```
print(intervals(fit))
```

This leads to the following output:

```
Approximate 95% confidence intervals

Fixed effects:
                lower       est.      upper
X           0.68637207 0.77011154 0.85385101
Xage        0.02586448 0.02971670 0.03356891
Xethnicity 0.01121194 0.07549794 0.13978393
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: groupVec
                       lower       est.      upper
sd(ZSpline - 1) 0.01028272 0.01725978 0.02897093
  Level: idNum
          lower      est.     upper
sd(1) 0.2221770 0.2440963 0.2681781

 Within-group standard error:
     lower       est.      upper
0.04788011 0.05162773 0.05566867
```

Execution of the above code should lead to Figure 2.14.

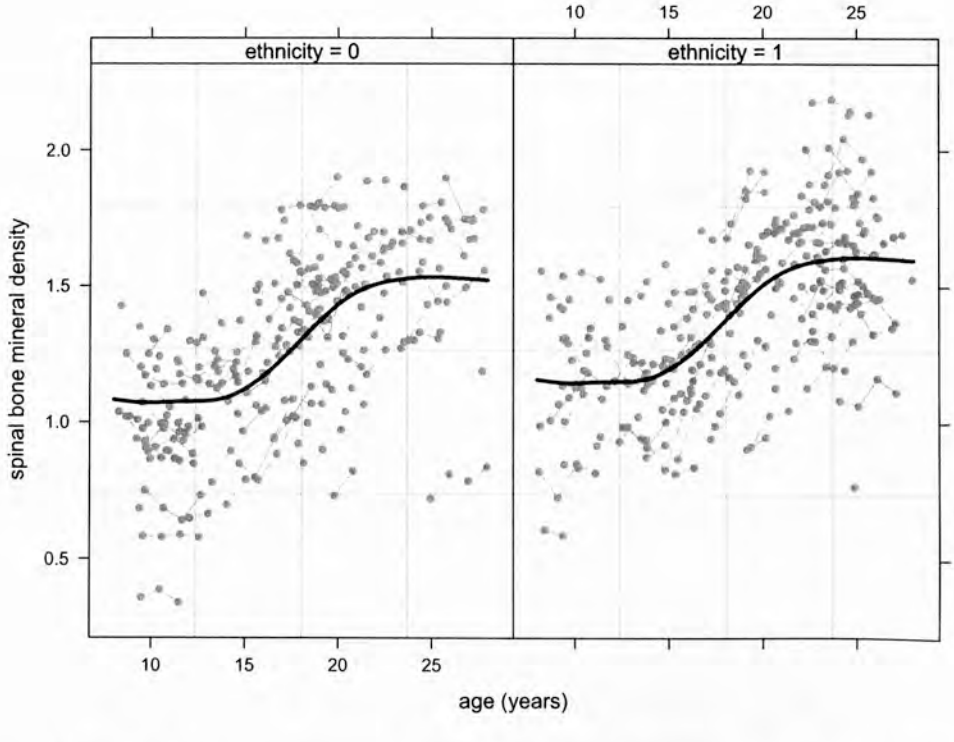


Figure 2.14: *Plot obtained from execution of the last chunk of code in this Appendix.*

CHAPTER 3

Parsimonious Classification via Generalised Linear Mixed Models [1]

3.1 Introduction

Classification is a very old and common problem, where training data are used to guide the classification of future objects into two or more classes based on observed predictors. Examples include clinical diagnosis based on patient symptoms, handwriting recognition based on digitised images and financial credit approval based on applicant attributes. Classification has an enormous number of applications; arising in most areas of science, but also in business as evidenced by the ongoing growth of industries such as data mining and fraud detection. The literature on classification methodology and theory is massive and mature. Contemporary statistical perspectives include Hastie, Tibshirani & Friedman (2001), Breiman (2001) and Hand (2006). A substantial portion of the classification literature is within the field of Computing Science, where 'classification' is usually called 'supervised learning' and 'predictors' often called 'features' or 'variables'.

There is a multitude of criteria that could be considered when tuning and assessing the quality of a classification algorithm. Numerical criteria include test error, Brier score and area under the curve of the receiver operating characteristic. A non-numerical quality criterion which, depending on the application, can be of utmost importance is *interpretability*. Hastie *et al.* (2001, Section 10.7) state that 'data mining applications generally require interpretable models' and that 'black box' classifiers with good numerical performance are 'far less useful'. Nevertheless, a good deal of classification theory and methodology, within both Statistics and Computing Science, is oblivious to interpretability. Some exceptions include tree-based approaches (e.g. Breiman, Friedman, Olshen & Stone, 1984; Hastie *et al.*, 2001) and additive model-based approaches (e.g. Hastie *et al.*, 2001). Related to interpretability is *parsimony*, where superfluous predictors are sifted out. This corresponds to pruning of tree-type classifiers and variable selection in those based on additive models. In Computing Science the topics of *variable selection* and *feature selection* (e.g. Guyon & Elisseeff, 2003) have similar aims.

Another often neglected quality measure is *speed*. Again, depending on the application, speed can be crucial. Speed is invariably tied to the size of the training data but

[1] This chapter corresponds to: Kauermann, G., Ormerod, J.T. & Wand, M.P. (2008), Parsimonious Classification via Generalised Linear Mixed Models. (submitted), representing joint research between G. Kauermann, J.T. Ormerod and M.P Wand.

there are huge differences, some involving several orders of magnitude, between existing classification algorithms in this respect.

In this chapter we develop a classification algorithm that strives for very good performance in terms of interpretation, parsimony and speed; while also achieving good classification performance. The algorithm, which we call KOW (after the authors of the corresponding paper), performs classification via a semiparametric logistic regression model after undergoing variable selection on the predictors. In this respect, KOW is similar in spirit to variable selection algorithms for additive models such as BRUTO (Hastie & Tibshirani, 1990), those based on versions of the `R` function `step.gam()` (Chambers & Hastie, 1992; Hastie, 2006; Wood, 2006), and Markov Chain Monte Carlo approaches such as that developed by Yau, Kohn & Wood (2002). The additive structure aids interpretation, but can also lead to improved test errors; see e.g. Section 12.3.4 of Hastie *et al.* (2001).

The KOW algorithm performs fast fitting and variable selection by borrowing ideas from generalised linear mixed models (GLMM). This is a relatively young, but rapidly growing, area of research that has its roots in biostatistical topics such as longitudinal data analysis and disease mapping; see e.g. Breslow & Clayton (1993), Verbeke & Molenberghs (2000) and Wakefield, Best & Waller (2000). However GLMMs can handle a much wider range of problems including generalised additive models (e.g. Zhao, Staudenmayer, Coull & Wand, 2006). The essence of KOW is to equate inclusion of a predictor with the significance of parameters in a GLMM. Linear terms correspond to fixed effect parameters, while non-linear terms correspond to variance components. KOW uses efficient score-based statistics, also known as *Rao statistics*, to choose among candidate predictors. A version of the Akaike Information Criterion is used to choose between fixed effect parameters and variance components, and also acts as a stopping rule. Unlike `step.gam()`, KOW has inbuilt automatic smoothing parameter selection for smooth function components.

When fitting a GLMM, whether for classification or not, the main obstacle is the presence of analytically intractable integrals in the likelihood. Currently available methods for fitting a GLMM fall into three general categories: quadrature, Monte Carlo methods and analytic approximation (e.g. McCulloch & Searle, 2001). Quadrature is not viable for the size of integrals arising in GLMMs with additive model structure. Monte Carlo methods are generally ruled out by their slowness. KOW makes use of a much faster Laplace-like approximation PQL (Breslow & Clayton, 1993). PQL approximations are sometimes criticised in GLMM analysis due to the substantial biases inherent in estimates of parameters of interest (e.g. McCulloch & Searle, 2000, p. 283). However, such issues are less crucial in the classification context.

We have tested KOW on several real and simulated data sets and compared it with other additive model-based classifiers. Our implementation of KOW fits a classifier to data sets with 5–10 possible predictors in a few seconds on a typical 2008 computer. If the number of predictors is in the tens then computation is in the order of minutes. The

penalised spline aspect of KOW means that training sample size only has a linear effect on computation times. KOW is generally much faster than `step.gam()`, although not as fast as BRUTO. However KOW can yield much better classification performance than BRUTO and is on par with `step.gam()`. Performances tend to be similar among algorithms in terms of interpretability and parsimony. On balance, we believe KOW has the potential for improved fast classification in contexts when interpretability and parsimony are important.

3.2 Fast Logistic Mixed Model Classifiers

Consider two-class classification with class labels denoted by $y \in \{0, 1\}$ and let $\mathbf{x} = (x_1, \ldots, x_d)$ be the set of possible predictors. Logistic regression-type classification is based on models of general form

$$\text{logit}\{P(y = 1|\mathbf{x})\} = \eta(\mathbf{x}). \tag{3.1}$$

Classification of a new observation with predictor vector $\mathbf{x}_{\text{new}}$ is performed according to

$$\text{sign}\{\widehat{\eta}(\mathbf{x}_{\text{new}})\}$$

where $\widehat{\eta}$ is an estimate of η based on training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. Here $\mathbf{x}_i$ is a d-variate vector representing the ith observation on $\mathbf{x}$.

A key element is appropriate modelling of $\eta(\mathbf{x})$. Given our interpretability goals, we work with sums of smooth low-dimensional functions, i.e. additive models (Hastie & Tibshirani, 1990) as described in Section 1.2.1. Models for $\boldsymbol{\eta} = (\eta(\mathbf{x}_1), \ldots, \eta(\mathbf{x}_n))$ can be written in the form

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \tag{3.2}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{u}$ is a vector of random effects, $\mathbf{X}$ contains a column of ones, together with a subset of the columns of $[\mathbf{x}_1, \ldots, \mathbf{x}_n]$, and $\mathbf{Z}$ are design matrices corresponding to spline bases. The covariance matrix of $\mathbf{u}$ takes the form

$$\mathbf{G}_{\boldsymbol{\sigma}^2} \equiv \underset{1 \le j \le v}{\text{blockdiag}}(\sigma_j^2 \mathbf{I}) \tag{3.3}$$

where $\boldsymbol{\sigma}^2 \equiv (\sigma_1^2, \ldots, \sigma_v^2)$ is the vector of variance components.

For the model defined by (3.1), (3.2) and (3.3) the log-likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ is

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) \;=\; \log \int \exp\left\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T \log(\mathbf{1} + e^{\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}})\right\} \\ \times (2\pi)^{-q/2} |\mathbf{G}_{\boldsymbol{\sigma}^2}|^{-1/2} \exp(-\tfrac{1}{2}\mathbf{u}^T \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1} \mathbf{u})\, d\mathbf{u} \end{aligned} \tag{3.4}$$

where q is the dimension of $\mathbf{u}$. The integral (3.4) cannot be calculated in analytic form. This is usually dealt with via Monte Carlo methods or analytic approximations. In the

interest of speed we work with the Laplace approximation of (3.4):

$$\begin{aligned}\ell_{\text{Laplace}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) &= -\tfrac{1}{2}\log|\mathbf{I} + \mathbf{Z}^T\mathbf{W}_{\boldsymbol{\beta},\hat{\mathbf{u}}}\mathbf{Z}\mathbf{G}_{\boldsymbol{\sigma}^2}| \\ &\quad + \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\hat{\mathbf{u}}) - \mathbf{1}^T\log(\mathbf{1} + e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\hat{\mathbf{u}}}) - \tfrac{1}{2}\hat{\mathbf{u}}^T\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\hat{\mathbf{u}}\end{aligned} \tag{3.5}$$

where

$$\mathbf{W}_{\boldsymbol{\beta},\mathbf{u}} \equiv \text{diag}\left\{\frac{e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}}}{(1+e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}})^2}\right\},$$

$\hat{\mathbf{u}}$ is the maximiser of the integrand in (3.5) (e.g. Breslow & Clayton, 1993) and $\boldsymbol{\sigma}^2 \geq \mathbf{0}$, i.e. satisfies

$$\begin{aligned}\mathbf{Z}^T\left(\mathbf{y} - \frac{e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\hat{\mathbf{u}}}}{1+e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\hat{\mathbf{u}}}}\right) - \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}\hat{\mathbf{u}} &= \mathbf{0} \\ \text{and } \boldsymbol{\sigma}^2 &\geq \mathbf{0}.\end{aligned} \tag{3.6}$$

Maximising (3.5) with respect to the remaining variables $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ is difficult due to non-linear expressions involving both $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ in the first and last terms of (3.5). We therefore pursue a backfitting idea by iteratively maximising (3.5) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$, respectively. Note that $\hat{\mathbf{u}}$ depends on $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$, so that the Laplace approximation has to be updated in each estimation iteration as well. We do this by updating the estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$ simultaneously. Let $\mathbf{B} \equiv \text{blockdiag}(\mathbf{0}, \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1})$, $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}, \mathbf{u})$, $\mathbf{C} \equiv [\mathbf{X}, \mathbf{Z}]$ and

$$df_j(\sigma_j^2) \equiv \text{tr}\{\mathbf{E}_j(\mathbf{Z}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{Z} + \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1})^{-1}\mathbf{Z}^T\mathbf{W}_{\hat{\boldsymbol{\beta}},\hat{\mathbf{u}}}\mathbf{Z}\}$$

where $\mathbf{E}_j$ is the diagonal matrix with ones in the diagonal positions corresponding to the spline basis functions for σ_j^2 and zeroes elsewhere. Note that $df_j(\sigma_j^2)$ has an 'effective degrees of freedom' (e.g. Buja, Hastie & Tibshirani, 1989) interpretation for the contribution from the spline terms attached to σ_j^2. We propose fitting logistic mixed model classifiers using Algorithm 1.

Algorithm 1 is similar to the algorithm developed by Breslow & Clayton (1993), commonly referred to as PQL (an acronym for Penalised Quasi-Likelihood) but differs in two respects. PQL uses Fisher scoring as the updating step for $\hat{\boldsymbol{\nu}}$ while Algorithm 1 uses a *repeated Hessian* Newton's method (see Appendix C). Here the Hessian is updated every second iteration and can be viewed as a slight modification of Fisher scoring. However, unlike PQL, the updating step for $\hat{\boldsymbol{\sigma}}^2$ uses a fixed point iteration in order to avoid calculating the Hessian matrix of derivatives with respect to $\boldsymbol{\sigma}^2$. The fixed point updating formula arises from differentiation of $\ell_{\text{Laplace}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ with respect to σ_j^2. The PQL approach to updating $\hat{\boldsymbol{\sigma}}^2$ is trickier to implement since more care is required to calculate the Hessian and ensuring positive definiteness in calculating Newton search directions for $\boldsymbol{\sigma}^2$.

Algorithm 1 is also quite fast compared to PQL. Solving for $\hat{\boldsymbol{\nu}}^{(s+1)}$ for a fixed $\hat{\boldsymbol{\sigma}}^2$ is a concave programming problem. Assuming that the function to be maximised has a Lipschitz continuous Hessian and the current iterate is sufficiently close to the solution then it is possible to show that the rate of convergence over two-steps of the algorithm is cubic (see Appendix C). Every odd iteration takes $O(nP^2 + P^3)$ while every even step only takes $O(nP + P^2)$ where P is the length of the $\hat{\boldsymbol{\nu}}$ vector. Solving for $\hat{\boldsymbol{\sigma}}^2$ can be

comprehended as a fixed-point iteration. Each σ^2 update can be computed in $O(nP^2 + P^3)$ operations.

Algorithm 1 Fast Fitting of a Logistic Mixed Model Classifier

1. **Initialise:** $\widehat{\boldsymbol{\nu}}^{(0)}$ and $\widehat{\boldsymbol{\sigma}}^{2(0)}$. Set L to be a small integer.
2. **Cycle:**

for $s = 1, 2, \ldots$ **do**
  if $s \mod L = 1$ **then**
    $\mathbf{K} = \mathbf{C}^T \mathbf{W}_{\widehat{\boldsymbol{\nu}}^{(s)}} \mathbf{C}$
  end if
  $\widehat{\boldsymbol{\nu}}^{(s+1)} = \widehat{\boldsymbol{\nu}}^{(s)} + (\mathbf{K} + \mathbf{B})^{-1} \left\{ \mathbf{C}^T \left(\mathbf{y} - \frac{e^{\mathbf{C}\widehat{\boldsymbol{\nu}}^{(s)}}}{1+e^{\mathbf{C}\widehat{\boldsymbol{\nu}}^{(s)}}} \right) - \mathbf{B}\widehat{\boldsymbol{\nu}}^{(s)} \right\}$
  for $t = 1, 2, \ldots$ **do**
    for $I \in \mathcal{I}$ **do**
      $\widehat{\sigma}_j^{2(t+1)} = \|\widehat{\mathbf{u}}_j^{(t)^T}\|^2 / df_j(\widehat{\sigma}_j^{2(t)})$
    end for
  end for
end for

until: $\max \left\{ \frac{\|\widehat{\boldsymbol{\nu}}^{(s+1)} - \widehat{\boldsymbol{\nu}}^{(s)}\|}{\|\widehat{\boldsymbol{\nu}}^{(s)}\|}, \frac{\|\widehat{\boldsymbol{\sigma}}^{2(t+1)} - \widehat{\boldsymbol{\sigma}}^{2(t)}\|}{\|\widehat{\boldsymbol{\sigma}}^{2(t)}\|} \right\}$ is below some small tolerance value.

3.3 Model Selection

We now address the problem of choosing between the various models for the classifier $\eta(\mathbf{x})$. Even for moderate d the number of such models can be very large. Our approach is driven by our previously stated goals of speed, parsimony and interpretability.

The fullest model has fixed effects component

$$\beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d.$$

However, smooth function terms will not be appropriate for all predictors. For example, some of the x_is may be binary. Let $\mathcal{S}$ be the subset of $\{1, \ldots, d\}$ such that x_i is to be modelled as smooth function for each $i \in \mathcal{I}$. Then let $\mathcal{S}$ be a partition of $\mathcal{I}$ that specifies the type of non-linear modelling in the fullest model. For example, if $d = 4$ and x_2 is binary then $\mathcal{S} = \{1, 3, 4\}$ corresponds to the fullest model being the additive model

$$\eta(x_1, x_2, x_3, x_4) = \beta_0 + s_1(x_1) + \beta_2 x_2 + s_3(x_3) + s_4(x_4),$$

while $\mathcal{S} = \{\{1, 3\}, 4\}$ corresponds to the model

$$\eta(x_1, x_2, x_3, x_4) = \beta_0 + s_{13}(x_1, x_3) + \beta_2 x_2 + s_4(x_4)$$

where s_1, s_3 and s_4 are smooth univariate functions of x_1, x_2 and x_4 respectively and s_{13} is a smooth bivariate function of x_1 and x_3. We will assume, for now, that $\mathcal{S}$ and $\mathcal{I}$ are

specified in advance. A recommended default choice is

$$\mathcal{S} = \text{all singleton sets of elements of } \mathcal{I}$$

corresponding to an additive model. Note that subscripting on the σ_j^2 corresponds to the elements of $\mathcal{I}$ rather than those of $\mathbf{x}$.

Description of our model selection strategy for the general set-up becomes notationally unwieldy. Therefore we will describe the algorithm via an example. Suppose that the set of possible predictors $\{x_1, x_2, x_3\}$ where x_1 is binary and x_2 and x_3 continuous, and that only additive models are to be considered. Then $\mathcal{S} = \{2, 3\}$ and $\mathcal{I} = \{2, 3\}$. The fullest model is

$$\eta(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sum_{k=1}^{q_2} u_{2k} Z_{2k}(x_2) + \sum_{k=1}^{q_3} u_{3k} Z_{3k}(x_3)$$

where u_{2k} i.i.d. $N(0, \sigma_1^2)$ and u_{3k} i.i.d. $N(0, \sigma_2^2)$. There are $2^5 = 32$ possible sub-models that include the intercept term. We propose the following forward selection approach to choosing among them:

1. Start with $\eta(x_1, x_2, x_3) = \beta_0$.
2. (a) Determine the 'best' linear component to add to the model from $\{\beta_1 x_1, \beta_2 x_2, \beta_3 x_3\}$. Let β_* denote the β_k corresponding to this choice.
   (b) Determine the 'best' non-linear (spline) component to add to the model from $\{\sum_{k=1}^{q_2} u_{2k} Z_{2k}(x_2), \sum_{k=1}^{q_3} u_{3k} Z_{3k}(x_3)\}$. Let σ_*^2 denote the σ_k^2 corresponding to this choice.
3. Add the component corresponding to β_* or σ_*^2 that leads to the bigger decrease in the marginal Akaike Information Criterion (mAIC). If there is no decrease or if there are no remaining components then stop and use the current model for classification. Otherwise, add the new component to the model and return to Step 2; modified to have one less component.

We propose to choose the 'best' linear and non-linear components using approximate score-type test statistics that do not require fitting of the candidate models. This has an obvious speed advantage. The details are given in Sections 3.3.1 and 3.3.2. The mAIC criterion is described in Section 3.3.3.

Before that we briefly remind the reader of some notation. For a general $d \times 1$ parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ with log-likelihood $\ell(\boldsymbol{\theta})$ the derivative vector of ℓ, $\mathsf{D}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$, is the $1 \times d$ with ith entry $\partial\ell(\boldsymbol{\theta})/\partial\theta_i$. The corresponding Hessian matrix is given by $\mathsf{H}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \mathsf{D}_{\boldsymbol{\theta}}\{\mathsf{D}_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})^T\}$. The information matrix of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is then $-\mathbb{E}\{\mathsf{H}_{\boldsymbol{\theta}}\ell(\widehat{\boldsymbol{\theta}})\}$.

3.3.1 *Choosing the 'best' linear component to add*

Let $(\boldsymbol{\beta}, \mathbf{u}, \sigma^2)$ define the current model, with fitted values $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}}, \widehat{\sigma}^2)$ as obtained via Algorithm 1, and let $\beta_k \mathbf{x}_k$ represent a generic linear component not already in the model. The log-likelihood corresponding to the new model with $\beta_k \mathbf{x}_k$ added is a modification of (3.4) with $\mathbf{X}\boldsymbol{\beta}$ replaced by $\mathbf{X}\boldsymbol{\beta} + \beta_k \mathbf{x}_k$ and is denoted by $\ell(\boldsymbol{\beta}, \sigma^2, \beta_k)$.

We propose to choose the 'best' $\beta_k \mathbf{x}_k$ among all candidates according to the maximum absolute *Rao statistic* (also known as the *score statistic*) (e.g. Rao, 1973, Chapter 6). Exact Rao statistics in GLMM are computationally expensive, so we make a number of convenient approximations. The first of these is to assume orthogonality between $(\boldsymbol{\beta}, \beta_k)$ and σ^2 in the information matrix of the joint parameters. Strictly speaking, these parameters are not orthogonal (Wand, 2007), but such orthogonality arises in the approximate log-likelihoods with which we work. Under orthogonality, the Rao statistic for the hypotheses $H_0 : \beta_k = 0$ versus $H_1 : \beta_k \neq 0$ is

$$R_{\beta_k} = [\mathsf{D}_{(\boldsymbol{\beta},\beta_k)}\ell(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, 0)]_{p+1} \Big/ \sqrt{1 \Big/ \left([\mathbb{E}_{\mathbf{y}}\{-\mathsf{H}_{(\boldsymbol{\beta},\beta_k)}\ell(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, 0)\}]^{-1} \right)_{p+1,p+1}}$$

where p is the length of $\boldsymbol{\beta}$. A practical approximation involves dropping the determinant term in (3.5) to obtain

$$\ell(\boldsymbol{\beta}, \sigma^2, \beta_k) \simeq \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{x}_k\beta_k + \mathbf{Z}\widehat{\mathbf{u}}) - \mathbf{1}^T \log(\mathbf{1} + e^{\mathbf{X}\boldsymbol{\beta}+\beta_k\mathbf{x}_k+\mathbf{Z}\widehat{\mathbf{u}}}) - \tfrac{1}{2}\widehat{\mathbf{u}}^T\mathbf{G}_{\sigma^2}^{-1}\widehat{\mathbf{u}}. \quad (3.7)$$

Vector calculus methods (e.g. Wand, 2002) applied to the right hand side of (3.7) lead to

$$\mathsf{D}_{(\boldsymbol{\beta},\beta_k)}\ell(\boldsymbol{\beta}, \sigma^2, \beta_k) \simeq \left(\mathbf{y} - \frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{x}_k\beta_k+\mathbf{Z}\widehat{\mathbf{u}}}}{\mathbf{1} + e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{x}_k\beta_k+\mathbf{Z}\widehat{\mathbf{u}}}} \right)^T [\mathbf{X}, \mathbf{x}_k].$$

Therefore the approximate numerator of R_{β_k} is the last entry of this vector with β_k set to zero:

$$[\mathsf{D}_{(\boldsymbol{\beta},\beta_k)}\ell(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, 0)]_{p+1} \simeq \mathbf{x}_k^T \left(\mathbf{y} - \frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{\mathbf{1} + e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}} \right).$$

The negative Hessian is approximately

$$-\mathsf{H}_{(\boldsymbol{\beta},\beta_k)}\ell(\boldsymbol{\beta}, \sigma^2, \beta_k) \simeq [\mathbf{X}, \mathbf{x}_k]^T \mathrm{diag} \left(\frac{e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{x}_k\beta_k+\mathbf{Z}\widehat{\mathbf{u}}}}{(\mathbf{1} + e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{x}_k\beta_k+\mathbf{Z}\widehat{\mathbf{u}}})^2} \right) [\mathbf{X}, \mathbf{x}_k].$$

The approximate denominator of R_{β_k} is the square root of the bottom right entry of this matrix with β_k set to zero and $\boldsymbol{\beta}$ set to its estimate at the current model. Standard results on the inverse of partitioned matrices lead to

$$R_{\beta_k} \simeq \mathbf{x}_k^T \left(\mathbf{y} - \frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{\mathbf{1} + e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}} \right) \Big/ \sqrt{\mathbf{x}_k^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}}\}\mathbf{x}_k}. \quad (3.8)$$

An advantage of this Rao statistic approach is that the candidate models corresponding to addition of the $\beta_k \mathbf{x}_k$ do not need to be fitted. This means that the R_{β_k} can be computed

quickly even when there is a large number of candidate linear components. This strategy has been used successfully in fitting regression spline models; see for example Stone, Hanson, Kooperberg & Truong (1997).

3.3.2 *Choosing the 'best' non-linear component to add*

As in Section 3.3.1, let $(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\sigma}^2)$ define the current model and let $\mathbf{Z}_k\mathbf{u}_k$, $\mathbf{u}_k \sim N(\mathbf{0}, \sigma_k^2\mathbf{I})$, represent a generic non-linear component not already in the model. The log-likelihood corresponding to the new model with σ_k^2 added is a modification of (3.4) with $\mathbf{Zu}$ replaced by $\mathbf{Zu} + \mathbf{Z}_k\mathbf{u}_k$ and is denoted by $\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_k^2)$.

The Rao statistic for $H_0 : \sigma_k^2 = 0$ versus $H_1 : \sigma_k^2 > 0$ is

$$\begin{aligned} R_{\sigma_k^2} &= [\mathsf{D}_{(\boldsymbol{\sigma}^2,\sigma_k^2)}\ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2, 0)]_{v+1} \Big/ \sqrt{\left[\mathbb{E}_{\mathbf{y}}\{-\mathsf{H}_{(\boldsymbol{\sigma}^2,\sigma_k^2)}\ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2, 0)\}\right]^{-1}_{v+1,v+1}} \\ &\equiv R^{\text{num}}_{\sigma_k^2} / R^{\text{den}}_{\sigma_k^2} \end{aligned} \tag{3.9}$$

where $R^{\text{num}}_{\sigma_k^2}$ and $R^{\text{den}}_{\sigma_k^2}$ respectively denote the numerator and denominator in $R_{\sigma_k^2}$ and r is the length of $\boldsymbol{\sigma}^2$. Test statistics of this type were studied by Cox & Koh (1989), Gray (1994), Lin (1997) and Zhang & Lin (2003), for example. We use the largest approximate $R_{\sigma_k^2}$ to choose the 'best' non-linear component not already in the model.

For practical reasons, we work with the Laplace approximation to $\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_k^2)$:

$$\begin{aligned} \ell_{\text{Laplace}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_k^2) &= -\frac{1}{2}\log|\mathbf{I} + [\mathbf{Z}, \mathbf{Z}_k]^T\mathbf{W}_{\boldsymbol{\beta},\hat{\mathbf{u}},\hat{\mathbf{u}}_k}[\mathbf{Z}, \mathbf{Z}_k]\text{blockdiag}(\mathbf{G}_{\boldsymbol{\sigma}^2}, \sigma_k^2\mathbf{I})| \\ &\quad + \mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\hat{\mathbf{u}} + \mathbf{Z}_k\hat{\mathbf{u}}_k) - \mathbf{1}^T\log(\mathbf{1} + e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\hat{\mathbf{u}}+\mathbf{Z}_k\hat{\mathbf{u}}_k}) \\ &\quad - \frac{1}{2}\hat{\mathbf{u}}^T\mathbf{G}^{-1}_{\boldsymbol{\sigma}^2}\hat{\mathbf{u}} - \frac{\|\hat{\mathbf{u}}_k\|^2}{2\sigma_k^2} \end{aligned} \tag{3.10}$$

where $(\hat{\mathbf{u}}, \hat{\mathbf{u}}_k)$ maximises

$$\mathbf{y}^T(\mathbf{Zu} + \mathbf{Z}_k\mathbf{u}_k) - \mathbf{1}^T\log(\mathbf{1} + e^{\mathbf{X}\boldsymbol{\beta}+\mathbf{Zu}+\mathbf{Zu}_k}) - \frac{1}{2}\mathbf{u}^T\mathbf{G}^{-1}_{\boldsymbol{\sigma}^2}\mathbf{u} - \frac{\|\hat{\mathbf{u}}_k\|^2}{2\sigma_k^2}. \tag{3.11}$$

The dependence of $\mathbf{W}_{\boldsymbol{\beta},\hat{\mathbf{u}},\hat{\mathbf{u}}_k}$ on $(\boldsymbol{\sigma}^2, \sigma_k^2)$ is ignored in the differentiation. Vector calculus methods (e.g. Wand, 2002) applied to the right hand side of (3.7) lead to

$$\begin{aligned} &[\mathsf{D}_{(\boldsymbol{\sigma}^2,\sigma_k^2)}\ell_{\text{Laplace}}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_k^2)]_j \\ &\quad = -\frac{1}{2}\text{tr}\left\{\mathbf{E}_j\left[\mathbf{I} + \widetilde{\mathbf{Z}}^T\mathbf{W}_{\boldsymbol{\beta},\hat{\mathbf{u}},\hat{\mathbf{u}}_k}\widetilde{\mathbf{Z}}\widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2}\right]^{-1}\widetilde{\mathbf{Z}}^T\mathbf{W}_{\boldsymbol{\beta},\hat{\mathbf{u}},\hat{\mathbf{u}}_k}\widetilde{\mathbf{Z}}\right\} + \frac{\|\hat{\mathbf{u}}_j\|^2}{2\sigma_j^2}. \end{aligned} \tag{3.12}$$

Noting that $(\hat{\mathbf{u}}, \hat{\mathbf{u}}_k)$ maximise (3.11) we get the relationships

$$\mathbf{G}_{\boldsymbol{\sigma}^2}\mathbf{Z}\left(\mathbf{y} - \frac{e^{\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{Z}\hat{\mathbf{u}}+\mathbf{Z}_k\hat{\mathbf{u}}_k}}{1 + e^{\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{Z}\hat{\mathbf{u}}+\mathbf{Z}_k\hat{\mathbf{u}}_k}}\right) = \hat{\mathbf{u}} \qquad \text{and} \qquad \sigma_k^2\mathbf{Z}_k\left(\mathbf{y} - \frac{e^{\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{Z}\hat{\mathbf{u}}+\mathbf{Z}_k\hat{\mathbf{u}}_k}}{1 + e^{\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{Z}\hat{\mathbf{u}}+\mathbf{Z}_k\hat{\mathbf{u}}_k}}\right) = \hat{\mathbf{u}}_k.$$

The second of these gives

$$\frac{\|\widehat{\mathbf{u}}_j\|^2}{\sigma_j^2} = \left\| \mathbf{Z}_j^T \left(\mathbf{y} - \frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{\mathbf{1}+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}} \right) \right\|^2 .$$

Substitution of this equation into (3.12) and setting $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2)$, $\sigma_k^2 = 0$, $\mathbf{u}_k = \mathbf{0}$ and $j = v+1$ then leads to

$$\begin{aligned}
&[\mathsf{D}_{(\boldsymbol{\sigma}^2,\sigma_k^2)} \ell_{\text{Laplace}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2, 0)]_{v+1} \\
&\quad \simeq -\frac{1}{2}\text{tr}\left\{ \mathbf{E}_{v+1} \left[\mathbf{I} + \widetilde{\mathbf{Z}}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \widetilde{\mathbf{Z}} \,\text{blockdiag}(\mathbf{G}_{\boldsymbol{\sigma}^2}, \mathbf{0}) \right]^{-1} \widetilde{\mathbf{Z}}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \widetilde{\mathbf{Z}} \right\} \\
&\qquad + \frac{1}{2} \left\| \mathbf{Z}_k^T \left(\mathbf{y} - \frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{\mathbf{1}+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}} \right) \right\|^2 .
\end{aligned}$$

Note that $\{\mathbf{I} + \widehat{\mathbf{Z}}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \widehat{\mathbf{Z}} \,\text{blockdiag}(\mathbf{G}_{\boldsymbol{\sigma}^2}, \mathbf{0})\}^{-1} \widehat{\mathbf{Z}}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \widetilde{\mathbf{Z}}$ has the explicit expression

$$\begin{bmatrix} \mathbf{I} + \mathbf{Z}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \mathbf{Z} \mathbf{G}_{\boldsymbol{\sigma}^2} & \mathbf{0} \\ \mathbf{Z}_k^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \mathbf{Z} \mathbf{G}_{\boldsymbol{\sigma}^2} & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Z}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \mathbf{Z} & \mathbf{Z}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \mathbf{Z}_k \\ \mathbf{Z}_k^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \mathbf{Z} & \mathbf{Z}_k^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \mathbf{Z}_k \end{bmatrix}.$$

Hence the expression

$$\begin{aligned}
R_{\sigma_k^2}^{\text{num}} \;\simeq\; & -\frac{1}{2}\text{tr}\left[\mathbf{Z}_k^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \{\mathbf{I} - \mathbf{Z}\mathbf{G}_{\boldsymbol{\sigma}^2}(\mathbf{I} + \mathbf{Z}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \mathbf{Z}\mathbf{G}_{\boldsymbol{\sigma}^2})^{-1} \mathbf{Z}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \} \mathbf{Z}_k \right] \\
& + \frac{1}{2} \left\| \mathbf{Z}_k^T \left(\mathbf{y} - \frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}}{\mathbf{1}+e^{\mathbf{X}\widehat{\boldsymbol{\beta}}+\mathbf{Z}\widehat{\mathbf{u}}}} \right) \right\|^2 .
\end{aligned} \tag{3.13}$$

then follows from standard results on the inverse of a partitioned matrix and some straightforward matrix algebra. Expression (3.13) has the computational advantage that the matrix inversion pertains to the current model and only needs to be performed once for selecting the 'best' non-linear component.

We now provide a computationally efficient expression for $R_{\sigma_k^2}^{\text{den}}$. Let

$$\mathcal{K}(\boldsymbol{\sigma}^2, \sigma_k^2) = \mathbb{E}_{\mathbf{y}}\{-\mathsf{H}_{(\boldsymbol{\sigma}^2,\sigma_k^2)} \ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_k^2)\}.$$

This can be approximated using the arguments in Section 2.4 of Breslow & Clayton (1993) leading to

$$\begin{aligned}
\mathcal{K}_{ij}(\boldsymbol{\sigma}^2, \sigma_k^2) \;\equiv\; & \frac{1}{2}\text{tr}\Big\{ \mathbf{E}_i (\mathbf{I} + \widetilde{\mathbf{Z}}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \widetilde{\mathbf{Z}} \widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2})^{-1} \widetilde{\mathbf{Z}}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \widetilde{\mathbf{Z}} \\
& \times \mathbf{E}_j (\mathbf{I} + \widetilde{\mathbf{Z}}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \widetilde{\mathbf{Z}} \widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2})^{-1} \widetilde{\mathbf{Z}}^T \mathbf{W}_{\widehat{\boldsymbol{\beta}},\widehat{\mathbf{u}}} \widetilde{\mathbf{Z}} \Big\},
\end{aligned}$$

where $\widetilde{\mathbf{Z}} \equiv [\mathbf{Z}, \mathbf{Z}_k]$, $\widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2} \equiv \text{blockdiag}(\mathbf{G}_{\boldsymbol{\sigma}^2}, \sigma_k^2 \mathbf{I})$ and $\mathbf{E}_1, \ldots, \mathbf{E}_{v+1}$ are the diagonal matrices, with zeroes and ones on the diagonal, defined by $\widetilde{\mathbf{G}}_{\boldsymbol{\sigma}^2,\sigma_k^2} = \sum_{i=1}^{v} (\boldsymbol{\sigma}^2)_j \mathbf{E}_j + \sigma_k^2 \mathbf{E}_{v+1}$. The formula $R_{\sigma_k^2}^{\text{den}}$ can now be written as

$$R_{\sigma_k^2}^{\text{den}} \simeq \sqrt{1/[\mathcal{K}(\boldsymbol{\sigma}^2, 0)^{-1}]_{v+1,v+1}}$$

To calculate $R^{\text{den}}_{\sigma^2_k}$, first partition $\mathcal{K}(\boldsymbol{\sigma}^2, \sigma^2_k)$ as

$$\mathcal{K}(\boldsymbol{\sigma}^2, \sigma^2_k) = \begin{bmatrix} \mathcal{K}_{11}(\boldsymbol{\sigma}^2, \sigma^2_k) & \mathcal{K}_{12}(\boldsymbol{\sigma}^2, \sigma^2_k) \\ \mathcal{K}_{12}(\boldsymbol{\sigma}^2, \sigma^2_k)^T & \mathcal{K}_{22}(\boldsymbol{\sigma}^2, \sigma^2_k) \end{bmatrix}$$

where $\mathcal{K}_{11}(\boldsymbol{\sigma}^2, \sigma^2_k)$ is the $v \times v$ upper left-hand block corresponding to the current model. Then

$$R^{\text{den}}_{\sigma^2_k} \simeq \{\mathcal{K}_{22}(\widehat{\boldsymbol{\sigma}}^2, 0) - \mathcal{K}_{12}(\widehat{\boldsymbol{\sigma}}^2, 0)^T \mathcal{K}_{11}(\widehat{\boldsymbol{\sigma}}^2, 0)^{-1} \mathcal{K}_{12}(\widehat{\boldsymbol{\sigma}}^2, 0)\}^{1/2}.$$

Note that the matrix inversion $\mathcal{K}_{11}(\widehat{\boldsymbol{\sigma}}^2, 0)^{-1}$ needs only be done once for the current model for each candidate model.

3.3.3 *The mAIC criterion*

For the model defined by $(\beta, \mathbf{u}, \boldsymbol{\sigma}^2)$ the marginal Akaike Information Criterion (mAIC) is

$$\text{mAIC}(\beta, \mathbf{u}, \boldsymbol{\sigma}^2) = -2\ell(\widehat{\beta}, \widehat{\boldsymbol{\sigma}}^2) + 2\{\dim(\beta) + \dim(\boldsymbol{\sigma}^2)\}$$

where $\dim(\mathbf{v})$ denotes the dimension, or length, of the vector $\mathbf{v}$. In practice we replace ℓ by ℓ_{Laplace}. The word 'marginal' is used to distinguish the criterion from conditional AIC (cAIC) introduced to mixed model analysis by Vaida & Blanchard (2005). In smooth function contexts, cAIC differs from mAIC in that the former used an 'effective degrees of freedom' measure (e.g. Buja *et al.*, 1989) in the second term rather than the number of fixed effects and variance components. Recently, Wager, Vaida & Kauermann (2007) compared mAIC and cAIC for model selection in Gaussian response models and concluded comparable performance in that context. While similar comparisons are yet to be made in the logistic context it is unlikely that one will significantly dominate the other. Our decision to use mAIC in the default KOW algorithm is driven by the high premium we are placing on computational speed.

3.3.4 *Variants and extensions*

The algorithm described near the start of this section, with details as laid out in Sections 3.1–3, is the 'default' version of the KOW algorithm for building a parsimonious classifier; optimised for speed and implementation simplicity. There are a number of variants and extensions that could be considered — albeit at the expense of speed and simplicity. Some of these are:

- Replace the mAIC-based model selection strategy with one that uses hypothesis testing and p-values. This involves approximate distribution theory for the Rao statistics.
- Replace the simple forward selection algorithm with a more elaborate scheme. One option is to have forward selection up to the fullest model, followed by a backward selection phase, using Wald statistics, back to the smallest model. Such a strategy is used by Stone *et al.* (1997), for example.

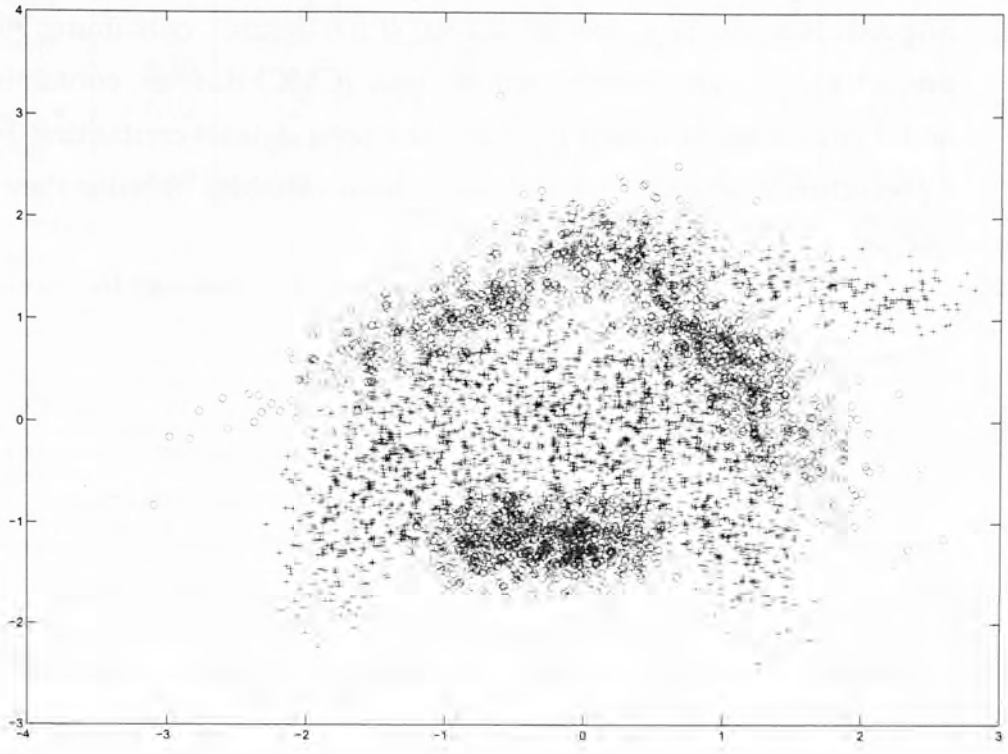


Figure 3.1: *Test sample 1 of 4900 data points from the Banana dataset.*

- Automate the choice between univariate and multivariate functions of the continuous predictors corresponding to the set S. The default version requires the user to either specify S or use only univariate functions.
- Decide whether a component should be added to the model based on criteria other than largest decrease in mAIC. Options include cAIC and versions of generalised cross-validation (e.g. Kooperberg, Bose & Stone, 1997).
- Insist that all non-linear components have a corresponding linear term. So if the non-linear component for x_k is selected for addition to the model then also add $\beta_k x_k$ if it is not already present.

3.4 Numerical Experience

We used both real and simulated datasets to test the effectiveness of the KOW algorithm. Two simulated datasets, *Orange* and *Banana*, were used for comparison. In *Orange* ten predictors $X_1, \ldots, X_{10}$ are simulated from a univariate standard normal distribution with one class having the first four predictors conditioned on $9 \leq \sum_{i=1}^{4} X_i^2 \leq 16$ (Hastie *et al.*, 2001). Thus *Orange* has 4 real predictors and 6 noise predictors. *Banana* is a 2 class 2-dimensional dataset simulated such that the points from four overlapping clusters two of which are banana-shaped. A sub-sample of these points are displayed in Figure 3.1. For the *Banana* dataset we added 6 standard normal noise predictors to make a total of 8 predictors for the dataset used for testing. Note that the data from the *Banana* dataset is not simulated from an additive model structure.

The four real datasets used were the *spam* dataset, containing 4601 observations and 57 predictors, the *pima indians diabetes* (PID) dataset, containing 768 observations and 8 predictors, the *contraceptive method choice* (CMC) dataset, containing 1473 observations and 9 covariates of mixed type and the *yeast* dataset containing 1484 observations and 8 predictors 2 of which we treat as ordinal variables because they have 2 and 3 unique values.

All datasets were obtained from the following Internet locations in 2008:

Name	Location
banana	`users.rsise.anu.edu.au/~raetsch/data/index.html`
CMC	`archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice`
PID/spam	`cran.au.r-project.org/src/contrib/mlbench_1.1-0.tar.gz`
orange	`www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/orange`
Yeast	`archive.ics.uci.edu/ml/datasets/Yeast`

Testing on the real datasets was conducted using 10-fold cross-validation. This involves splitting the dataset into 10 different parts. For the ith part we fit the model using the other 9 parts of the data, and calculate the prediction error of the model when predicting the ith part of the data. We did this for all 10 parts and averaged the 10 estimates to obtain the test error.

For each continuous variable we used a univariate O-spline basis as described in Wand & Ormerod (2008) (see also Chapter 2). We used 15 interior knots for each O-spline spaced equally with respect to the quantiles of each continuous variable.

3.4.1 *Illustrations*

We will first illustrate the KOW algorithm using the CMC dataset. This dataset is a subset of the 1987 national Indonesia contraceptive prevalence survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. The problem is to predict the current contraceptive method choice (no use or some use) of a woman based on her demographic and socio-economic characteristics. The covariates are of mixed data types and are listed below

1. Wife's age (`wife age`, continuous).
2. Wife's education (`wife edu`, ordinal: 1=low, 2, 3, 4=high).
3. Husband's education (`hus edu`, ordinal:1=low, 2, 3, 4=high).
4. Number of children ever born (`num chil`, continuous).
5. Wife's religion (`wife rel`, binary: 0=Non-Islam, 1=Islam).
6. Wife's now working? (`wife wor`, binary: 0=No, 1=Yes).
7. Husband's occupation (`hus occu`, nominal: 1, 2, 3, 4).
8. Standard-of-living index (`SOL`, ordinal: 1=low, 2, 3, 4=high).
9. Media exposure (`media ex`, binary: 0=Not Good, 1=Good).

We note that `num chil` is strictly speaking discrete taking the values 0 to 16. However treating `num chil` as a continuous variable and using smoothing methodology simplifies the analysis.

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
$\beta_{\{\text{wife age}\}}$	3.57	3.20	**8.34**	0.76	0.60	0.59	0.58	0.59
$\beta_{\{\text{wife edu}\}}$	**−9.39**							
$\beta_{\{\text{hus edu}\}}$	−6.02	−0.27	−0.88	−0.90	−0.60	0.22	0.39	0.44
$\beta_{\{\text{num chil}\}}$	−4.59	**−6.71**						
$\beta_{\{\text{wife rel=1}\}}$	2.80	0.66	0.82	1.93	2.03	1.75	**1.80**	
$\beta_{\{\text{wife wor=Y}\}}$	−1.51	−2.16	−1.64	−1.55	−0.07	−0.24	−0.16	−0.15
$\beta_{\{\text{hus occu=2}\}}$	2.15	1.41	1.54	1.65	1.32	1.32	1.34	**1.61**
$\beta_{\{\text{hus occu=3}\}}$	0.88	−1.64	−2.17	−0.91	−0.61	−1.13	−1.03	−1.03
$\beta_{\{\text{hus occu=4}\}}$	0.59	−0.61	−1.08	−0.98	−1.46	−1.32	−1.48	−1.53
$\beta_{\{\text{SOL}\}}$	−6.11	−2.98	−2.54	**−3.92**	**−3.84**			
$\beta_{\{\text{media ex=Y}\}}$	5.52	2.56	3.17	2.78	2.81	**2.10**		
mAIC_β	**1927.5**	**1882.9**	1813.0	1792.0	**1676.7**	**1674.2**	**1672.6**	1672.9
$s_{\{\text{wife age}\}}$	**2.50**	**1.88**	**3.88**					
$s_{\{\text{num chil}\}}$	0.32	0.85	−0.29	**15.43**				
mAIC_s	1968.9	1893.1	**1806.2**	**1690.3**				

Table 3.4.1: *Illustration of the steps taken by the KOW algorithm on the* contraceptive method choice *(CMC) dataset. Rao scores for each predictor, mAIC for the 'best linear' predictor mAIC_β and mAIC for the 'best nonlinear' predictor mAIC_s for each stage of the algorithm are listed in the columns. The best 'best linear' predictors, 'best nonlinear' predictors and lowest mAIC values are highlighted in bold.*

We wish to predict to response "Contraceptive used" which is modified to be binary by letting `contraceptive use` take the values 0 for "No-use" and 1 for "Long-term" or "Short-term" use. The largest possible model is

$$
\begin{aligned}
&\text{logit}\{\mathbb{P}(y_i = 1)\} \\
&= \beta_0 + \texttt{wife edu}_i\beta_{\{\text{wife edu}\}} + \texttt{hus edu}_i\beta_{\{\text{hus edu}\}} + \mathbb{I}_{\{\texttt{wife rel}_i=1\}}\beta_{\{\text{wife rel=1}\}} \\
&\quad + \mathbb{I}_{\{\texttt{wife wor}_i=Y\}}\beta_{\{\text{wife wor=Y}\}} + \mathbb{I}_{\{\texttt{media ex}_i=Y\}}\beta_{\{\text{media ex=Y}\}} + \texttt{SOL}_i\beta_{\{\text{SOL}\}} \\
&\quad + \mathbb{I}_{\{\texttt{hus occu}_i=2\}}\beta_{\{\text{hus occu=2}\}} + \mathbb{I}_{\{\texttt{hus occu}_i=3\}}\beta_{\{\text{hus occu=3}\}} + \mathbb{I}_{\{\texttt{hus occu}_i=4\}}\beta_{\{\text{hus occu=4}\}} \\
&\quad + \texttt{wife age}_i\beta_{\{\text{wife age}\}} + s_{\{\text{wife age}\}}(\texttt{wife age}_i) + \texttt{num chil}_i\beta_{\{\text{num chil}\}} \\
&\quad + s_{\{\text{num chil}\}}(\texttt{num chil}_i)
\end{aligned}
$$

which contains 11 possible predictors. The 8 steps of the KOW algorithm including Rao scores and mAIC values are illustrated in Table 3.4.1.

Table 3.4.1 corresponds to a running of the KOW algorithm taking 1.2 seconds, using 6 out of the 11 possible predictors and, as we will see, highly interpretable. In each of the four panels in Figure 3.2 cross-sections from the fitted additive function $\widehat{\eta}(\mathbf{x})$ for the CMC dataset are illustrated. The cross-section for each predictor corresponds to all other continuous predictors set to their medians. The effect of the values for wife's religion and media exposure are also illustrated by dropping or lifting $\widehat{\eta}(\mathbf{x})$ according to their values.

Based on the fit obtained from the KOW algorithm on the CMC dataset the following interpretations might be made for women in the study

- Increasing either wife's education or standard of living increases the chances contraceptives are used.

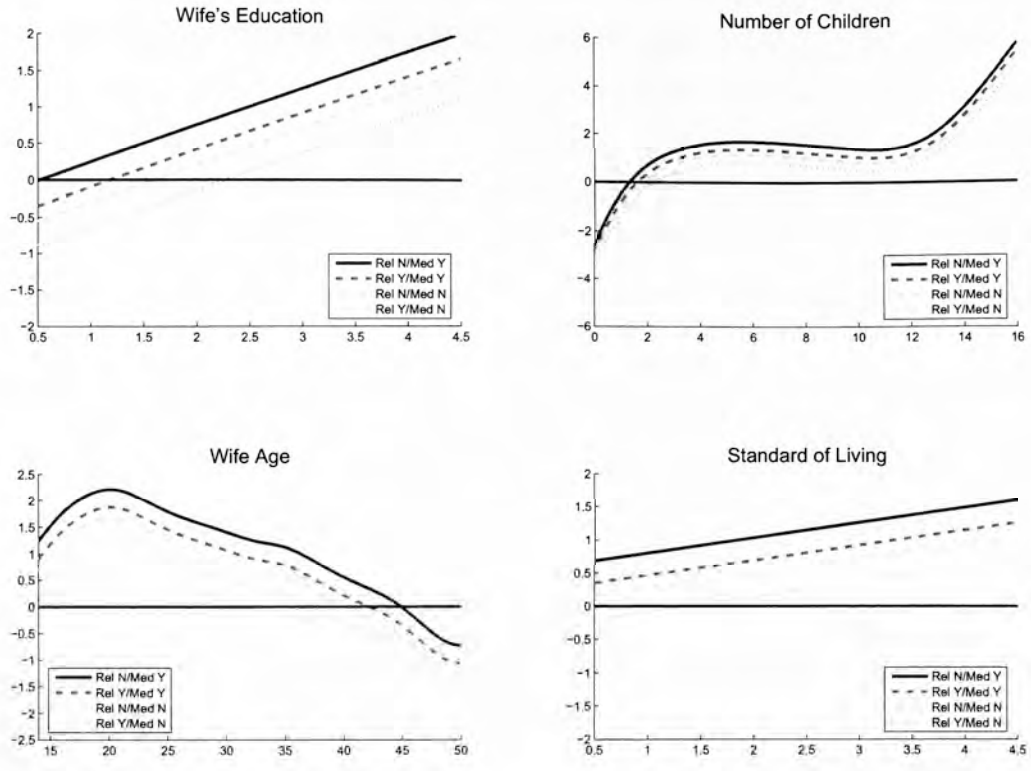


Figure 3.2: *The final model produced by the KOW algorithm for the* contraceptive method choice *(CMC) dataset. The cross-section for each predictor corresponds to all other continuous predictors set to their medians. Note that we have used the abbreviations wife's religion (REL) and media exposure (MED) above.*

- The predictors wife's age and number of children both have a nonlinear effect on contraceptive use.
- Islamic women are more likely to use contraceptives than non-Islamic women.
- Women who have had media exposure are more likely to use contraceptives.
- Contraceptive use peaks for women in their early 20's and decreases as they get older.
- Women without children or those with only one child are less likely to use contraceptives. Women with 3 to 12 children have similar chances of using contraceptives. Increasing the number of children above 12 proportionally increases the use of contraceptives.

Figure 3.3 illustrates cross-sections from the fitted additive function $\widehat{\eta}(\mathbf{x})$ for the *spam* dataset. The cross-section for each predictor corresponds to all other predictors set to their medians. When the curve moves above the zero line e-mails are more likely to be spam and when the curve moves below the zero line e-mails are less likely to be spam e-mails. For example when the proportion of number of times *business* is used to the total number of words is less than 2 there is nearly no effect but after the proportion is above 2 the probability that the e-mail is spam appears to increase (roughly) linearly. Curves that hover around the zero curve, for example the variable *our*, do not have a large effect on the predicted value.

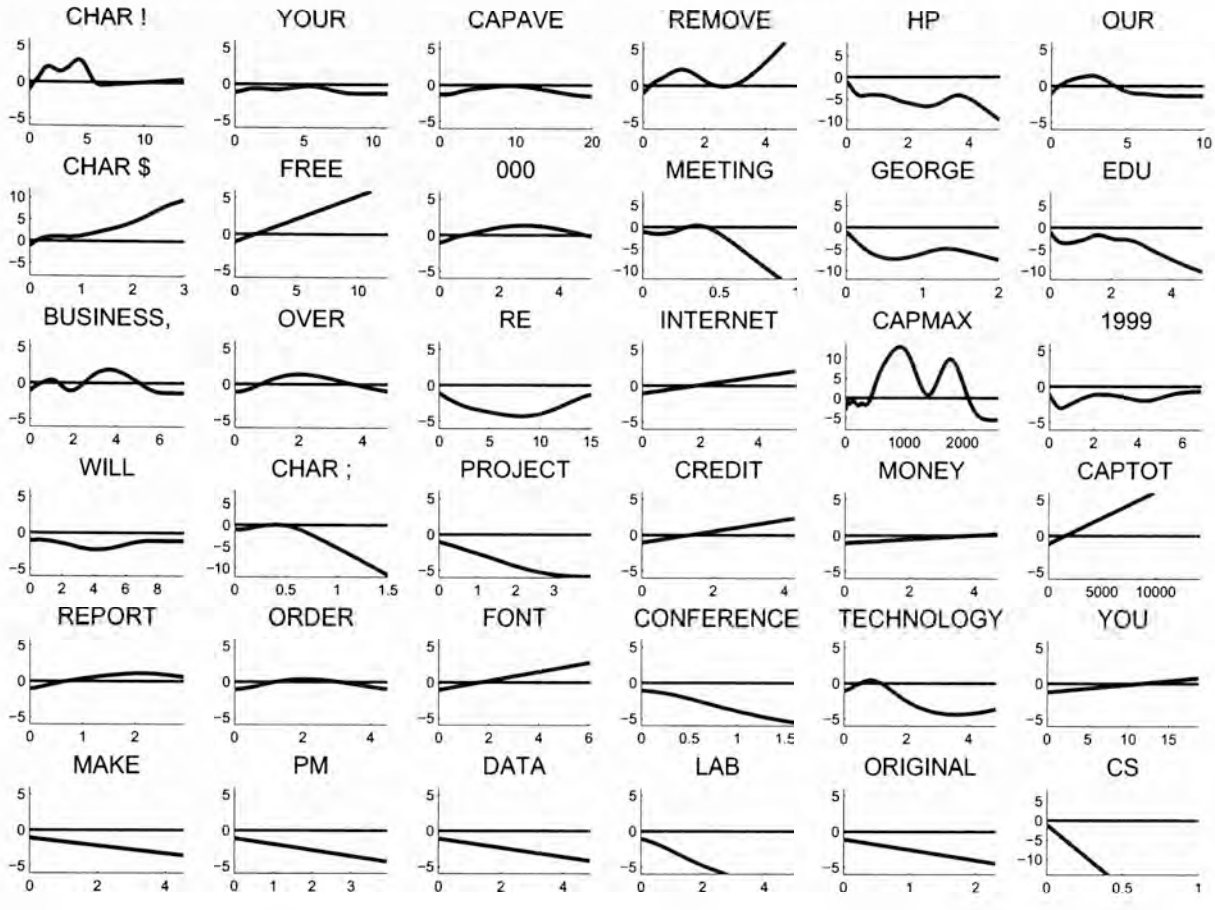


Figure 3.3: *A plot of fitted model for the* spam *dataset using the predictors as chosen by the KOW algorithm. The cross-section for each predictor corresponds to all other continuous predictors set to their medians.*

3.4.2 *Comparative Performance*

We now compare KOW with algorithms similar in their aims including: BRUTO (Hastie & Tibshirani, 1990) and the versions of the `R` function `step.gam()` (Chambers & Hastie, 1992; Hastie, 2006; Wood, 2006). The comparisons are made with respect to test error, parsimony and speed.

The `mgcv` package performs smoothing and model selection via optimisation of the generalised cross-validation (GCV) criteria. However `mgcv` does not perform variable selection as such but uses the related concept of shrinkage (see Hastie *et al.*, 2001, Chapter 3 for instance). For the purposes of testing we treat variables with an estimated effective degrees of freedom smaller than 0.01 as not included in the model.

The `step.gam()` function in the `gam` package requires the user to specify a list of possible degrees of freedom, or schemes, to use for each variable. In every dataset except the *spam* dataset, for reasons we will later state, we experimented with a number of schemes for each variable. The `step.gam()` method sequentially adds list elements from left to right for each variable and stops when the AIC fails to decrease. We specified these lists to allow for smoothing with 2, 4, 6 and 8 degrees of freedom, 8, 6, 4 and 2

degrees of freedom, 3, 6, 9 and 12 degrees of freedom or 12, 9, 6 and 3 degrees of freedom and also allowed for a linear fit or for the variable to be not included in the final fit. Thus, we specified schemes which started from larger degrees of freedom and tried to decrease the AIC by fitting models with smaller degrees of freedom and schemes which started from smaller degrees of freedom and tried to decrease the AIC by fitting models with larger degrees of freedom. The scheme with the smallest test error was recorded.

Finally, the BRUTO procedure uses least squares loss with smoothing splines where back-fitting model selection is based on an approximate GCV criteria.

For the *Orange* dataset each algorithm was run using 50 observations for each class (making a total of 100 observations), and the test error was attained by taking the average error from 50 simulations containing 500 observations for each class. For the *Banana* dataset each algorithm was run using 400 observations and the test error was attained by taking the average error from 100 simulations containing 4900 observations altogether.

Dataset	Method	Without Noise Test Error (%)	With Noise Test Error (%)	Real	Noise	Mean Time (seconds)
Banana	mgcv	28.12 (0.15)	29.06 (0.16)	2.00	3.41	22.74 (1.25)
	gam	30.25 (0.17)	30.69 (0.17)	2.00	0.71	5.64 (0.14)
	BRUTO	28.13 (0.12)	28.29 (0.13)	1.85	0.35	0.81 (0.00)
	KOW	28.11 (0.15)	28.76 (0.15)	1.87	1.07	1.08 (0.05)
Orange	mgcv	13.18 (0.86)	12.00 (0.85)	4.00	1.10	57.46 (2.69)
	gam	9.34 (0.29)	10.24 (0.35)	4.00	0.30	46.40 (3.35)
	BRUTO	8.58 (0.65)	9.10 (0.71)	4.00	0.30	0.14 (0.00)
	KOW	9.45 (0.39)	11.92 (0.87)	3.92	0.78	1.82 (0.06)

Table 3.4.2: *Averages (standard deviation) results for the* Banana *and* Orange *study described in Section 3.4.*

Examining Table 3.4.2 we see that all methods are fairly robust classifiers when noise variables are added. Furthermore all methods appear to be fairly good at discerning the real predictors from the noise predictors. KOW appears to select more noise predictors than all of the other methods accept `mgcv`. BRUTO appears to give slightly better classification rates on the *Orange* dataset.

The `gam` and BRUTO procedures failed on the full *spam* dataset and BRUTO failed on the *contraceptive method choice* dataset. The `step.gam()` procedure failed on the *spam* dataset because it creates an object indicating whether each of the possible 6^d candidate models had been fitted. For high d the size of this object becomes too large. We could not ascertain why the BRUTO procedure failed.

To allow comparison of all 4 methods we also worked with a reduced version of the *spam* dataset based on the 29 variables most often selected by KOW. We also simplified the model selection scheme used by the `step.gam()` method, for this case we allowed for either the variable to not be included or to be fit with approximately 4 degrees of freedom.

Examining Table 3.4.3 we see that KOW seems to gives similar, possibly slightly better, classification errors compared to `mgcv` and `gam`, with the aforementioned settings,

Dataset	Method	Test Error (%)	Mean No. Predictors Included	Mean Time (seconds)
Contraceptive Method Choice	mgcv	31.25 (1.30)	6.1	0.39 (0.03)
	gam	30.64 (1.17)	6.7	9.05 (0.37)
	BRUTO	failed	N/A	N/A
	KOW	30.77 (0.84)	6.6	0.89 (0.09)
Pima Indians Diabetes	mgcv	23.43 (1.90)	6.9	14.27 (1.08)
	gam	22.92 (2.23)	5.7	13.13 (1.18)
	BRUTO	50.64 (1.80)	5.3	0.12 (0.00)
	KOW	22.92 (1.62)	6.0	2.51 (0.11)
Spam	mgcv	5.89 (0.34)	50.7	21278.00 (4466.75)
	gam	failed	N/A	N/A
	BRUTO	failed	N/A	N/A
	KOW	5.38 (0.20)	37.6	1033.05 (98.93)
Reduced Spam	mgcv	6.15 (0.37)	28.4	4076.51 (694.35)
	gam	6.42 (0.22)	28.2	7521.10 (1467.74)
	BRUTO	16.86 (0.73)	25.7	1.01 (0.01)
	KOW	5.57 (0.25)	27.3	590.06 (62.13)
Yeast	mgcv	29.36 (4.00)	5.4	0.39 (0.03)
	gam	28.69 (3.42)	6.8	61.22 (3.62)
	BRUTO	53.50 (2.51)	5.6	0.03 (0.00)
	KOW	28.14 (3.48)	6.9	14.01 (0.64)

Table 3.4.3: *Means (standard deviations) for the test errors, number of predictors and running times using* `mgcv`, `gam`, *BRUTO and KOW methods on the* contraceptive method choice, Pima Indians diabetes, spam *and* yeast *datasets described in Section 3.4.*

and usually in less time. Based on the results for the *spam* dataset KOW seems to scale better to moderately sized datasets than all of the other methods considered. Furthermore, for the aforementioned reasons, the `gam` procedure becomes infeasible when a large number of predictors are used. Also when many predictors are used the computational time for `mgcv` may rule out its use on large data mining problems. BRUTO was faster than KOW, however the classification performance enjoyed by BRUTO on the simulated datasets did not seem to carry onto any of the real datasets where it failed miserably. We speculate that this is due to the fact that BRUTO models responses as Gaussian.

Finally, we should issue a note of caution on interpreting the test errors. For each method it is possible that lower test errors may be obtained by changing various settings, e.g. splines used, knot selection and model scheme (for `gam`) to name a few. For this reason all we can only conclude from Table 3.4.3 is that KOW, `mgcv` and `gam` have similar test errors for each given dataset.

3.5 Discussion

The KOW classification algorithm represents an appealing application of statistical inferential techniques to data mining and related problems. Parsimony and interpretability are delivered using likelihood-based inference ideas. Speed is obtained via Laplace-like approximations. Generalised linear mixed models, which have mainly been the providence of regression-type analyses of data from biostatistical studies, can be seen to have wider applicability.

While, in this chapter, we have concentrated on classification and logistic mixed models the methods presented are directly extendible to more general mixed models; e.g. those appropriate for count data, and non-classification problems such as variable selection in generalised additive model analyses. We envisage several useful by-products of the KOW algorithm for semiparametric analysis of multi-predictor data.

CHAPTER 4

Grid-Based Variational Posterior Approximations

4.1 Introduction

The problem of calculating integrals and summations is ubiquitous in the field of statistics. In statistical analysis of real world problems we model the uncertainty of unobserved or complex aspects of the system under observation. Taking expectations via integration or summation averages out this uncertainty, leaving us to deal with other aspects of the system. In frequentist statistics, amongst other situations, integrals occur when calculating moments, Fisher information, averaging over random effects or unobserved values or calculating confidence intervals. Similarly, from a Bayesian perspective integrals occur when calculating virtually anything including marginal distributions, posteriors distributions and credible intervals.

When these integrals or summations become analytically intractable we need to approximate them in some way. Traditionally these integrals were approximated using asymptotic methods typified by Edgeworth expansions, saddlepoint expansions and Laplace's method (Barndorff-Nielsen & Cox, 1989, 1994), or numerical quadrature methods (for example, Abramowitz & Stegun 1964, Chapter 25). Unfortunately there are situations where these methods are either inaccurate or prohibitively slow. With computing power continually increasing, Monte Carlo methods (Clayton, 1996; Robert & Casella, 1999; Gilks, Richardson & Spiegelhalter, 1996) can be used to increase accuracy. Unfortunately Monte Carlo methods can also be prohibitively slow for example when the Markov chain does not mix quickly or in importance sampling where the sampling distribution is not close to the target distribution.

Variational methods are a class of analytic approximations which have recently been applied to statistical problems in the machine learning literature. They are gaining popularity due to their computational speed, flexibility and simplicity; see for instance Jordan, Ghahramani, Jaakkola, & Saul (1999), Corduneanu & Bishop (2001), Ueda & Ghahramani (2002), Bishop & Winn (2003), Winn & Bishop (2005) and McGrory & Titterington (2007).

In the greater context of mathematics, the name *variational methods* corresponds to the classical set of methods known as the "calculus of variations" which can be used to find the extremum of an integral depending on an unknown function and its derivatives. In modern contexts variational methods describe a class of techniques where a problem is either transformed into an optimisation problem or directly formulated as an optimisation problem (Jaakkola, 2001). In this chapter we consider the use of variational methods

to transform integral problems that arise in Statistics into optimisation problems, but usually with some approximation.

These approximations typically involve a parameterized lower bound on the integral, which is then maximised over the parameters in order to tighten the bound. These lower bounds are generally constructed either by directly exploiting convexity properties of the integrand (Jordan *et al.*, 1999) or by the use of Jensen's inequality. In the latter case, Jensen's inequality can be used to develop a generalisation of the expectation maximisation (EM) algorithm of Dempster, Laird & Rubin (1977). Indeed Neal & Hilton (1998) showed that EM and several variants can be interpreted as a variational method which minimises the free energy (or equivalently the Kullback-Leibler divergence between two distributions). Later Attias (2000), inspired by the work of MacKay (1995) and Neal & Hilton (1998), used the free energy principal to generalise the EM algorithm, which became known as the variational expectation maximisation (VEM) or variational Bayes algorithm (although the same technique can be used in non-Bayesian contexts). This generalisation approximates the marginal likelihood by minimising the Kullback-Leibler divergence between the true posterior distribution and a convenient approximate posterior distribution. Ghahramani & Beal (2000) and Beal (2003) expanded upon this work to the class of conjugate-exponential models.

In each of the papers referenced above, variational approximations perform fairly well at the practical level. Unfortunately, the theoretical properties of the methods have received comparatively little attention, although a number of important theoretical contributions have been made by Humphreys & Titterington (2000), Hall, Humphreys & Titterington (2002) and Wang & Titterington (2003a, 2003b, 2004, 2005, 2006). These results include conditions for which variational methods are consistent, in various settings including missing value problems. As noted by Humphreys & Titterington (2000), Wang & Titterington (2005) and Consonni & Marin (2007) in various contexts, interval estimates corresponding to VEM approximations are typically "too small" because posterior variances are underestimated.

In this chapter we will make the following contributions:

1. An alternative variational approach for approximating posterior distributions is developed. We call this approach *grid-based variational posterior approximation* (GBVPA). This method is more accurate, sometimes considerably, than the typical variational method for approximating posteriors.
2. Discuss some alternative approaches to the optimisation approaches that arise in the implementation of the VEM algorithm.
3. The GBVPA algorithm is illustrated in two main examples: Bayesian linear regression and a Bayesian missing binary covariate model. The variational approximation to the later model is novel, compares well to Markov Chain Monte Carlo (MCMC) methods and scales well to large datasets ($> 10^6$) observations.

4. We show that for a frequentist missing continuous covariate model the EM and VEM algorithms deliver the same results. However the VEM algorithm does so more simply.
5. We demonstrate asymptotic consistency of a variational approximation of the Bayesian linear regression model.

4.2 Variational Approximations in Statistics

Suppose that we have observed $\mathbf{y} = (y_1, \ldots, y_n)$ which we have modelled via the joint density $[\mathbf{y}, \vartheta; \theta]$. In frequentist statistics ϑ is a vector of latent variables and θ are model parameters. Integrating out ϑ we obtain the likelihood

$$\mathcal{L}(\theta) = [\mathbf{y}; \theta] = \int [\mathbf{y}, \vartheta; \theta] d\vartheta. \tag{4.1}$$

In Bayesian analysis the joint density arises from the product $[\mathbf{y}, \vartheta; \theta] = [\mathbf{y}|\vartheta][\vartheta; \theta]$ where $[\mathbf{y}|\vartheta]$ is the sampling distribution and $[\vartheta; \theta]$ is the prior distribution. In this case ϑ is a vector of model parameters and θ is a vector of fixed prior hyperparameters which characterise knowledge about ϑ. The calculation of the posterior of ϑ requires calculating

$$[\vartheta|\mathbf{y}; \theta] = \frac{[\mathbf{y}, \vartheta; \theta]}{\int [\mathbf{y}, \vartheta; \theta] d\vartheta} \tag{4.2}$$

which is also possibly analytically intractable.

Note that when the ϑ are discrete we replace the integrals in (4.1) and (4.2) by summations. Summing over all combinations of the values for ϑ is can be computationally challenging due to exponential growth in the number of terms to be calculated (see for example equation (4.39)). Henceforth we will write $\int$ for simplicity for continuous and discrete ϑ.

The variational approximations to (4.1) or (4.2) which we will consider will identify parameterized (typically lower) bounds to the integrals and then optimise over any free parameters in order to tighten this bound, i.e.

$$[\vartheta|\mathbf{y}; \theta] \geq [\vartheta|\mathbf{y}; \theta, \xi]_L$$

where ξ are additional parameter and the subscript L denotes lower bound.

We will consider two types of variational approximations which have been applied to statistical problems and can be used separately or in combination. We will call these *tangent transforms* and *density transforms*.

4.2.1 *Tangent Transforms*

The first of the general methods we will look at for finding lower bounds is a simple illustration of variational approximations. The idea is to take advantage of the fact that any tangent to a convex function is a lower bound of that function (Rockafeller, 1972). We then use this lower bound to simplify the integral in such a way that the integrand becomes tractable.

Transform	Function	Variational Form	Optimal Value
exp	$\exp(x)$	$\exp(\xi) + \exp(\xi)(x - \xi)$	$\xi = x$
log	$-\log(x)$	$\log(\xi) + 1 - \xi x$	$\xi = x^{-1}$
ξ	$-\log(e^{-\frac{x}{2}} + e^{\frac{x}{2}})$	$-\log(e^{-\frac{\xi}{2}} + e^{\frac{\xi}{2}}) - \frac{\tanh(\xi/2)}{4\xi}(x^2 - \xi^2)$	$\xi = \pm x$

Table 4.2.1: *Some univariate variational forms. Each function in the second column is greater than the variational form in the third column for all values of x and ξ. The function is restored by substituting the optimal value in the fourth column into the variational form. The first column contains specific names for each of the tangent transforms.*

Suppose that $f(\mathbf{x})$ is a convex differentiable function in $\mathbf{x} \in \mathbb{R}^n$ for some integer n. Then

$$f(\mathbf{x}) \geq f(\boldsymbol{\xi}) + (\mathsf{D}_{\mathbf{x}} f(\boldsymbol{\xi}))^T (\mathbf{x} - \boldsymbol{\xi}) \text{ for all } \mathbf{x}, \boldsymbol{\xi} \in \mathbb{R}^n. \tag{4.3}$$

Indeed

$$f(\mathbf{x}) = \max_{\boldsymbol{\xi}} \left\{ f(\boldsymbol{\xi}) + (\mathsf{D}_{\mathbf{x}} f(\boldsymbol{\xi}))^T (\mathbf{x} - \boldsymbol{\xi}) \right\}. \tag{4.4}$$

Similarly, if $f(\mathbf{x})$ is a concave differentiable function we reverse the direction of the inequality in (4.3) and minimise rather than maximise in (4.4).

Jordan *et al.* (1999) offers an insight into a more general approach to variational approximations based on the duality theory of convex analysis. Some examples of lower bounds we will encounter in the upcoming chapters are summarised in Table 4.2.1.

Example 4.1: As an illustrative example of the simplicity of these techniques, consider the model

$$\begin{aligned} y|\lambda &\sim \text{Poisson}(\lambda), \qquad \lambda > 0, y = 0, 1, 2, \ldots \\ \lambda &\sim \text{Gamma}(\alpha, \beta), \quad \alpha, \beta > 0, \end{aligned} \tag{4.5}$$

where y is a single observation and

$$[\lambda; \alpha, \beta] = \beta^\alpha \lambda^\alpha e^{-\beta\lambda} / \Gamma(\alpha).$$

and $\Gamma(\cdot)$ is the gamma function (see Abramowitz & Stegun, 1964, Chapter 6). If we integrate out the random parameter λ we obtain the negative binomial distribution:

$$[y; \alpha, \beta] = \int_0^\infty \left(\frac{\lambda^y}{y!} e^{-\lambda} \right) \left(\frac{\lambda^{\alpha-1} \beta^\alpha \exp(-\lambda\beta)}{\Gamma(\alpha)} \right) d\lambda = \frac{\Gamma(y + \alpha)}{\Gamma(y + 1)\Gamma(\alpha)} \beta^\alpha (1 + \beta)^{-(y+\alpha)}. \tag{4.6}$$

Suppose, for the purposes of illustration, that the expression for the marginal likelihood (4.6) has no closed form. A tangent transform might "simplify" the distribution $[\lambda]$ and hence the integrand in (4.6) by using the tangent bound

$$\exp(-x) \geq \exp(-\xi) - \exp(-\xi)(x - \xi)$$

to obtain

$$[\lambda; \alpha, \beta] \geq [\lambda; \alpha, \beta, \xi]_L \equiv \frac{\lambda^{\alpha-1} \beta^\alpha \left(e^{-\xi} - e^{-\xi}(-\lambda\beta - \xi) \right)}{\Gamma(\alpha)}$$

which holds for all λ, α, β and ξ. Hence

$$\begin{aligned}
[y;\alpha,\beta] &\geq [y;\alpha,\beta,\xi]_L \\
&= \int_0^\infty [y|\lambda][\lambda;\alpha,\beta,\xi]_L d\lambda \\
&= \int_0^\infty \left(\frac{\lambda^y}{y!}e^{-\lambda}\right)\left(\frac{\lambda^{\alpha-1}\beta^\alpha\left(e^{-\xi}-e^{-\xi}(-\lambda\beta-\xi)\right)}{\Gamma(\alpha)}\right) d\lambda \\
&= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}\beta^\alpha \exp(-\xi)(1-\beta y-\beta\alpha+\xi).
\end{aligned}$$

Maximising $[y;\alpha,\beta,\xi]_L$ with respect to ξ decreases the gap between $[y;\alpha,\beta]$ and $[y;\alpha,\beta,\xi]_L$. It is easy to show that $[y;\alpha,\beta,\xi]_L$ is maximised when $\widehat{\xi} = -\beta(y+\alpha)$. Substituting this value for $\widehat{\xi}$ back into $[y;\alpha,\beta,\xi]_L$ we obtain

$$[y;\alpha,\beta] \geq [y;\alpha,\beta,\widehat{\xi}]_L = \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}\beta^\alpha \exp(-\beta(y+\alpha)). \tag{4.7}$$

This bound can be verified by the fact that $(1+x)^{-a} \geq e^{-ax}$ for $a, x > 0$.

Figure 4.1 illustrates the likelihood and (4.7) as a function of $\alpha = \beta$ for 100 simulated points for true $\log(\alpha) = \log(\beta) \in \{-4,-3,-2,-1\}$. We can see from this figure that the variational approximation is more accurate for smaller values of $\alpha = \beta$.

4.2.2 *Expectation Maximisation as a Variational method*

A major development of variational approximations is based on a modification of the EM algorithm. The EM algorithm developed by Dempster *et al.* (1977) is a simple algorithm for maximum likelihood estimation which pivots between an "expectation step" and a "maximisation step". The EM algorithm is listed in Algorithm 2.

Algorithm 2 Expectation Maximisation

1. Initialise $\widehat{\boldsymbol{\theta}}$.
2. **Cycle**
   - **2.1. Expectation Step (E-step)**
     Calculate
     $$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{old}) \equiv \mathbb{E}_{\boldsymbol{\vartheta}|\mathbf{y}} \log [\mathbf{y}, \boldsymbol{\vartheta}; \boldsymbol{\theta}_{old}] \tag{4.8}$$
     where $\boldsymbol{\theta}_{old} = \widehat{\boldsymbol{\theta}}$.
   - **2.2. Maximisation Step (M-step)**
     Replace $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$ where
     $$\widehat{\boldsymbol{\theta}} := \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{old}).$$

   Until convergence.

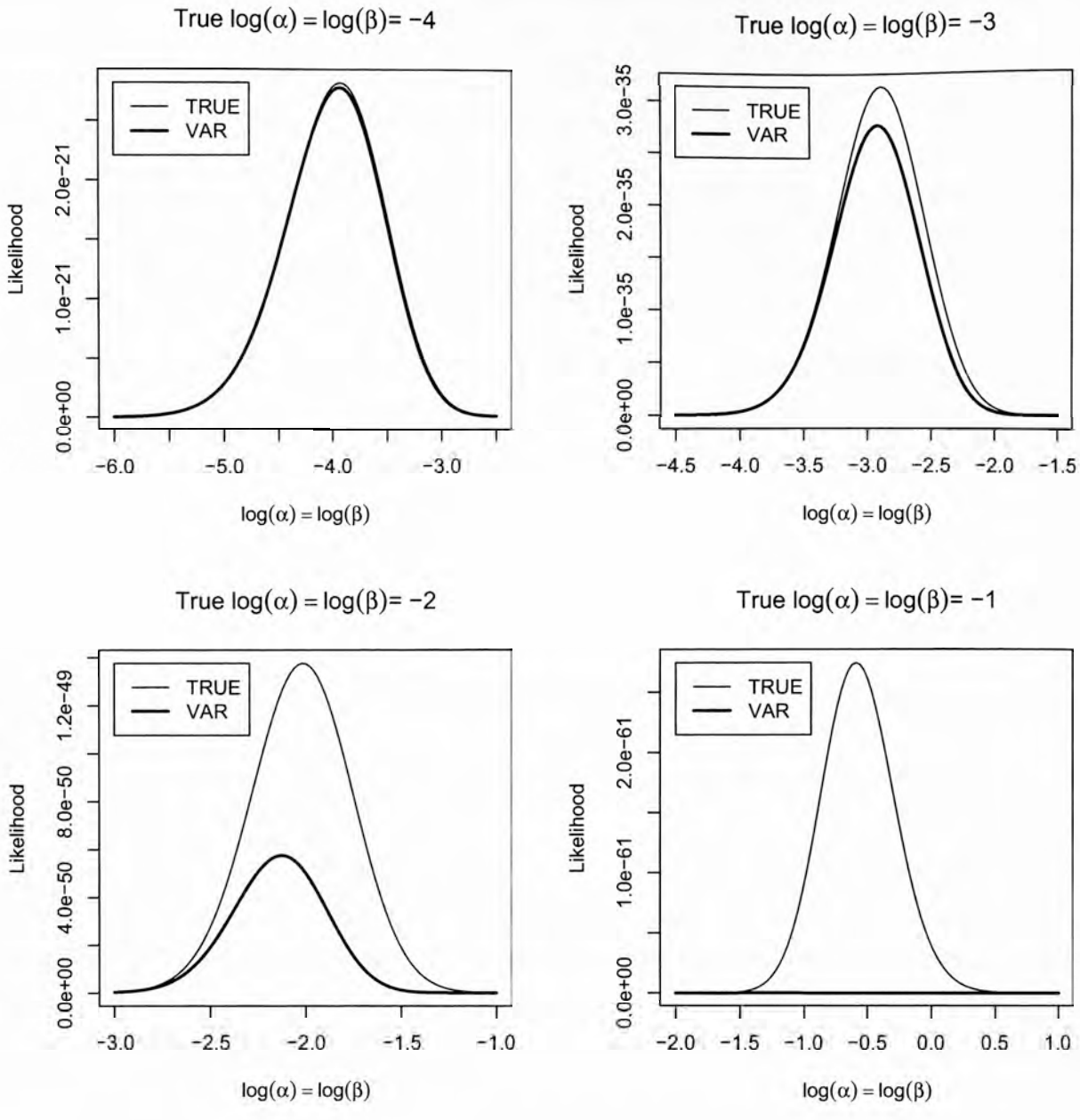


Figure 4.1: *Illustration of the likelihood and variational approximation for model (4.5) for different values of* $\alpha = \beta$.

Dempster *et al.* (1977) showed that if $\boldsymbol{\theta}$ is chosen by iteratively maximising $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{old})$ with respect to $\boldsymbol{\theta}$, the $\widehat{\boldsymbol{\theta}}$ will converge to a local maximum of the likelihood. The quantity Q arises as follows:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log[\mathbf{y};\boldsymbol{\theta}] \\
&= \log \int [\mathbf{y},\boldsymbol{\vartheta};\boldsymbol{\theta}] d\boldsymbol{\vartheta} \\
&= \log \int \frac{[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}][\mathbf{y},\boldsymbol{\vartheta};\boldsymbol{\theta}]}{[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]} d\boldsymbol{\vartheta} \\
&\geq \int [\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}] \log\left(\frac{[\mathbf{y},\boldsymbol{\vartheta};\boldsymbol{\theta}]}{[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]}\right) d\boldsymbol{\vartheta} \\
&= \mathbb{E}_{\boldsymbol{\vartheta}|\mathbf{y}}\left(\log[\mathbf{y},\boldsymbol{\vartheta};\boldsymbol{\theta}]\right) - \mathbb{E}_{\boldsymbol{\vartheta}|\mathbf{y}}\left(\log[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]\right) \\
&= Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{old}) + \mathcal{H}_{\boldsymbol{\vartheta}|\mathbf{y}} \\
&= \ell_L(\boldsymbol{\theta})
\end{aligned} \tag{4.9}$$

where $\ell_L(\boldsymbol{\theta})$ is a lower bound for $\ell(\boldsymbol{\theta})$ and the inequality follows from the use of Jensen's inequality $\varphi(\mathbb{E}(X)) \geq \mathbb{E}(\varphi(X))$ which holds for all concave functions $\varphi(\cdot)$ (with the inequality being reversed if $\varphi(\cdot)$ is a convex function). Here $X = \frac{[\mathbf{y},\boldsymbol{\vartheta};\boldsymbol{\theta}]}{[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]}$ and expectations are taken with respect to $[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]$, $\mathbb{E}_{\boldsymbol{\vartheta}|\mathbf{y}}$ denotes expectations with respect to $\boldsymbol{\vartheta}|\mathbf{y}$, and $\mathcal{H}_{\boldsymbol{\vartheta}|\mathbf{y}}$ is the entropy of $\boldsymbol{\vartheta}|\mathbf{y}$ where the Shannon's entropy (henceforth simply entropy) of a density $[\boldsymbol{\vartheta}]$ is given by

$$\mathcal{H}_{\boldsymbol{\vartheta}} \equiv -\int [\boldsymbol{\vartheta}] \log[\boldsymbol{\vartheta}] d\boldsymbol{\vartheta} = -\mathbb{E}_{\boldsymbol{\vartheta}} (\log[\boldsymbol{\vartheta}]) . \tag{4.10}$$

Note that $\mathcal{H}_{\boldsymbol{\vartheta}|\mathbf{y}}$ is a constant function of $\boldsymbol{\theta}$ and can be ignored in the M-step of the EM algorithm.

Neal & Hilton (1998) describe the EM algorithm and several variants in terms of free energy minimisation. In general the difference between the lower bound $\ell_L(\boldsymbol{\theta})$ and the likelihood $\ell(\boldsymbol{\theta})$ is given by the Kullback-Leibler (KL) divergence, relative entropy or free energy between $[\boldsymbol{\vartheta}|\mathbf{y}]$ and $[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]$, i.e.

$$\begin{aligned} \ell(\boldsymbol{\theta}) - \ell_L(\boldsymbol{\theta}) &= \log [\mathbf{y};\boldsymbol{\theta}] - \mathbb{E}_{\boldsymbol{\vartheta}|\mathbf{y}} (\log [\mathbf{y},\boldsymbol{\vartheta};\boldsymbol{\theta}]) + \mathbb{E}_{\boldsymbol{\vartheta}|\mathbf{y}} (\log[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]) \\ &= \mathbb{E}_{\boldsymbol{\vartheta}|\mathbf{y}} (\log[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]) - \mathbb{E}_{\boldsymbol{\vartheta}|\mathbf{y}} (\log [\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}]) \\ &= KL([\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]||[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}]) \end{aligned} \tag{4.11}$$

where

$$\begin{aligned} KL(f_1||f_2) &\equiv -\int f_1(\boldsymbol{\vartheta}) \log \left(\frac{f_2(\boldsymbol{\vartheta})}{f_1(\boldsymbol{\vartheta})} \right) d\boldsymbol{\vartheta} = Q(f_1||f_2) - \mathcal{H}_{f_1} \\ Q(f_1||f_2) &\equiv -\int f_1(\boldsymbol{\vartheta}) \log (f_2(\boldsymbol{\vartheta})) \, d\boldsymbol{\vartheta}, \text{ and} \end{aligned}$$

for any two densities $f_1(\boldsymbol{\vartheta})$ and $f_2(\boldsymbol{\vartheta})$, the quantity $Q(f_1||f_2)$ is the cross-entropy between f_1 and f_2 and $\mathcal{H}_{f_1}$ is the entropy of f_1. It is possible to show that $KL(f_1||f_2) \geq 0$ and $KL(f_1||f_2) = 0$ if and only if $f_1 \equiv f_2$ (e.g. Shorack & Wellner, 1986; Csiszár & Shields, 2004). Thus, if $\boldsymbol{\theta} = \boldsymbol{\theta}_{old}$ then $\ell(\boldsymbol{\theta}) = \ell_L(\boldsymbol{\theta})$. Intuitively we can think of $KL(f_1||f_2)$ being a measure of similarity between f_1 and f_2. Note that KL is neither symmetric nor satisfies the triangle inequality and is therefore not a metric. Thus using (4.11) we can interpret the EM algorithm as a variational approximation because it replaces the integral in (4.9) with a sequence of maximisation problems.

4.2.3 *Variational Expectation Maximisation and Density Transforms*

Unfortunately for many problems of interest the calculation of (4.8) is no easier than calculating the log-likelihood (4.1). See for example Ruppert *et al.* (2003, Section 10.8.5). This is because in order to calculate $[\boldsymbol{\vartheta}|\mathbf{y}]$ we need to be able to calculate (4.1).

Variational expectation maximisation (VEM, see Algorithm 3) is a generalisation of the EM algorithm of Dempster *et al.*, (1977) where the expectation step in the EM algorithm, which in itself provides a lower bound for the marginal log-likelihood, is modified to obtain a more general class of lower bounds. Attais (2000), noted that the same logic

in (4.9) applies if we replace $[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}_{old}]$ with *any* density $\delta(\boldsymbol{\vartheta};\boldsymbol{\xi})$ i.e.

$$\ell(\boldsymbol{\theta}) \geq \mathbb{E}_\delta\left(\log[\mathbf{y},\boldsymbol{\vartheta};\boldsymbol{\theta}]\right) - \mathbb{E}_\delta\left(\log(\delta(\boldsymbol{\vartheta};\boldsymbol{\xi}))\right) = Q(\boldsymbol{\theta}|\boldsymbol{\xi}) + \mathcal{H}_\delta = \ell_L(\boldsymbol{\theta}|\boldsymbol{\xi}) \tag{4.12}$$

where $\ell_L(\boldsymbol{\theta}|\boldsymbol{\xi})$ is a lower bound for $\ell(\boldsymbol{\theta})$, $\boldsymbol{\xi}$ are variational parameters (which may include elements of $\boldsymbol{\theta}$ and/or $\boldsymbol{\theta}_{old}$) and

$$Q(\boldsymbol{\theta}|\boldsymbol{\xi}) = \mathbb{E}_\delta\left(\log\left[\mathbf{y},\boldsymbol{\vartheta};\boldsymbol{\theta}\right]\right).$$

Again, using the same logic we have

$$\ell(\boldsymbol{\theta}) - \ell_L(\boldsymbol{\theta};\boldsymbol{\xi}) \quad = KL(\delta(\boldsymbol{\vartheta};\boldsymbol{\xi}) \;||\; [\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}]). \tag{4.13}$$

For simplicity we use the following notation

- For compactness, unless there is room for confusion, $\delta(\boldsymbol{\vartheta};\boldsymbol{\xi})$ is denoted as $\delta(\boldsymbol{\vartheta})$ or simply δ.
- Jaakkola & Jordan (2000) refer to (4.12) as the q-transform of the log-likelihood (using q-densities instead of δ-densities above. We will instead refer to (4.12) as the more descriptive *density transform*). Furthermore, we name the density transform after the approximating distribution, e.g. if $\delta(\boldsymbol{\vartheta};\boldsymbol{\xi}) = \phi_{\boldsymbol{\Sigma}}(\boldsymbol{\vartheta}-\boldsymbol{\mu})$ where $\phi_{\boldsymbol{\Sigma}}(\boldsymbol{\vartheta}-\boldsymbol{\mu})$ is the multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. In this case we will denote the approximating distribution of $\boldsymbol{\vartheta}|\mathbf{y}$ as

$$\boldsymbol{\vartheta}|\mathbf{y} \sim_\delta N(\boldsymbol{\mu},\boldsymbol{\Sigma}).$$

There are a number of points to be made: The "closer" $\delta(\boldsymbol{\vartheta};\boldsymbol{\xi})$ is to $[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}]$ (i.e the smaller the KL-divergence between $\delta(\boldsymbol{\vartheta};\boldsymbol{\xi})$ and $[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}]$), the "closer" $\ell_L(\boldsymbol{\theta};\boldsymbol{\xi})$ is to $\ell(\boldsymbol{\theta})$; Maximising $\ell_L(\boldsymbol{\theta};\boldsymbol{\xi})$ is equivalent to minimizing the KL-divergence between $\delta(\boldsymbol{\vartheta};\boldsymbol{\xi})$ and $[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}]$; If $[\boldsymbol{\vartheta}|\mathbf{y};\boldsymbol{\theta}] = \delta(\boldsymbol{\vartheta};\widehat{\boldsymbol{\xi}})$ for some $\widehat{\boldsymbol{\xi}}$ then $\ell_L(\boldsymbol{\theta};\widehat{\boldsymbol{\xi}}) = \ell(\boldsymbol{\theta})$.

Also note that the EM algorithm essentially minimises the cross entropy term Q and ignores the entropy term $\mathcal{H}$ since it does not depend on parameters to be optimised. For the VEM algorithm we need both Q and $\mathcal{H}$ terms since the entropy term depends on the variational parameters $\boldsymbol{\xi}$.

In practice we choose $\delta(\boldsymbol{\vartheta})$ from a convenient class of distributions so that we can easily calculate expectations of the joint log-likelihood with respect to the density transform. Suppose that we partition $\boldsymbol{\vartheta}$ into $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1,\boldsymbol{\vartheta}_2)$. Such partitions are usually natural within the context of particular models. For example, in generalised linear mixed models (see Chapter 5) it is natural to classify variables as random effects, variance components or nuisance parameters; in graphical models it is natural to partition the model by nodes (see, for example, Jordan *et al.*, 1999; Beal, 2003; Jaakkola & Jordan, 200l; Jordan, 2004). Often within the variational literature to factorise the density δ to correspond to this grouping, i.e. $\delta(\boldsymbol{\vartheta}) = \delta_1(\boldsymbol{\vartheta}_1)\delta_2(\boldsymbol{\vartheta}_2)$. The extreme case where $\boldsymbol{\vartheta} = (\vartheta_1,\ldots,\vartheta_m)$ and $\delta(\boldsymbol{\vartheta}) = \prod_{i=1}^m \delta_i(\vartheta_1)$ is called the *mean-field approximation* (see Beal 2003, Chapter 2).

Most variational approximations in this thesis use this idea (see Example 4.2 below and Sections 4.5 and 4.6 for examples). Factorisations such as these amounts to assuming independence between parameters for the approximating density δ. Mean-field approximations have been studied by, amongst others, (Saul, Jaakkola & Jordan, 1996; Jordan *et al.*, 1999; Ghahramani & Beal, 2001; Humphreys & Titterington, 2001; Jaakkola, 2001; Hall *et al.*, 2002; Wang & Titterington, 2003; Jordan, 2004; Consonni & Marin, 2007). In particular Jaakkola & Jordan (1998) considered mixtures of such densities to improve mean-field approximations.

Finally some cases it may be difficult to calculate (4.12) because of the density transform chosen. In such cases, Jensen's inequality may to find lower bounds to (4.12) and hence simplify calculations. In this case we minimise an upper bound on the KL-divergence.

Algorithm 3 Variational Expectation Maximisation

---

1. Variational Step

Select a density $\delta(\vartheta;\boldsymbol{\xi})$ to approximate $[\vartheta|\mathbf{y}]$. Initialise $\widehat{\boldsymbol{\theta}}$.

2. Cycle

2.1. Expectation Step

Calculate

$$\widehat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi}}{\operatorname{argmax}}\, \ell_L(\widehat{\boldsymbol{\theta}}|\boldsymbol{\xi}) \tag{4.14}$$

2.2. Maximisation Step

Replace $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$ where

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell_L(\boldsymbol{\theta}|\widehat{\boldsymbol{\xi}}).$$

Until convergence.

Example 4.2: Suppose we have the triplets of observations $(y_i, x_{1,i}, x_{2,i})$, $1 \leq i \leq n$ and hypothesise a linear relationship between the y_is and both the $x_{1,i}$s and the $x_{2,i}$s. However some of the x_2s are missing completely at random (MCAR) and we also suspect a linear relationship exists between the x_1s and the x_2s. A likelihood based model for a this situation might be

$$\begin{aligned} y_i|x_{2,i} &\sim N(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}, \sigma_y^2) \\ \text{and} \qquad x_{2,i} &\sim N(\widetilde{\beta}_0 + \widetilde{\beta}_1 x_{1,i}, \sigma_x^2). \end{aligned}$$

Suppose that x_2 contains missing values. First let us partition x_2 as $\mathbf{x}_2 = (\mathbf{x}_{2,obs}, \mathbf{x}_{2,mis})$ where $\mathbf{x}_{2,obs} = (\mathbf{x}_{2,obs,1}, \ldots, \mathbf{x}_{2,obs,n_{obs}})$ and $\mathbf{x}_{2,mis} = (\mathbf{x}_{2,mis,1}, \ldots, \mathbf{x}_{2,mis,n_{mis}})$. Similarly, let $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis})$ and $\mathbf{x}_1 = (\mathbf{x}_{1,obs}, \mathbf{x}_{1,mis})$ be partitions of y and x_1 coinciding with the "missingness" of x_2.

We wish to fit the parameters of interest $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_y^2, \widetilde{\boldsymbol{\beta}}, \sigma_x^2)$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ and $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_0, \widetilde{\beta}_1)$. To this end the EM approach updates $\boldsymbol{\theta}$ via the equation

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \mathbb{E}_{\mathbf{x}_{2,mis}|\mathbf{y},\mathbf{x}_{2,obs};\boldsymbol{\theta}^{(t)}} \log[\mathbf{y}, \mathbf{x}_{2,obs}, \mathbf{x}_{2,mis}; \boldsymbol{\theta}] \right\}.$$

We can calculate $[\mathbf{x}_{2,mis}|\mathbf{y}, \mathbf{x}_{2,obs}; \boldsymbol{\theta}]$ using results from Appendix A.2 from Wand & Jones (1993, 1995) to obtain

$$\begin{aligned}
&[\mathbf{x}_{2,mis}|\mathbf{y}, \mathbf{x}_{2,obs}; \boldsymbol{\theta}] \\
&= \frac{\phi_{\sigma_y^2\mathbf{I}}(\mathbf{y}_{mis} - \mathbf{X}_{mis}\boldsymbol{\beta})\phi_{\sigma_x^2\mathbf{I}}(\mathbf{x}_{2,mis} - \widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}})}{\int \phi_{\sigma_y^2\mathbf{I}}(\mathbf{y}_{mis} - \mathbf{X}_{mis}\boldsymbol{\beta})\phi_{\sigma_x^2\mathbf{I}}(\mathbf{x}_{2,mis} - \widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}})d\mathbf{x}_{2,mis}} \\
&= \frac{\phi_{\sigma_y^2\beta_2^{-2}\mathbf{I}}(\mathbf{x}_{2,mis} - \beta_2^{-1}(\mathbf{y}_{mis} - \mathbf{X}_{mis,-2}\boldsymbol{\beta}_{-2}))\phi_{\sigma_x^2\mathbf{I}}(\mathbf{x}_{2,mis} - \widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}})}{\int \phi_{\sigma_y^2\mathbf{I}}(\mathbf{y}_{mis} - \mathbf{X}_{mis}\boldsymbol{\beta})\phi_{\sigma_x^2\mathbf{I}}(\mathbf{x}_{2,mis} - \widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}})d\mathbf{x}_{2,mis}} \\
&= \frac{\phi_{\sigma_y^2\beta_2^{-2}\mathbf{I}}(\mathbf{x}_{2,mis} - \beta_2^{-1}(\mathbf{y}_{mis} - \mathbf{X}_{mis,-2}\boldsymbol{\beta}_{-2}))\phi_{\sigma_x^2\mathbf{I}}(\mathbf{x}_{2,mis} - \widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}})}{\phi_{\sigma_y^2\beta_2^{-2}\mathbf{I}+\sigma_x^2\mathbf{I}}(\widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}} - \beta_2^{-1}(\mathbf{y}_{mis} - \mathbf{X}_{mis,-2}\boldsymbol{\beta}_{-2}))} \\
&= \phi_{\boldsymbol{\Sigma}}(\mathbf{x}_{2,mis} - \boldsymbol{\mu})
\end{aligned} \tag{4.15}$$

where $\mathbf{X}_{mis} = [\mathbf{1}, \mathbf{x}_{1,mis}, \mathbf{x}_{2,mis}]$, $\mathbf{X}_{mis,-2} = [\mathbf{1}, \mathbf{x}_{1,mis}]$, $\widetilde{\mathbf{X}}_{mis} = [\mathbf{1}, \mathbf{x}_{1,mis}]$ and

$$\boldsymbol{\mu} = \frac{\sigma_x^2\beta_2(\mathbf{y}_{mis} - \mathbf{X}_{mis,-2}\widetilde{\boldsymbol{\beta}}_{-2}) + \sigma_y^2\widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}}}{\sigma_y^2 + \beta_2^2\sigma_x^2} \qquad \boldsymbol{\Sigma} = \frac{\sigma_y^2\sigma_x^2}{\beta_2^2\sigma_x^2 + \sigma_y^2}\mathbf{I} \tag{4.16}$$

and $\mathbf{I}$ is the identity matrix.

The $\boldsymbol{\mu}$ can be interpreted as a vector of "imputed" values with associated covariance matrix $\boldsymbol{\Sigma}$. From these two equations we can see how the current estimated variances σ_y^2 and σ_x^2 affect the value "imputed" when performing the expectation step. If σ_x^2 is large then the value "imputed" will be closer to $\mathbf{y}_{mis} - \mathbf{X}_{mis,-2}\boldsymbol{\beta}_{-2}$, i.e. the residual of the estimated error for $\mathbf{y}_{mis}$ without the term $\beta_2\mathbf{x}_{2,mis}$. Similarly if σ_y^2 is large then the value "imputed" will be closer to $\widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}}$.

The variational approach we will consider simplifies some of the above steps. This is done by assuming a distributional form for $[\mathbf{x}_{2,mis}|\mathbf{y}, \mathbf{x}_{2,obs}; \boldsymbol{\theta}]$ and then fitting any free parameters by minimizing the KL-divergence with respect to them. Thus instead of calculating (4.15), we use the density transform $\mathbf{x}_{2,mis} \sim_\delta \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are additional variational parameters so that

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &\geq \ell_L(\boldsymbol{\theta}; \boldsymbol{\xi}) \\
&= -\frac{n}{2}\log(\sigma_y^2) - \frac{n}{2}\log(\sigma_x^2) - \frac{\|\mathbf{y}_{obs} - \mathbf{X}_{obs}\boldsymbol{\beta}\|^2}{2\sigma_y^2} - \frac{\|\mathbf{x}_{2,obs} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}\|^2}{2\sigma_x^2} \\
&\quad - \frac{\|\mathbf{y}_{mis} - \mathbf{X}_{mis-2}\boldsymbol{\beta}_{-2} - \beta_2\boldsymbol{\mu}\|^2 + \beta_2^2\mathrm{tr}(\boldsymbol{\Sigma})}{2\sigma_y^2} - \frac{\|\boldsymbol{\mu} - \widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma})}{2\sigma_x^2} \\
&\quad + \frac{1}{2}\log|2e\pi\boldsymbol{\Sigma}|.
\end{aligned} \tag{4.17}$$

From (4.13) we note that maximising (4.17) is equivalent to minimizing the KL-divergence between $[\mathbf{x}_{2,mis}|\mathbf{y}, \mathbf{x}_{2,obs}; \boldsymbol{\theta}]$ and $\delta(\mathbf{x}_{2,mis})$.

Differentiating $\ell_L(\boldsymbol{\theta};\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$ we have

$$\begin{aligned} \mathsf{D}_{\boldsymbol{\mu}}\ell_L &= \frac{\beta_2(\mathbf{y}_{mis} - \mathbf{X}_{mis,-2}\boldsymbol{\beta}_{-2} - \beta_2\boldsymbol{\mu})}{\sigma_y^2} - \frac{(\boldsymbol{\mu} - \widetilde{\mathbf{X}}_{mis}\widetilde{\boldsymbol{\beta}})}{\sigma_x^2} \\ \mathsf{D}_{\Sigma_{ij}}\ell_L &= \frac{1}{2}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{-1} - \frac{\sigma_y^2\sigma_x^2}{\beta_2^2\sigma_x^2 + \sigma_y^2}\mathbf{I}\right)\mathbf{E}_{ij}\right) \end{aligned} \tag{4.18}$$

where $\mathbf{E}_{ij}$ is a matrix of zeros, except for the (i,j)th entry which is one and has the same dimensions as $\boldsymbol{\Sigma}$. Thus first order optimality conditions (see Appendix C) imply the same values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as (4.16). Thus, if we evaluate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ at $\boldsymbol{\theta}^{(t)}$, the equation $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ will be identical for EM and VEM algorithms.

Completing the calculations, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is given by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = -\frac{n}{2}\log(\sigma_x^2\sigma_y^2) - \frac{\|\mathbf{y} - \widehat{\mathbf{X}}\boldsymbol{\beta}\|^2 + \beta_2^2\mathrm{tr}(\boldsymbol{\Sigma})}{2\sigma_y^2} - \frac{\|\widehat{\mathbf{x}}_2 - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma})}{2\sigma_x^2} \tag{4.19}$$

where $\widehat{\mathbf{x}}_2 = (\mathbf{x}_{2,obs}, \boldsymbol{\mu})$, $\widehat{\mathbf{X}}_{mis} = [\mathbf{1}, \mathbf{x}_{1,mis}, \boldsymbol{\mu}]$, and $\widehat{\mathbf{X}} = (\mathbf{X}_{obs}, \widehat{\mathbf{X}}_{mis})$.

Maximising $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$ the update equations are

$$\begin{aligned} \sigma_y^2 &:= \frac{\|\mathbf{y} - \widehat{\mathbf{X}}\boldsymbol{\beta}\|^2 + \beta_2^2\mathrm{tr}(\boldsymbol{\Sigma})}{n} & \sigma_x^2 &:= \frac{\|\widehat{\mathbf{x}}_2 - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma})}{n} \\ \boldsymbol{\beta} &:= (\widehat{\mathbf{X}^T\mathbf{X}})^{-1}\widehat{\mathbf{X}}^T\mathbf{y} & \widetilde{\boldsymbol{\beta}} &:= (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\widehat{\mathbf{x}}_2 \end{aligned}$$

where $\widetilde{\mathbf{X}} = [\mathbf{1}, \mathbf{x}_1]$ and

$$\widehat{\mathbf{X}^T\mathbf{X}} = \begin{bmatrix} n & \mathbf{1}^T\mathbf{x}_1 & \mathbf{1}^T\widehat{\mathbf{x}}_2 \\ \mathbf{1}^T\mathbf{x}_1 & \mathbf{x}_1^T\mathbf{x}_1 & \mathbf{x}_1^T\widehat{\mathbf{x}}_2 \\ \mathbf{1}^T\widehat{\mathbf{x}}_2 & \mathbf{x}_1^T\widehat{\mathbf{x}}_2 & \widehat{\mathbf{x}}_2^T\widehat{\mathbf{x}}_2 + \mathrm{tr}(\boldsymbol{\Sigma}) \end{bmatrix}.$$

Comparing the practical ease of the EM and variational approaches, we note that the equations (4.15) for the EM algorithm were harder to calculate than (4.17) and (4.18) for the variational approach. This is because we assume that the posterior distribution $\mathbf{x}_{2,mis}|\mathbf{y}, \mathbf{x}_{2,obs}, \mathbf{x}_1; \boldsymbol{\theta}^{(t)}$ is a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$) and deduce their values by minimizing the KL-divergence between the true and unknown posterior distributions. On the other hand, for the EM approach we needed to calculate this conditional distribution in order to implement this algorithm.

4.3 Some Comments on Optimisation

One problem with the VEM algorithm as described in Algorithm 3 is that the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ are optimised separately. This has the potential to slow down convergence. In optimisation circles the extreme case where each parameter is optimised separately in a cyclic manner is called the cyclic coordinate descent method. While cyclic coordinate descent are often easier to describe and implement they can converge even slower than steepest decent methods which converge only linearly (Nocedal & Wright, 1999, Section

3.3). We therefore propose Algorithm 4 as a slight modification of the VEM algorithm where optimisation of parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ is performed simultaneously.

Algorithm 4 Variational Expectation Maximisation (Modification)

1. **Select Step:**
   Select a density $\delta(\vartheta;\boldsymbol{\xi})$ to approximate $\vartheta|\mathbf{y}$.
2. **Expectation Step:**
   Calculate
   $$\ell_L(\boldsymbol{\theta}|\boldsymbol{\xi}) = Q(\boldsymbol{\theta}|\boldsymbol{\xi}) + \mathcal{H}_\delta$$
3. **Maximisation Step:**
   Estimate $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$ where
   $$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\xi}}) = \underset{(\boldsymbol{\theta},\boldsymbol{\xi})}{\operatorname{argmax}}\, \ell_L(\boldsymbol{\theta}|\boldsymbol{\xi})$$

Furthermore, often in the literature on variational methods (as often in EM literature) $\ell_L(\boldsymbol{\theta};\boldsymbol{\xi})$ is maximised via a series of fixed point updates. Let $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\xi})$ and let us write $\ell_L(\boldsymbol{\theta};\boldsymbol{\xi})$ as $\ell_L(\vartheta)$. Suppose that $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_p)$ is some partition of $\boldsymbol{\zeta}$ into p subvectors. As discussed at the end of Section 4.2.3 such partitions occur naturally within the context of various models. The first order optimality conditions require

$$\begin{aligned} \mathsf{D}_{\boldsymbol{\zeta}_1}\ell_L &= \mathbf{0} \\ &\vdots \\ \mathsf{D}_{\boldsymbol{\zeta}_p}\ell_L &= \mathbf{0}. \end{aligned} \tag{4.20}$$

Often it is possible to find an explicit solution to each individual equation $\mathsf{D}_{\boldsymbol{\zeta}_i}\ell_L = \mathbf{0}$ of the form

$$\begin{aligned} \boldsymbol{\zeta}_1^{(t+1)} &:= g_1(\boldsymbol{\zeta}_1^{(t)}) \\ &\vdots \\ \boldsymbol{\zeta}_p^{(t+1)} &:= g_p(\boldsymbol{\zeta}_p^{(t)}) \end{aligned} \tag{4.21}$$

form some functions $g_i(\cdot)$, $1 \leq i \leq p$. Many of the optimisation problems in this thesis are treated in this manner.

Borrowing terminology from difference equation theory, a fixed point $\mathbf{x}^*$ of a function $g(\cdot)$ satisfies $\mathbf{x}^* = g(\mathbf{x}^*)$. Analogously we call (4.21) fixed point updates. Difference equation theory can then be used to analyse the convergence properties of (4.21). Based on the form of (4.21), most methods in optimisation, for example, the Newton-Raphson, quasi-Newton and steepest descent methods can be viewed as fixed point updates. However, we will use fixed point updates to describe updates (4.21) which are direct solutions to (4.20).

On the upside, fixed point equations:

1. are simple to implement,
2. require no additional matrix manipulation to perform updates (4.21) and
3. can give an intuitive feel about the solution to $\max_{\boldsymbol{\zeta}} \ell(\boldsymbol{\zeta})$.

The second point is particularly important when the length of ζ is $O(n)$ because it means we do not need to store Hessian matrices (or approximate Hessian matrices for quasi-Newton methods) of dimensions $O(n) \times O(n)$ (see Appendix C). We will see this in an example in Section 4.6.

On the downside, fixed point equations:

1. may not exist or may be difficult to determine,
2. may converge extremely slowly if $\zeta^{(0)}$ is too far from the solution to the maximisation problem (or not at all), and
3. may have difficulty enforcing implicit constraints on variables (e.g. positive variances).

Thus if $\dim(\zeta) \ll n$ then Newton-Raphson or quasi-Newton methods are preferred. Later in Chapter 5 we will consider a hybrid quasi-Newton/fixed-point method which outperforms either of these alone.

4.4 Grid-Based Variational Posterior Approximations

In Bayesian analysis it is common to approximate univariate marginal posterior densities. The posterior density for a single parameter ϑ_i is given by

$$[\vartheta_i|\mathbf{y}] = \frac{[\mathbf{y}, \vartheta_i]}{[\mathbf{y}]} = \frac{\int [\mathbf{y}, \boldsymbol{\vartheta}] d\boldsymbol{\vartheta}_{-i}}{\int [\mathbf{y}, \boldsymbol{\vartheta}] d\boldsymbol{\vartheta}} \tag{4.22}$$

where $\boldsymbol{\vartheta}_{-i} = (\vartheta_1, \ldots, \vartheta_{i-1}, \vartheta_{i+1}, \ldots, \vartheta_m)$, i.e. the vector $\boldsymbol{\vartheta}$ with the ith element removed.

A common variational approach to approximating posteriors is to select a density $\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi})$ to approximate $[\boldsymbol{\vartheta}|\mathbf{y}]$ where $\boldsymbol{\xi}$ are variational parameters. The density $\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi})$ is used in turn to approximate the marginal likelihood, i.e.

$$\log[\mathbf{y}] = \log \int [\mathbf{y}, \boldsymbol{\vartheta}] d\boldsymbol{\vartheta} \geq \mathbb{E}_\delta \log[\mathbf{y}, \boldsymbol{\vartheta}] + \mathcal{H}_\delta. \tag{4.23}$$

Then the posterior for $\boldsymbol{\vartheta}$ is best approximated, in the KL-divergence sense, by

$$[\boldsymbol{\vartheta}|\mathbf{y}] \approx \delta(\boldsymbol{\vartheta}; \widehat{\boldsymbol{\xi}}) \qquad \text{where} \qquad \widehat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi}}{\operatorname{argmax}} [\mathbf{y}; \boldsymbol{\xi}]_L,$$

which tightens the bound $[\mathbf{y}] \geq [\mathbf{y}; \boldsymbol{\xi}]_L$, or equivalently minimises the KL-divergence between $[\boldsymbol{\vartheta}|\mathbf{y}]$ and $\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi})$ (see Section 4.2.3). The posterior distribution for individual variables $\vartheta_i|\mathbf{y}$ are approximated by the marginals of $\delta(\boldsymbol{\vartheta}; \widehat{\boldsymbol{\xi}})$ for ϑ_i, i.e.

$$\delta(\vartheta_i; \widehat{\boldsymbol{\xi}}) = \int \delta(\boldsymbol{\vartheta}; \widehat{\boldsymbol{\xi}}) d\boldsymbol{\vartheta}_{-i}. \tag{4.24}$$

We call posteriors based on (4.24) variational posterior approximations (VPA).

Humphreys & Titterington (2000), Wang & Titterington (2005) and Consonni & Marin (2007) noted (in various settings) that $\delta(\boldsymbol{\vartheta}; \widehat{\boldsymbol{\xi}})$ typically underestimates the true posterior covariance $[\boldsymbol{\vartheta}|\mathbf{y}]$, sometimes dramatically. Thus interval estimates based on $\delta(\vartheta_i; \widehat{\boldsymbol{\xi}})$ inadequate because they are too small.

Instead of approximating (4.24) by (4.23) we will consider alternative approximations for individual posteriors based directly on (4.23). Let us suppose that ϑ_i is continuous. Using a density transform $\delta(\boldsymbol{\vartheta}_{-i};\boldsymbol{\xi})$ a variational lower bound to $\log[\mathbf{y},\vartheta_i]$ can be simply obtained by

$$\log[\mathbf{y},\vartheta_i] \geq \log[\mathbf{y},\vartheta_i;\boldsymbol{\xi}]_L \equiv \mathbb{E}_{\delta(\boldsymbol{\vartheta}_{-i})}\left(\log[\mathbf{y},\boldsymbol{\vartheta}]\right) + \mathcal{H}_{\delta(\boldsymbol{\vartheta}_{-i})}. \tag{4.25}$$

In order to tighten this bound we maximise $\log[\mathbf{y},\vartheta_i;\boldsymbol{\xi}]_L$ with respect to $\boldsymbol{\xi}$. Let

$$\widehat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi}}{\operatorname{argmax}}\left\{\log[\mathbf{y},\vartheta_i;\boldsymbol{\xi}]_L\right\} \tag{4.26}$$

so that $[\mathbf{y},\vartheta_i;\widehat{\boldsymbol{\xi}}]_L$ is also a tight lower bound for $[\mathbf{y},\vartheta_i]$. Note that in general the values for the optimal variational parameters $\widehat{\boldsymbol{\xi}}$ implicitly depends on the value of ϑ_i through (4.26) so that we write $\widehat{\boldsymbol{\xi}}(\vartheta_i)$.

Given $[\mathbf{y},\vartheta_i;\widehat{\boldsymbol{\xi}}(\vartheta_i)]_L$ we could approximate the marginal likelihood by

$$[\mathbf{y}]_L \equiv \int [\mathbf{y},\vartheta_i;\widehat{\boldsymbol{\xi}}(\vartheta_i)]_L d\vartheta_i \tag{4.27}$$

which is a lower bound for the marginal density $[\mathbf{y}]$. Given $[\mathbf{y},\vartheta_i;\widehat{\boldsymbol{\xi}}(\vartheta_i)]_L$ and $[\mathbf{y}]_L$ an approximation to $[\vartheta_i|\mathbf{y}]$ is given by

$$[\vartheta_i|\mathbf{y}] \approx \frac{[\mathbf{y},\vartheta_i;\widehat{\boldsymbol{\xi}}(\vartheta_i)]_L}{[\mathbf{y}]_L}. \tag{4.28}$$

The complicated dependency of $\widehat{\boldsymbol{\xi}}$ on ϑ_i means that it may be impossible to find a closed form expression for $[\mathbf{y},\vartheta_i;\widehat{\boldsymbol{\xi}}(\vartheta_i)]_L$. Instead we evaluate

$$\widehat{\boldsymbol{\xi}}_j = \max_{\boldsymbol{\xi}}[\mathbf{y},\widehat{\vartheta}_{ij};\boldsymbol{\xi}]_L$$

for a grid of values

$$(\widehat{\vartheta}_{i1},\ldots,\widehat{\vartheta}_{iN}) \tag{4.29}$$

for some integer N. We then approximate $\log[\mathbf{y},\vartheta_i;\widehat{\boldsymbol{\xi}}(\vartheta_i)]_L$ by some curve $\log[\mathbf{y},\vartheta_i]_G$ (where the subscript G denotes a grid based approximation) such that

$$\log[\mathbf{y},\widehat{\vartheta}_{ij}]_G = \log[\mathbf{y},\widehat{\vartheta}_{ij};\widehat{\boldsymbol{\xi}}_j]_L \text{ for } 1 \leq j \leq N, \tag{4.30}$$

i.e. $\log[\mathbf{y},\widehat{\vartheta}_{ij}]_G$ interpolates the points $(\widehat{\vartheta}_{ij},\log[\mathbf{y},\widehat{\vartheta}_{ij};\widehat{\boldsymbol{\xi}}_j]_L)$ for $1 \leq j \leq N$. Finally a grid based variational posterior approximation (GBVPA) for $[\vartheta_i|\mathbf{y}]$ is given by

$$[\vartheta_i|\mathbf{y}]_G \equiv \frac{[\mathbf{y},\vartheta_i]_G}{[\mathbf{y}]_G} \tag{4.31}$$

where the one dimensional integral $[\mathbf{y}]_G \equiv \int [\mathbf{y},\vartheta_i]_G d\vartheta_i$ is evaluated numerically. This approximation is formalised in Algorithm 5.

Algorithm 5 Grid Based Variational Posterior Approximation

1. Select a grid of N points $(\widehat{\vartheta}_{i1}, \ldots, \widehat{\vartheta}_{iN})$ for ϑ_i.
2. Calculate $\log[\mathbf{y}, \widehat{\vartheta}_{ij}; \widehat{\boldsymbol{\xi}}]_L = \max_{\boldsymbol{\xi}} \log[\mathbf{y}, \widehat{\vartheta}_{ij}; \boldsymbol{\xi}]_L$ for $1 \leq j \leq N$.
3. Find a $\log[\mathbf{y}, \vartheta_i]_G$ which interpolates the points $(\vartheta_{ij}, \log[\mathbf{y}, \widehat{\vartheta}_{ij}; \widehat{\boldsymbol{\xi}}]_L)_{1 \leq j \leq N}$.
4. Numerically approximate $[\mathbf{y}]_G$ where

$$[\mathbf{y}]_G \equiv \int [\mathbf{y}, \vartheta_i]_G d\vartheta_i. \tag{4.32}$$

5. The posterior distribution of $\log[\vartheta_i|\mathbf{y}]$ is then approximated by

$$[\vartheta_i|\mathbf{y}]_G = \frac{[\mathbf{y}, \vartheta_i]_G}{[\mathbf{y}]_G}. \tag{4.33}$$

There are a number of details which are required for a practical implementation of Algorithm 5 including: the choice and number of grid values, type of interpolation used to approximate $\log[\mathbf{y}, \vartheta_i; \widehat{\boldsymbol{\xi}}(\vartheta_i)]_L$ and quadrature method to approximate $[\mathbf{y}, \vartheta_i]_G$. The choices we have made in the following examples are as follows:

- The grid values are based on artificially widened intervals based on VPA. Suppose that $(\vartheta_{iL}, \vartheta_{iR})$ is a 95% highest posterior density credible region for ϑ_i based on the density $\delta(\vartheta_i; \widehat{\boldsymbol{\xi}})$. Then we let $(\widehat{\vartheta}_{i1}, \ldots, \widehat{\vartheta}_{iN})$ be equally spaced on the interval $\vartheta_i \in (\vartheta_{iL} - \delta/2, \vartheta_{iR} + \delta/2)$ where $\delta = \vartheta_{iR} - \vartheta_{iL}$ which may be truncated to be within the allowable values for ϑ_i.
- We experimented with two types of interpolation to approximate $[\mathbf{y}, \vartheta_i]_G$. We used interpolation using a polynomial of degree $N - 1$ and natural splines. The later case was implemented using the function `spline()` in the standard `R` library. Both types of interpolation worked well in practice.
- A 5,000 point composite trapezoid rule was used to approximate $[\mathbf{y}]_G$ on the interval $\vartheta_i \in (\vartheta_{iL} - \delta/2, \vartheta_{iR} + \delta/2)$. Other quadrature methods could be used, for example Gaussian quadrature, which could be both faster and more accurate, but the composite trapezoidal rule worked reasonably well and took a negligible amount of time. We note that higher point rules and/or adaptive quadrature methods might be needed for general problems.

Assuming all marginal posterior densities need to be approximated, one possible downside of GBVPA is that $N \times \dim(\boldsymbol{\vartheta})$ optimisation problems of the form (4.26) need to be solved. Thus, in practice, we seek to choose the grid (4.29) with as few points as possible but enough points to ensure that we have a reasonable approximation for $[\vartheta_i|\mathbf{y}]_G$.

We note that GBVPA could potentially be improved by:

1. using derivatives of $\log[\mathbf{y}, \vartheta_i; \boldsymbol{\xi}]_L$ with respect to ϑ_i;
2. choosing the grid (4.29) adaptively in some way;

3. assuming the optimisation problems $\widehat{\boldsymbol{\xi}} = \text{argmax}_{\boldsymbol{\xi}} \log[\mathbf{y}, \vartheta_i = \widehat{\vartheta}_{ij}; \boldsymbol{\xi}]_L$ are solved consecutively, the $\widehat{\boldsymbol{\xi}}$ based on previous grid points could be used as initial values for the next grid point $\widehat{\vartheta}_{i,j+1}$.

However we propose GBVPA as a starting place for such improvements.

Based on the application of GBVPA on the models considered in Sections 4.4–5 we have found that the marginal posterior approximations $[\vartheta_i|\mathbf{y}]_G$ appear to be better than marginal posteriors based on VPA when compared to densities estimated using posterior samples obtained via MCMC, even for N as small as about 20 and still reasonable for N as small as 12. But this suggests the question: for a particular dataset, when is one posterior density approximation "better" than another?

To answer this question, we compare posterior density approximations using VPA and GBVPA with posterior density approximations provided by using kernel density estimation techniques (Scott, 1992; Wand & Jones, 1995) for posterior samples obtained via MCMC. The kernel density estimates use the Gaussian kernel with the bandwidth chosen via a direct plug-in method (Wand and Jones, 1995, Section 3.6) using the `R` package `KernSmooth`. Alternatively the Sheather-Jones method (Sheather & Jones, 1991) can deliver excellent results.

It has been well-established in kernel smoothing literature that the choice of kernel has little effect on density estimates (e.g. Marron & Nolan, 1988, Wand & Jones, Chapter 2). However, how the bandwidth is chosen does matter. Extensive simulation studies (e.g. Park & Turlach, 1992; Cao, Cuevas & Gonzalez-Manteiga, 1994; Jones, Marron & Sheather, 1996) have shown that, for large sample sizes and densities that are Gaussian in shape, automatic bandwidth methods such as the direct plug-in methods and the Sheather-Jones method lead to quite accurate density estimates.

In the following sections we performed some initial tests based on generating data from shapes similar to the marginal posterior densities in Figures 4.2, 4.4 and 4.5 and found that sample sizes of 10,000 were sufficient to give reasonable accuracy, while 100,000 were sufficient to give very good accuracy, with the main difference being the estimation of the densities near the peaks. Hence, in each of the examples in this chapter, we use chains of length 505,000 which includes a burn-in of 5,000 and applied a thinning factor of 5 for posterior samples of size 100,000.

Let $[\vartheta_i|\mathbf{y}]_{MCMC}$ be the marginal posterior approximation for ϑ_i based on kernel density estimates of posterior samples obtained from MCMC. Assuming that the Markov chain has converged and the number of posterior samples is sufficiently large, then $[\vartheta_i|\mathbf{y}]_{MCMC}$ should be close to the exact posterior $[\vartheta_i|\mathbf{y}]$. Thus, for practical purposes, we could compare different marginal posterior density approximations $f(\vartheta_i)$ using the integrated square error (ISE) defined by

$$\text{ISE}\left(f(\vartheta_i), [\vartheta_i|\mathbf{y}]_{MCMC}\right) = \int \left(f(\vartheta_i) - [\vartheta_i|\mathbf{y}]_{MCMC}\right)^2 d\vartheta_i \tag{4.34}$$

where $f(\vartheta_i)$ is either $\delta(\vartheta_i; \widehat{\boldsymbol{\xi}})$ or $[\vartheta_i|\mathbf{y}]_G$.

The above method is a little abstract and makes more sense within the context of specific models. In the following sections VPA and GBVPA are compared for two simple models: Bayesian linear regression and a Bayesian missing binary covariate model.

4.5 Bayesian Linear Regression

Consider the following Bayesian linear regression model. Suppose we observe the pairs (y_i, x_i), $1 \leq i \leq n$ and

$$\begin{aligned} y_i|\boldsymbol{\beta}, \sigma_y^2 &\sim N(\beta_0 + \beta_1 x_i, \sigma_y^2) \\ \boldsymbol{\beta} &\sim N(0, \sigma_\beta^2) \\ \sigma_y^2 &\sim IG(A_y, B_y) \end{aligned}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1)$ where IG is the inverse-gamma distribution (see Appendix A). The prior hyperparameters are $\sigma_\beta^2 = 10^8$ and $A_y = B_y = 10^{-2}$ characterising vague priors on $(\boldsymbol{\beta}, \sigma_y^2)$. Note that even for this model, one of the simplest Bayesian models, the marginal likelihood does not have a closed form expression.

4.5.1 *Variational Posterior Approximation*

First, consider the task of approximating posteriors using VPA. The common variational approach to this would be to choose a density transform which mirrors the priors used, such that the approximate marginal posterior for each variable is independent, i.e. using $\delta(\boldsymbol{\beta}, \sigma_y^2) = \delta_\beta(\boldsymbol{\beta})\delta_{\sigma_y^2}(\sigma_y^2)$ where

$$\begin{aligned} \boldsymbol{\beta}|\mathbf{y} &\sim_{\delta_\beta} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \sigma_y^2|\mathbf{y} &\sim_{\delta_{\sigma_y^2}} IG(\alpha_y, \beta_y) \end{aligned}$$

where $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_y, \beta_y)$ are additional variational parameters. Using this, the approximate marginal distribution is given by

$$\begin{aligned} [\mathbf{y}] &\geq [\mathbf{y}; \boldsymbol{\xi}]_L \\ &= \mathbb{E}_\delta \left(\log[\mathbf{y}|\boldsymbol{\beta}, \sigma_y^2][\boldsymbol{\beta}][\sigma_y^2] \right) + \mathcal{H}_\delta \\ &= \mathbb{E}_\delta \left(\log[\mathbf{y}|\boldsymbol{\beta}, \sigma_y^2] \right) + \mathbb{E}_\delta \left(\log[\boldsymbol{\beta}] \right) + \mathbb{E}_\delta \left(\log[\sigma_y^2] \right) + \mathcal{H}_{\delta_\beta} + \mathcal{H}_{\delta_{\sigma_y^2}} \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}_\delta \left(\log[\mathbf{y}|\boldsymbol{\beta}, \sigma_y^2] \right) &= -\frac{n}{2}(\log(2\pi\beta_y) - \psi(\alpha_y)) - \frac{\alpha_y}{\beta_y} \cdot \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{X})}{2}, \\ \mathbb{E}_\delta \left(\log[\boldsymbol{\beta}] \right) &= -\log(2\pi\sigma_\beta^2) - \frac{\|\boldsymbol{\mu}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma})}{2\sigma_\beta^2}, \\ \mathbb{E}_\delta \left(\log[\sigma_y^2] \right) &= A_y \log(B_y) - \log\Gamma(A_y) - (A_y + 1)(\log(\beta_y) - \psi(\alpha_y)) - B_y \frac{\alpha_y}{\beta_y}, \\ \mathcal{H}_{\delta_\beta} &= \frac{1}{2}\log|2e\pi\boldsymbol{\Sigma}| \\ \text{and } \mathcal{H}_{\delta_{\sigma_y^2}} &= \alpha_y + \log(\beta_y) + \log\Gamma(\alpha_y) - (\alpha_y + 1)\psi(\alpha_y). \end{aligned}$$

We have used the facts that $\mathbb{E}_\delta(\sigma_y^{-2}) = \alpha_y/\beta_y$, $\mathbb{E}_\delta(\log(\sigma_y^2)) = \log(\beta_y) - \psi(\alpha_y)$, $\mathbb{E}_\delta(\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}) = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma})$ for any appropriately sized matrix $\mathbf{A}$ and $\psi(\cdot)$ is the digamma function (see Abramowitz & Stegun, 1964, Chapter 6).

Differentiating $[\mathbf{y};\boldsymbol{\xi}]_L$ with respect to $\boldsymbol{\xi}$ we obtain

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{\mu}}[\mathbf{y};\boldsymbol{\xi}]_L &= \frac{\alpha_y}{\beta_y}\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\mu}) - \sigma_\beta^{-2}\boldsymbol{\mu} \\
\mathsf{D}_{\Sigma_{ij}}[\mathbf{y};\boldsymbol{\xi}]_L &= \mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{-1} - \frac{\alpha_y}{\beta_y}\mathbf{X}^T\mathbf{X} - \sigma_\beta^{-2}\mathbf{I}\right)\mathbf{E}_{ij}\right)/2 \\
\mathsf{D}_{\alpha_y}[\mathbf{y};\boldsymbol{\xi}]_L &= \left(A_y + \frac{n}{2} - \alpha_y\right)\psi'(\alpha_y) - \left(B_y + \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\mu}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{X})}{2}\right)\left(\frac{1}{\beta_y}\right) + 1 \\
\mathsf{D}_{\beta_y}[\mathbf{y};\boldsymbol{\xi}]_L &= \left(B_y + \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\mu}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{X})}{2}\right)\frac{\alpha_y}{\beta_y^2} - \left(A_y + \frac{n}{2}\right)\frac{1}{\beta_y}.
\end{aligned} \tag{4.35}$$

Thus solving the first order optimality conditions for $\boldsymbol{\xi}$ we obtain the following fixed point updates

$$\begin{aligned}
\boldsymbol{\Sigma} &:= \left(\frac{\alpha_y}{\beta_y}\mathbf{X}^T\mathbf{X} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1} \\
\boldsymbol{\mu} &:= \frac{\alpha_y}{\beta_y}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{y} \\
\alpha_y &:= A_y + \frac{n}{2} \\
\beta_y &:= B_y + \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\mu}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{X})}{2}.
\end{aligned}$$

These updates are applied sequentially until convergence is obtained.

If $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\alpha}_y, \widehat{\beta}_y)$ then the variational posterior approximations are

$$\begin{aligned}
\beta_i|\mathbf{y} &\sim N(\widehat{\mu}_i, \widehat{\Sigma}_{ii}), \\
\sigma_y^2|\mathbf{y} &\sim IG(\widehat{\alpha}_y, \widehat{\beta}_y)
\end{aligned} \tag{4.36}$$

and the posterior means are

$$\begin{aligned}
\widehat{\sigma}_y^2 = \mathbb{E}_\delta(\sigma_y^2) = \quad \frac{\widehat{\beta}_y}{\widehat{\alpha}_y - 1} &= \frac{2B_y + \|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\mu}}\|^2 + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\mathbf{X}^T\mathbf{X})}{2A_y + n - 2} \\
\widehat{\beta}_i = \mathbb{E}_\delta(\beta_i) = \quad \widehat{\mu}_i &= \left[\left(\mathbf{X}^T\mathbf{X} + \frac{\widehat{\sigma}_\epsilon^2}{\sigma_\beta^2}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}\right]_i.
\end{aligned}$$

The maximum likelihood estimators for $\boldsymbol{\beta}$ and σ_y^2 from frequentist linear regression are

$$\begin{aligned}
\widehat{\sigma}_{y,ML}^2 &\equiv \frac{\|\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\mu}}\|^2}{n} \\
\widehat{\boldsymbol{\beta}}_{ML} &\equiv \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.
\end{aligned} \tag{4.37}$$

We see that as $\sigma_\beta^2 \to \infty$, $A_y, B_y \to 0$ and $n \to \infty$ that $\widehat{\boldsymbol{\beta}} \to \widehat{\boldsymbol{\beta}}_{ML}$ and $\widehat{\sigma}_y^2 \to \widehat{\sigma}_{y,ML}^2$. Since $\widehat{\boldsymbol{\beta}}_{ML}$ and $\widehat{\sigma}_{y,ML}^2$ are asymptotically consistent estimators so are $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}_y^2$ as $\sigma_\beta^2 \to \infty$ and $A_y, B_y \to 0$.

4.5.2 *Grid Based Variational Posterior Approximation for β_i*

We now consider the task of approximating the marginal posterior density of β_i using GBVPA. First, in order to approximate the β_i posteriors we fix $\beta_i = \widehat{\beta}_i$ and replace $[\boldsymbol{\beta}]$ and

$\delta(\boldsymbol{\beta})$ with $\beta_{-i} \sim N(0, \sigma_\beta^2)$ and $\beta_{-i}|\mathbf{y} \sim_{\delta_{\beta_{-i}}} N(\mu_{-i}, \sigma^2)$. The density transform of the joint likelihood for $\mathbf{y}$ and $\beta_i = \widehat{\beta}_i$ is

$$\begin{aligned}
[\mathbf{y}, \widehat{\beta}_i; \boldsymbol{\xi}]_L &= -\frac{n}{2}(\log(2\pi\beta_y) - \psi(\alpha_y)) - \frac{\alpha_y}{\beta_y} \cdot \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \sigma^2[\mathbf{X}^T\mathbf{X}]_{-i,-i}}{2} \\
&\quad -\frac{1}{2}\log(2\pi\sigma_\beta^2) - \frac{\mu_{-i}^2 + \sigma^2}{2\sigma_\beta^2} \\
&\quad +A_y \log(B_y) - \log\Gamma(A_y) - (A_y + 1)(\log(\beta_y) - \psi(\alpha_y)) - B_y\frac{\alpha_y}{\beta_y} \\
&\quad +\frac{1}{2}\log(2e\pi\sigma^2) + \alpha_y + \log(\beta_y) + \log\Gamma(\alpha_y) - (\alpha_y + 1)\psi(\alpha_y)
\end{aligned}$$

where $\mu_i = \widehat{\beta}_i$. Differentiating $[\mathbf{y}, \widehat{\beta}_i; \boldsymbol{\xi}]_L$ with respect to $\boldsymbol{\xi}$ we obtain similar derivatives to (4.35) and using similar algebra we arrive at the following fixed point updates

$$\begin{aligned}
\sigma^2 &:= \left(\frac{\alpha_y}{\beta_y}[\mathbf{X}^T\mathbf{X}]_{-i,-i} + \sigma_\beta^{-2}\right)^{-1}, \\
\mu_{-i} &:= \frac{\alpha_y}{\beta_y}\sigma^2\left([\mathbf{X}^T\mathbf{y}]_{-i} - [\mathbf{X}^T\mathbf{X}]_{-i,i}\widehat{\beta}_i\right), \\
\alpha_y &:= A_y + \frac{n}{2} \\
\text{and } \beta_y &:= B_y + \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \text{tr}(\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{X})}{2}.
\end{aligned}$$

Again, these updates are applied in order until convergence is obtained.

Suppose that $\widehat{\boldsymbol{\xi}} = (\widehat{\mu}_{-i}, \widehat{\sigma}^2, \widehat{\alpha}_y, \widehat{\beta}_y)$ are the values of the variational parameters at the convergence of these iterates. For a fixed $\beta_i = \widehat{\beta}_i$ we can calculate $[\mathbf{y}, \widehat{\beta}_i; \widehat{\boldsymbol{\xi}}]_L$ which is sufficient information to implement a GBVPA for $[\beta_i|\mathbf{y}]$.

4.5.3 *Grid Based Variational Posterior Approximation for σ_y^2*

We now consider the task of approximating the marginal posterior density of σ_y^2 using GBVPA. First, in order to approximate the σ_y^2 posteriors we fix $\sigma_y^2 = \widehat{\sigma}_y^2$ and remove the prior on σ_y^2 and $\delta_{\sigma_y^2}$ from δ. The density transform of the joint likelihood of $\mathbf{y}$ and $\sigma_y^2 = \widehat{\sigma}_y^2$ is

$$\begin{aligned}
[\mathbf{y}, \widehat{\sigma}_y^2; \boldsymbol{\xi}]_L &= -\frac{n}{2}\log(2\pi\widehat{\sigma}_y^2) - \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \text{tr}(\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{X})}{2\widehat{\sigma}_y^2} \\
&\quad - \log(2\pi\sigma_\beta^2) - \frac{\|\boldsymbol{\mu}\|^2 + \text{tr}(\boldsymbol{\Sigma})}{2\sigma_\beta^2} + \frac{1}{2}\log|2e\pi\boldsymbol{\Sigma}|.
\end{aligned}$$

Differentiating $[\mathbf{y}, \widehat{\sigma}_y^2; \boldsymbol{\xi}]_L$ with respect to $\boldsymbol{\xi}$ we arrive at the following update equations

$$\begin{aligned}
\boldsymbol{\Sigma} &:= \left(\widehat{\sigma}_y^{-2}\mathbf{X}^T\mathbf{X} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1} \\
\text{and } \boldsymbol{\mu} &:= \widehat{\sigma}_y^{-2}\boldsymbol{\Sigma}\mathbf{X}^T\mathbf{y}.
\end{aligned}$$

Again, these updates are applied in order until convergence is obtained.

Suppose that $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ are the values of the variational parameters at the convergence of these iterates. For a fixed $\sigma_y^2 = \widehat{\sigma}_y^2$ we can calculate $[\mathbf{y}, \widehat{\sigma}_y^2; \widehat{\boldsymbol{\xi}}]_L$ which is sufficient information to implement a GBVPA for $[\sigma_y^2|\mathbf{y}]$.

4.5.4 *Numerical Comparisons*

In order to test the effectiveness of the variational approximations for the Bayesian missing covariate model, we consider datasets where points (y_i, x_i), $1 \leq i \leq n$ are randomly generated from $x_i \sim \text{Unif}(0,1)$ and $y_i \sim N(\beta_0 + \beta_1 x, \sigma_y^2)$ where the parameters β_0, β_1 and σ_y^2 are fixed at $\beta_0 = -1$, $\beta_1 = 1$ and $\sigma_y^2 = 2$ respectively for some n. The marginal posterior approximations using VPA and GBVPA for a particular dataset with $n = 40$ points are illustrated in the first and second row of panels of Figure 4.2 respectively. The GBVPA used $N = 30$ grid points.

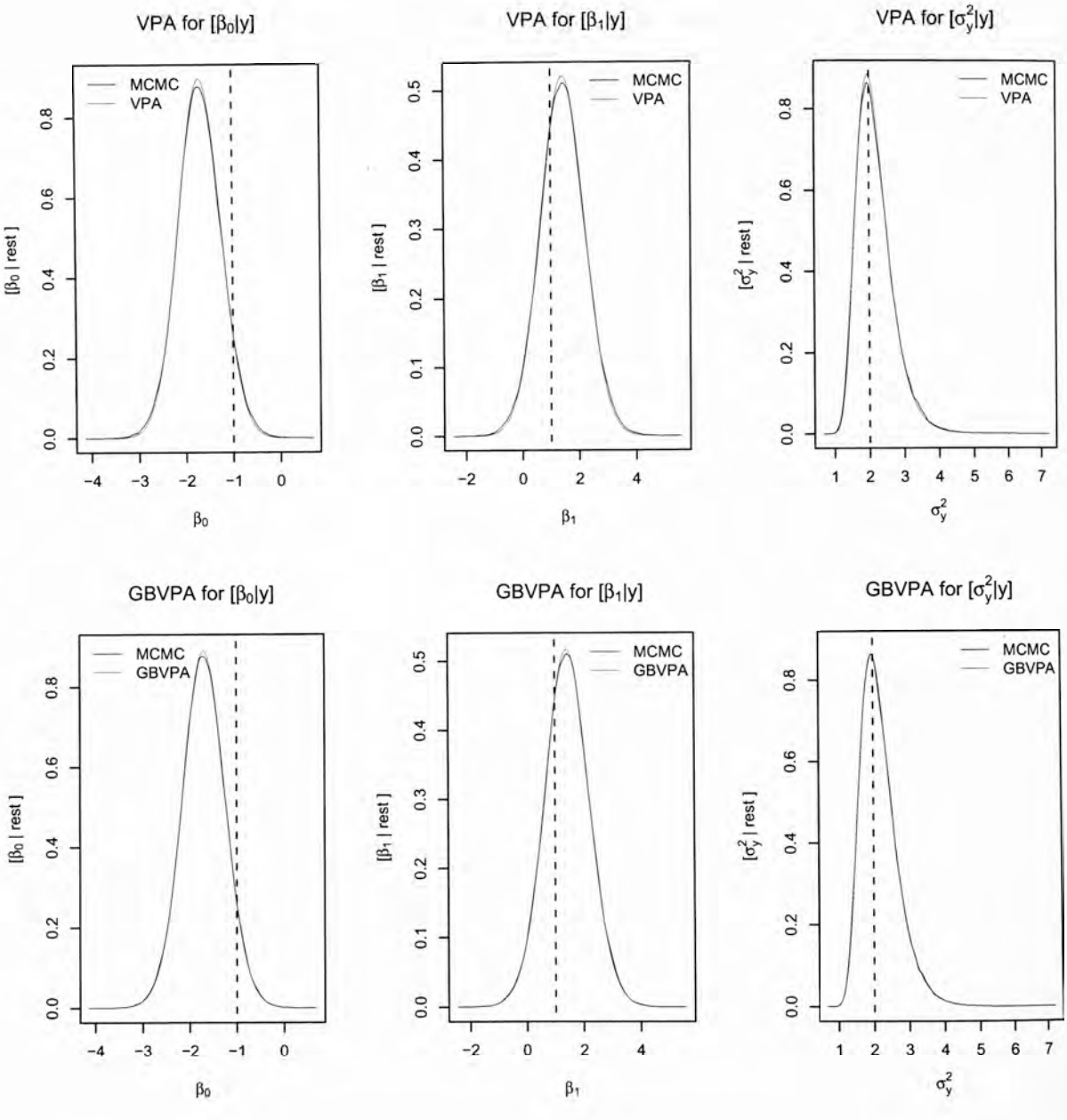


Figure 4.2: *Marginal posterior approximations for Bayesian linear regression model using VPA and GBVPA. The dashed vertical lines represent the "true" values used in the simulation.*

In Figure 4.2 the VPAs are very reasonable when compared to the MCMC posterior density approximations and do not significantly underestimate the posterior variances. Comparing VPA and GBVPA approximations we see that GBVPA has a slight advantage over VPA in terms of accuracy. There are slight differences between the MCMC and

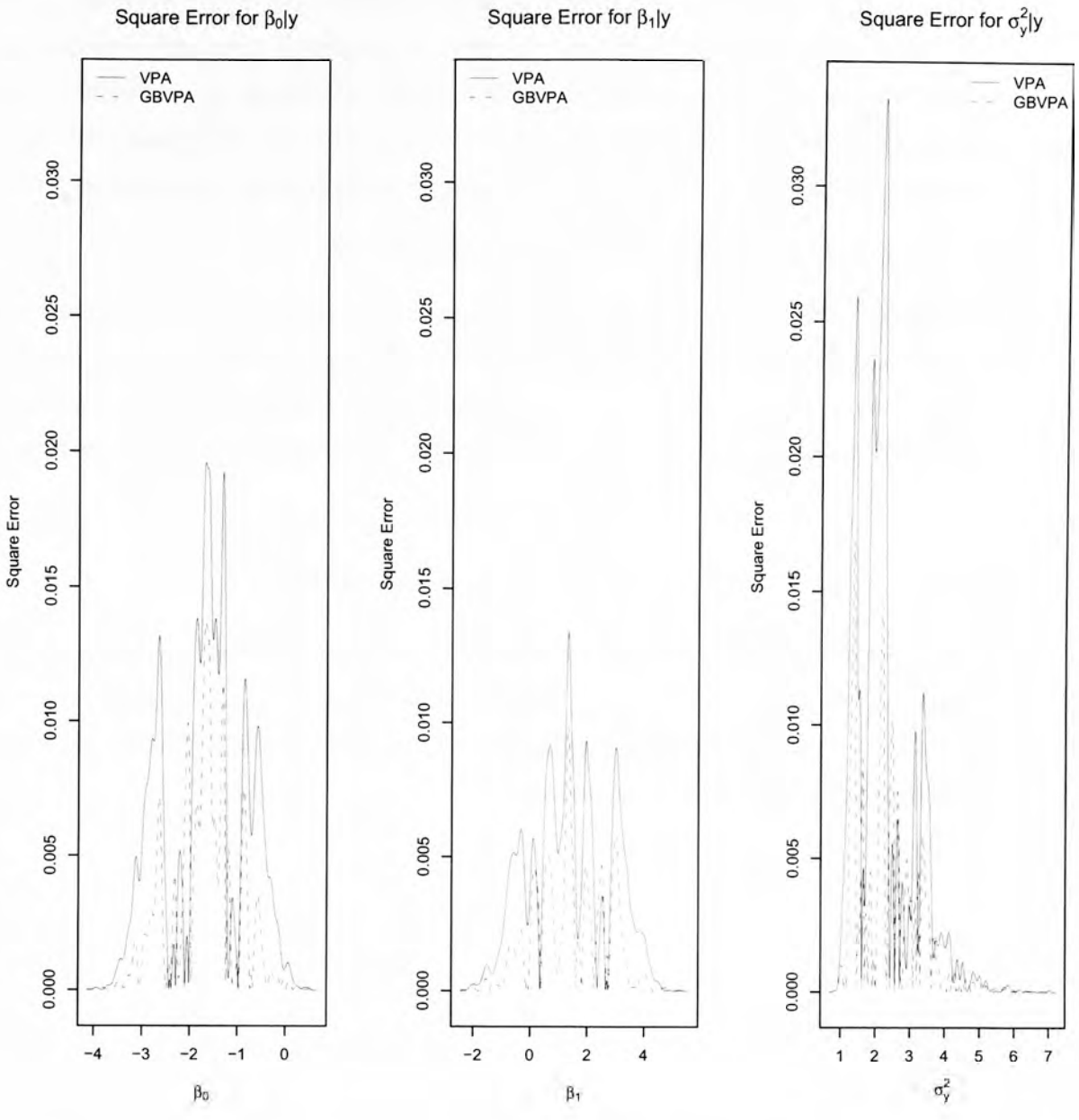


Figure 4.3: *Differences of VPA and GBVPA with kernel density approximations obtained from MCMC posterior samples for Bayesian linear regression model.*

VPA approximations whereas the differences are slightly less noticeable than those GBVPA particularly for the σ_y^2 posterior approximation where the main differences can be put down to inaccuracies due to the use of kernel density estimation. The differences are more visible in Figure 4.3 where the differences between GBVPA and VPA with the MCMC posterior approximation are illustrated.

To compare accuracy of GBVPA and VPA in terms of the ISE we compared the average ISE for 30 simulated datasets for eight combinations of parameters where $\beta_0 \in \{-1, 0\}$, $\beta_1 \in \{0, 1\}$ and $\sigma_y^2 \in \{1/2, 4\}$ and $n = 100$. These results are summarised in Table 4.2. We also note that for individual simulations using $C = 5 \times 10^5$ for posterior samples of size 10^5 the relative ISEs for each parameter were similar.

Unfortunately, for this simple example, based on Figure 4.2, the gains made by using GBVPA instead of VPA are almost imperceptible to the eye. On the other hand looking

at Table 4.2 we see that in terms of the ISE the GBVPA is between about 2.8 and 6.0 times more accurate than VPA for this example. The average time taken to approximate these posteriors using VPA is less than 0.005 seconds, GBVPA for all posteriors took on average about 0.15 seconds while MCMC fits took on average 21.55 seconds. We now consider a more complicated model which give more significant improvements in terms of ISE.

4.6 Bayesian Missing Binary Covariate Model

Consider the complication for a Bayesian linear regression model where a binary covariate is missing completely at random (MCAR). Suppose we have one covariates x for a response y where some of the xs are MCAR and we suspect that x has a fixed probability p of being 0 or 1. Thus we might consider the likelihood approach based on

$$\begin{aligned} y_i|x_i, \boldsymbol{\beta}, \sigma_y^2 &\sim N(\beta_0 + \beta_1 x_i, \sigma_y^2) \\ x_i|p &\sim \text{Bernoulli}(p). \end{aligned} \tag{4.38}$$

For convenience we make the partition $\mathbf{x} = (\mathbf{x}_{obs}, \mathbf{x}_{mis})$ where $\mathbf{x}_{obs} = (\mathbf{x}_{obs,1}, \ldots, \mathbf{x}_{obs,n_{obs}})$ and $\mathbf{x}_{mis} = (\mathbf{x}_{mis,1}, \ldots, \mathbf{x}_{mis,n_{mis}})$ such that $n = n_{obs} + n_{mis}$. Similarly let $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis})$ be a partition of $\mathbf{y}$ coinciding with the "missingness" of $\mathbf{x}$. Now we place the following additional priors on the $\boldsymbol{\beta}$, σ_y^2 and p

$$\begin{aligned} \boldsymbol{\beta} &\sim N(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \mathbf{I}), \\ \sigma_y^2 &\sim IG(A_y, B_y) \\ \text{and } p &\sim \text{Beta}(1, 1) \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and the prior hyperparameters are $\sigma_{\boldsymbol{\beta}}^2 = \sigma_{\tilde{\boldsymbol{\beta}}}^2 = 10^8$ and $A_y = B_y = 10^{-2}$ to characterise the priors for the parameters $\boldsymbol{\beta}, \sigma_y^2$ and p as vague. We wish to fit posteriors for the parameters of interest $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \sigma_y^2, p)$.

In this model the missing values are discrete and so integrating out the missing xs are replaced by multiple summations. After integrating out σ_y^2 and p the joint likelihood for $\mathbf{y}$, $\mathbf{x}_{obs}$ and $\boldsymbol{\beta}$ is proportional to

$$\begin{aligned} [\mathbf{y}, \mathbf{x}_{obs}, \boldsymbol{\beta}] \propto & \sum_{x_{mis,1}=0}^{1} \cdots \sum_{x_{mis,n_{mis}}=0}^{1} \Gamma(\mathbf{1}^T\mathbf{x} + 1)\Gamma(n - \mathbf{1}^T\mathbf{x} + 1) \\ & \times \exp\left\{ -\frac{\|\boldsymbol{\beta}\|^2}{2\sigma_{\boldsymbol{\beta}}^2} - \left(A_y + \frac{n}{2}\right) \log\left(B_y + \frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2}\right) \right\}. \end{aligned} \tag{4.39}$$

The summation involves $2^{n_{mis}}$ terms. Thus unless n_{mis} is small then summation over all $2^{n_{mis}}$ values of $\mathbf{x}_{mis}$ is not computationally feasible. Similar models are fitted using MCMC by Ibrahim, Chen & Lipsitz (2001) and a combination of the EM algorithm and PQL approximations by French & Wand (2004). Instead we pursue variational approximations.

Variational approximations have recently been applied to a number of missing value problems in several contexts (MacKay, 1997; Attias, 1999, 2000; Ghahramani & Beal, 2000;

CASE $(\beta_0, \beta_1, \sigma_y^2)$	Median $10^3 \times$ISE for VPA of β_0	Median $10^3 \times$ISE GBVPA of of β_0	Ratio β_0	Median $10^3 \times$ISE for VPA of β_1	Median $10^3 \times$ISE for GBVPA of β_1	Ratio β_1	Median $10^3 \times$ISE for VPA of σ_y^2	Median $10^3 \times$ISE for GBVPA of σ_y^2	Ratio σ_y^2	Mean Time (s) VPA	Mean Time (s) GBVPA	Mean Time (s) MCMC
$(-1, 0, 1/2)$	0.5174	0.1844	2.8059	0.2615	0.0760	3.4408	1.4913	0.4001	3.7273	<0.005	0.15	21.53
$(0, 0, 1/2)$	0.1774	0.0632	2.8070	0.0896	0.0260	3.4462	0.1753	0.0288	6.0868	0.01	0.15	21.49
$(-1, 1, 1/2)$	0.4953	0.1765	2.8062	0.2503	0.0727	3.4429	1.3666	0.2556	5.3466	<0.005	0.15	21.59
$(0, 1, 1/2)$	0.1842	0.0656	2.8079	0.0931	0.0270	3.4481	0.1890	0.0311	6.0772	0.01	0.15	21.53
$(-1, 0, 2)$	0.4976	0.1773	2.8065	0.2515	0.0730	3.4452	1.3794	0.2626	5.2529	0.01	0.14	21.63
$(0, 0, 2)$	0.1774	0.0632	2.8070	0.0897	0.0261	3.4368	0.1754	0.0288	6.0903	<0.005	0.16	21.58
$(-1, 1, 2)$	0.5202	0.1854	2.8058	0.2629	0.0764	3.4411	1.5077	0.4481	3.3647	<0.005	0.15	21.50
$(0, 1, 2)$	0.1798	0.0641	2.8050	0.0908	0.0264	3.4394	0.1800	0.0296	6.0811	<0.005	0.16	21.58
COMBINED	0.3172	0.1130	2.8071	0.1603	0.0466	3.4399	0.5968	0.0991	6.0222	>0.005	0.15	21.55

Table 4.5.2: *Integrated Square Errors (ISE, see equation (4.34)) and times for variational posterior approximations (VPA) and grid based variational posterior approximations for Bayesian linear regression model (see Section 4.5). One hundred trials of points* (y_i, x_i), $1 \leq i \leq n$ *were simulated from* $x_i \sim Unif(0, 1)$ *and* $y_i \sim N(\beta_0 + \beta_1 x, \sigma_y^2)$ *where* $n = 100$ *and the values for* $(\beta_0, \beta_1, \sigma_y^2)$ *are in the first column.*

Penny & Roberts, 2000; Humphreys & Titterington, 2000, 2001; Beal & Ghahramani, 2002; Beal, 2003; Celeux, Forbes, Robert & Titterington, 2006; Consonni & Marin, 2007). The approximations in the above papers were shown to be practically efficient and effective. Few theoretical results are available although Hall *et al.* (2002) and Wang & Titterington (2004) have been able to prove some important results. In particular Hall *et al.* (2002) were able to show for certain Markov models, the parameter estimator obtained by maximising the variational lower bound function is asymptotically consistent, provided the proportion of all values that are missing tends to zero. Wang & Titterington (2004) investigated the consistency properties of both mean field and variational Bayesian estimators in the context of linear state space models and proved that the mean-field approximations are asymptotically consistent when the variances of the noise variables in the system are sufficiently small.

4.6.1 *Variational Posterior Approximations*

Consider the task of approximating posteriors using VPA. Choosing the density transform which mirrors the priors, i.e using $\delta(\boldsymbol{\beta}, p, \sigma_y^2, \mathbf{x}_{mis}) = \delta_{\boldsymbol{\beta}}(\boldsymbol{\beta})\delta_{\sigma_y^2}(\sigma_y^2)\delta_p(p)\delta_{\mathbf{x}_{mis}}(\mathbf{x}_{mis})$ where

$$\begin{aligned}
\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}_{obs} &\sim_{\delta_{\boldsymbol{\beta}}} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\sigma_y^2|\mathbf{y}, \mathbf{x}_{obs} &\sim_{\delta_{\sigma_y^2}} IG(\alpha_y, \beta_y) \\
p|\mathbf{y}, \mathbf{x}_{obs} &\sim_{\delta_p} \text{Beta}(\alpha_p, \beta_p) \\
x_{mis,i}|\mathbf{y}, \mathbf{x}_{obs} &\sim_{\delta_{\mathbf{x}_{mis}}} \text{Bernoulli}(\rho_i)
\end{aligned}$$

Furthermore note that the form of $\delta(\boldsymbol{\beta}, p, \sigma_y^2, \mathbf{x}_{mis})$ assumed independence of the variables $\boldsymbol{\beta}$, σ_y^2, p and $\mathbf{x}_{mis}$ which is not necessarily the case for the true posterior density $\boldsymbol{\beta}, p, \sigma_y^2, \mathbf{x}_{mis}|\mathbf{y}, \mathbf{x}_{obs}$. The additional variational parameters are $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_y, \beta_y, \alpha_p, \beta_p, \boldsymbol{\rho})$ where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{n_{mis}})$.

The density transform of the marginal log-likelihood is given by

$$\begin{aligned}
[\mathbf{y}, \mathbf{x}_{obs}; \boldsymbol{\xi}]_L &= \mathbb{E}_\delta(\log[\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma_y^2]) + \mathbb{E}_\delta(\log[\boldsymbol{\beta}]) + \mathbb{E}_\delta(\log[\sigma_y^2]) + \mathbb{E}_\delta(\log[\mathbf{x}|p]) + \mathbb{E}_\delta(\log[p]) \\
&\quad + \mathcal{H}_{\delta_{\boldsymbol{\beta}}} + \mathcal{H}_{\delta_{\sigma_y^2}} + \mathcal{H}_{\delta_p} + \mathcal{H}_{\delta_{\mathbf{x}_{mis}}}
\end{aligned}$$

where the relevant expectations and entropies are given by

$$
\begin{aligned}
\mathbb{E}_\delta(\log[\mathbf{y}|\mathbf{x},\boldsymbol{\beta},\sigma_y^2]) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\left(\log(\beta_y) - \psi(\alpha_y)\right) \\
&\quad -\frac{\alpha_y}{\beta_y}\cdot\frac{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\widehat{\mathbf{X}}\boldsymbol{\mu} + \boldsymbol{\mu}^T\widehat{\mathbf{X}^T\mathbf{X}}\boldsymbol{\mu} + \mathrm{tr}\left(\boldsymbol{\Sigma}\widehat{\mathbf{X}^T\mathbf{X}}\right)}{2}, \\
\mathbb{E}_\delta(\log[\boldsymbol{\beta}]) &= -\log(2\pi\sigma_\beta^2) - \frac{\|\boldsymbol{\mu}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma})}{2\sigma_\beta^2}, \\
\mathbb{E}_\delta(\log[\sigma_y^2]) &= A_y\log(B_y) - \log\Gamma(A_y) - (A_y+1)(\log(\beta_y) - \psi(\alpha_y)) - B_y\frac{\alpha_y}{\beta_y}, \\
\mathbb{E}_\delta(\log[\mathbf{x}|p]) &= (\mathbf{1}^T\widehat{\mathbf{x}})\psi(\alpha_p) + (n - \mathbf{1}^T\widehat{\mathbf{x}})\psi(\beta_p) - n\psi(\alpha_p+\beta_p), \\
\mathbb{E}_\delta(\log[p]) &= 0, \\
\mathcal{H}_{\delta_\beta} &= \frac{1}{2}\log|2e\pi\boldsymbol{\Sigma}|, \\
\mathcal{H}_{\delta_{\sigma_y^2}} &= \alpha_\epsilon + \log(\beta_y) + \log\Gamma(\alpha_y) - (\alpha_y+1)\psi(\alpha_y), \\
\mathcal{H}_{\delta_p} &= \log B(\alpha_p,\beta_p) - (\alpha_p-1)\psi(\alpha_p) - (\beta_p-1)\psi(\beta_p) \\
&\quad +(\alpha_p+\beta_p-2)\psi(\alpha_p+\beta_p) \\
\text{and } \mathcal{H}_{\delta_{\mathbf{x}_{mis}}} &= -\boldsymbol{\rho}^T\log(\boldsymbol{\rho}) - (\mathbf{1}-\boldsymbol{\rho})^T\log(\mathbf{1}-\boldsymbol{\rho})
\end{aligned}
\tag{4.40}
$$

where $B(\alpha_p,\beta_p) = \frac{\Gamma(\alpha_p)\Gamma(\beta_p)}{\Gamma(\alpha_p+\beta_p)}$ is the Beta function, $\widehat{\mathbf{x}} = (\mathbf{x}_{obs},\boldsymbol{\rho})$, $\widehat{\mathbf{X}} = [\mathbf{1},\widehat{\mathbf{x}}]$ and

$$
\widehat{\mathbf{X}^T\mathbf{X}} = \begin{bmatrix} n & \mathbf{1}^T\mathbf{x}_{obs} + \mathbf{1}^T\boldsymbol{\rho} \\ \mathbf{1}^T\mathbf{x}_{obs} + \mathbf{1}^T\boldsymbol{\rho} & \mathbf{x}_{obs}^T\mathbf{x}_{obs} + \mathbf{1}^T\boldsymbol{\rho} \end{bmatrix} = \begin{bmatrix} n & \mathbf{1}^T\widehat{\mathbf{x}} \\ \mathbf{1}^T\widehat{\mathbf{x}} & \mathbf{1}^T\widehat{\mathbf{x}} \end{bmatrix}
$$

which follows from $\mathbf{x}_{obs}$ being binary (implying $\mathbf{x}_{obs}^T\mathbf{x}_{obs} = \mathbf{1}^T\mathbf{x}_{obs}$). Here have used the facts $\mathbb{E}_\delta(\log(p)) = \psi(\alpha_p) - \psi(\alpha_p+\beta_p)$ and $\mathbb{E}_\delta(\log(1-p)) = \psi(\beta_p) - \psi(\alpha_p+\beta_p)$. These may be verified either by direct integration, integration by parts or by using a symbolic computing package.

By taking derivatives with respect to all variational parameters and equating to zero, fixed point updates for each parameter can be derived.

$$
\begin{aligned}
\boldsymbol{\Sigma} &:= \left(\frac{\alpha_y}{\beta_y}\widehat{\mathbf{X}^T\mathbf{X}} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1}, \\
\boldsymbol{\mu} &:= \frac{\alpha_y}{\beta_y}\boldsymbol{\Sigma}\widehat{\mathbf{X}}^T\mathbf{y}, \\
\alpha_y &:= A_y + \frac{n}{2}, \\
\beta_y &:= B_y + \left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\widehat{\mathbf{X}}\boldsymbol{\mu} + \boldsymbol{\mu}^T\widehat{\mathbf{X}^T\mathbf{X}}\boldsymbol{\mu} + \mathrm{tr}\left(\boldsymbol{\Sigma}\widehat{\mathbf{X}^T\mathbf{X}}\right)\right)/2, \\
\alpha_p &:= 1 + \mathbf{1}^T\widehat{\mathbf{x}}, \\
\beta_p &:= 1 + n - \mathbf{1}^T\widehat{\mathbf{x}}, \\
\rho_i &:= \frac{1}{1+\exp(-\eta_i)}, \\
\eta_i &:= \psi(\alpha_p) - \psi(\beta_p) + \frac{\alpha_y}{\beta_y}\left(y_{mis,i}\mu_1 - \frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}\mathbf{A}\right)\right) \\
\text{and } \mathbf{A} &= \frac{\partial\widehat{\mathbf{X}^T\mathbf{X}}}{\partial\rho_i} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.
\end{aligned}
\tag{4.41}
$$

These updates are applied until the variational parameters converge.

This form of optimisation is computationally more efficient than using Newton-Raphson or quasi-Newton updates since, if n_{mis} is large, then the Hessian (or approximate Hessian matrix for quasi-Newton methods) is of size $O(n_{mis}) \times O(n_{mis})$ whereas the space cost for the above iterations is $O(n_{mis})$ and updates can be performed in $O(n)$ time. Unfortunately, these iterates can sometimes converge slowly. Using these equations, it is relatively easy to calculate the posterior distributions for $\boldsymbol{\beta}$, σ_y^2 and p. An alternative method which might used is the steepest descent approach.

Suppose that $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\alpha}_y, \widehat{\beta}_y, \widehat{\alpha}_p, \widehat{\beta}_p, \widehat{\boldsymbol{\rho}})$ are the values of the variational parameters at convergence. Then the approximate marginal posterior densities are

$$\begin{aligned} \beta_i|\mathbf{y}, \mathbf{x}_{obs} &\sim_{\delta_\beta} N(\widehat{\mu}_i, \widehat{\Sigma}_{ii}) \\ \sigma_y^2|\mathbf{y}, \mathbf{x}_{obs} &\sim_{\delta_{\sigma_y^2}} IG(\widehat{\alpha}_y, \widehat{\beta}_y) \\ p|\mathbf{y}, \mathbf{x}_{obs} &\sim_{\delta_p} \text{Beta}(\widehat{\alpha}_p, \widehat{\beta}_p) \end{aligned} \tag{4.42}$$

and the posterior means for the parameters $(\boldsymbol{\beta}, \sigma_y^2, p)$ can be approximated using the formula for the means of the approximate posterior densities. In this case

$$\begin{aligned} \mathbb{E}_\delta(\boldsymbol{\beta}) &= \widehat{\boldsymbol{\mu}}, \\ \mathbb{E}_\delta(\sigma_y^2) &= \frac{\widehat{\beta}_y}{\widehat{\alpha}_y - 1} \\ \text{and } \mathbb{E}_\delta(p) &= \frac{\widehat{\alpha}_p}{\widehat{\alpha}_p + \widehat{\beta}_p}. \end{aligned}$$

We will now consider a number of simulated experiments based on data points (y_i, x_i), $1 \leq i \leq n$ where $x_i \sim \text{Bern}(p)$ and $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_y^2)$ with a given percentage of x_is removed completely at random. The approximate marginal posterior densities (4.42) are illustrated in Figure 4.4 where we notice that the posterior variances are all underestimated.

4.6.2 *Grid Based Variational Posterior Approximation for β_i*

For a fixed $\beta_i = \widehat{\beta}_i$ we use $\beta_{-i} \sim N(0, \sigma_\beta^2)$, the density transform $\delta(\beta_{-i}, p, \sigma_y^2, \mathbf{x}_{mis}) = \delta_{\beta_{-i}}(\beta_{-i})\delta_{\sigma_y^2}(\sigma_y^2)\delta_p(p)\delta_{\mathbf{x}_{mis}}(\mathbf{x}_{mis})$, and $\beta_{-i}|\mathbf{y}, \mathbf{x}_{obs} \sim_\delta N(\mu_{-i}, \sigma^2)$. Applying this density transform we obtain

$$\begin{aligned} [\mathbf{y}, \mathbf{x}_{obs}, \beta_i; \boldsymbol{\xi}]_L &= \mathbb{E}_\delta(\log[\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma_y^2]) + \mathbb{E}_\delta(\log[\beta_{-i}]) + \mathbb{E}_\delta(\log[\sigma_y^2]) + \mathbb{E}_\delta(\log[\mathbf{x}|p]) \\ &\quad + \mathcal{H}_{\delta_{\beta_{-i}}} + \mathcal{H}_{\delta_{\sigma_y^2}} + \mathcal{H}_{\delta_p} + \mathcal{H}_{\delta_{\mathbf{x}_{mis}}} \end{aligned}$$

where the relevant expectations and entropies are given in (4.40) except

$$\begin{aligned} \mathbb{E}_\delta \log[\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma_y^2] &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}(\log(\beta_y) - \psi(\alpha_y)) \\ &\quad - \frac{\alpha_y}{\beta_y} \cdot \frac{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\widehat{\mathbf{X}}\boldsymbol{\mu} + \boldsymbol{\mu}^T\widehat{\mathbf{X}^T\mathbf{X}}\boldsymbol{\mu} + \sigma^2[\widehat{\mathbf{X}^T\mathbf{X}}]_{-i,-i}}{2}, \\ \mathbb{E}_\delta \log[\beta_{-i}] &= -\frac{1}{2}\log(2\pi\sigma_\beta^2) - \frac{\mu_{-i}^2 + \sigma^2}{2\sigma_\beta^2} \\ \text{and } \mathcal{H}_{\delta_{\beta_{-i}}} &= \frac{1}{2}\log(2e\pi\sigma^2) \end{aligned}$$

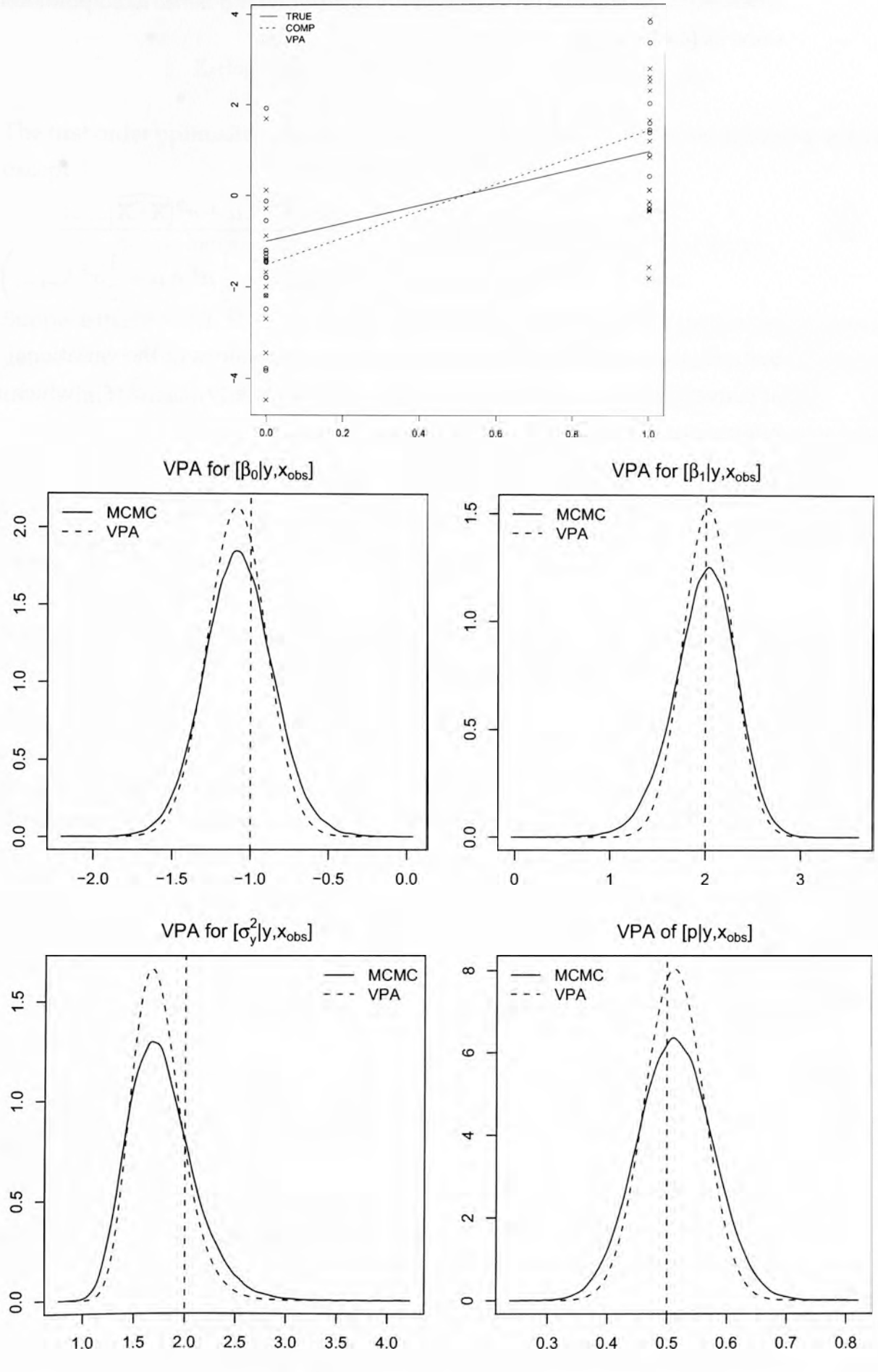


Figure 4.4: *Fitted line (top panel) and variational posterior approximations (bottom panels) for the Bayesian missing binary covariate model. Data points* (y_i, x_i), $1 \leq i \leq n$ *were generated where* $x_i \sim Bern(p)$ *and* $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_y^2)$ *where* $n = 100, p = 0.5, \beta_0 = -1, \beta_1 = 2, \sigma_y^2 = 2$ *and then* 50% *of the* x_i*s were removed at random. The dashed vertical lines represent the "true" values used in the simulation.*

where $\mu_i = \widehat{\beta}_i$. The first order optimality conditions and hence fixed point iterates are the same as (4.41) except

$$\begin{aligned}
\sigma^2 &:= \left(\frac{\alpha_y}{\beta_y}\widehat{[\mathbf{X}^T\mathbf{X}]}_{-i,-i} + \sigma_\beta^{-2}\right)^{-1}, \\
\mu_{-i} &:= \frac{\alpha_y}{\beta_y}\sigma^2\left([\widehat{\mathbf{X}}^T\mathbf{y}]_{-i} - \widehat{[\mathbf{X}^T\mathbf{X}]}_{i,-i}\widehat{\beta}_i\right), \\
\beta_y &:= B_y + \frac{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\widehat{\mathbf{X}}\boldsymbol{\mu} + \boldsymbol{\mu}^T\widehat{\mathbf{X}^T\mathbf{X}}\boldsymbol{\mu} + \sigma^2\widehat{[\mathbf{X}^T\mathbf{X}]}_{-i,-i}}{2} \\
\text{and } \eta_i &:= \psi(\alpha_p) - \psi(\beta_p) + \frac{\alpha_y}{\beta_y}\left(y_{mis,i}\mu_1 - \frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} - \frac{1}{2}\sigma^2\mathbf{A}_{-i,-i}\right).
\end{aligned}$$

Suppose that $\widehat{\boldsymbol{\xi}} = (\widehat{\mu}_{-i}, \widehat{\sigma}^2, \widehat{\alpha}_y, \widehat{\beta}_y, \widehat{\alpha}_p, \widehat{\beta}_p, \widehat{\boldsymbol{\rho}})$ are the values of the variational parameters at convergence. For a fixed $\beta_i = \widehat{\beta}_i$ we can calculate $[\mathbf{y}, \mathbf{x}_{obs}, \widehat{\beta}_i; \widehat{\boldsymbol{\xi}}]_L$ which is sufficient information to implement a GBVPA for $[\beta_i|\mathbf{y}]$.

4.6.3 *Grid Based Variational Posterior Approximation for σ_y^2*

For a fixed $\sigma_y^2 = \widehat{\sigma}_y^2$ we use the density transform $\delta(\boldsymbol{\beta}, p, \mathbf{x}_{mis}) = \delta_{\beta_{-i}}(\beta_{-i})\delta_p(p)\delta_{\mathbf{x}_{mis}}(\mathbf{x}_{mis})$ to obtain a lower bound for the joint likelihood for $\mathbf{y}$, $\mathbf{x}_{obs}$ and $\sigma_y^2 = \widehat{\sigma}_y^2$ given by

$$\begin{aligned}
\log[\mathbf{y}, \mathbf{x}_{obs}, \widehat{\sigma}_y^2; \boldsymbol{\xi}]_L &= \mathbb{E}_\delta(\log[\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \widehat{\sigma}_y^2]) + \mathbb{E}_\delta(\log[\boldsymbol{\beta}]) + \log[\widehat{\sigma}_y^2] + \mathbb{E}_\delta(\log[\mathbf{x}|p]) + \mathbb{E}_\delta(\log[p]) \\
&\quad + \mathcal{H}_{\delta_{\beta_{-i}}} + \mathcal{H}_{\delta_p} + \mathcal{H}_{\delta_{\mathbf{x}_{mis}}}
\end{aligned}$$

where the relevant expectations and entropies are given in (4.40) except

$$\begin{aligned}
\mathbb{E}_\delta(\log[\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \widehat{\sigma}_y^2]) &= -\frac{n}{2}\log(2\pi\widehat{\sigma}_y^2) - \frac{\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\widehat{\mathbf{X}}\boldsymbol{\mu} + \boldsymbol{\mu}^T\widehat{\mathbf{X}^T\mathbf{X}}\boldsymbol{\mu} + \operatorname{tr}\left(\boldsymbol{\Sigma}\widehat{\mathbf{X}^T\mathbf{X}}\right)}{2\widehat{\sigma}_\epsilon^2} \\
\text{and } \log[\sigma_y^2] &= A_y\log(B_y) - \log\Gamma(A_y) - (A_y+1)\log(\widehat{\sigma}_y^2) - B_y\widehat{\sigma}_y^{-2}.
\end{aligned}$$

The first order optimality conditions and hence fixed point iterates are the same as (4.41) except

$$\begin{aligned}
\boldsymbol{\Sigma} &:= \left(\widehat{\sigma}_y^{-2}\widehat{\mathbf{X}^T\mathbf{X}} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1}, \\
\boldsymbol{\mu} &:= \widehat{\sigma}_y^{-2}\boldsymbol{\Sigma}\widehat{\mathbf{X}}^T\mathbf{y} \\
\text{and } \eta_i &:= \psi(\alpha_p) - \psi(\beta_p) + \frac{\alpha_y}{\beta_y}\left(y_{mis,i}\mu_1 - \frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}\mathbf{A})\right).
\end{aligned}$$

Suppose that $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\alpha}_p, \widehat{\beta}_p, \widehat{\boldsymbol{\rho}})$ are the values of the variational parameters at convergence. For a fixed $\sigma_y^2 = \widehat{\sigma}_y^2$ can calculate $\log[\mathbf{y}, \mathbf{x}_{obs}, \widehat{\sigma}_y^2; \widehat{\boldsymbol{\xi}}]_L$ which is sufficient information to implement a GBVPA for $[\sigma_y^2|\mathbf{y}, \mathbf{x}_{obs}]$.

4.6.4 *Grid Based Variational Posterior Approximation for p*

For a fixed $p = \widehat{p}$ we use the density transform $\delta(\boldsymbol{\beta}, \sigma_y^2, \mathbf{x}_{mis}) = \delta_{\boldsymbol{\beta}}(\boldsymbol{\beta})\delta_{\sigma_y^2}(\sigma_y^2)\delta_{\mathbf{x}_{mis}}(\mathbf{x}_{mis})$ to obtain a lower bound for the joint likelihood for $\mathbf{y}$, $\mathbf{x}_{obs}$ and $p = \widehat{p}$ given by

$$\begin{aligned}
[\mathbf{y}, \mathbf{x}_{obs}, \widehat{p}; \boldsymbol{\xi}]_L &= \mathbb{E}_\delta(\log[\mathbf{y}|\mathbf{x}, \boldsymbol{\beta}, \sigma_y^2]) + \mathbb{E}_\delta(\log[\boldsymbol{\beta}]) + \mathbb{E}_\delta(\log[\sigma_y^2]) + \mathbb{E}_\delta(\log[\mathbf{x}|p]) \\
&\quad + \mathcal{H}_{\delta_{\boldsymbol{\beta}}} + \mathcal{H}_{\delta_{\sigma_y^2}} + \mathcal{H}_{\delta_{\mathbf{x}_{mis}}}
\end{aligned}$$

where the relevant expectations and entropies are given in (4.40) except

$$\mathbb{E}_\delta(\log[\mathbf{x}|p]) \quad = (\mathbf{1}^T\widehat{\mathbf{x}})\log(p) + (n - \mathbf{1}^T\widehat{\mathbf{x}})\log(1-p).$$

The first order optimality conditions and hence fixed point iterates are the same as (4.41) except

$$\eta_i \quad := \log(p) - \log(1-p) + \frac{\alpha_y}{\beta_y}\left(y_{mis,i}\mu_1 - \frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}\mathbf{A}\right)\right).$$

Suppose that $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\alpha}_y, \widehat{\beta}_y, \widehat{\boldsymbol{\rho}})$ are the values of the variational parameters at convergence. For a fixed $\sigma_y^2 = \widehat{\sigma}_y^2$ we can calculate $\log[\mathbf{y}, \mathbf{x}_{obs}, \widehat{p}; \widehat{\boldsymbol{\xi}}]_L$ which is sufficient information to implement a GBVPA for $[p|\mathbf{y}, \mathbf{x}_{obs}]$.

Figure 4.5 illustrates the marginal posterior densities for the Bayesian binary missing value model for the same setting in Figure 4.4 using GBVPA. We notice from this figure that the GBVPAs are much closer to the kernel densities estimated from posterior samples via MCMC than for VPA.

4.6.5 Numerical Comparisons

In order to test the effectiveness of the variational approximations for the Bayesian missing covariate model we consider randomly generated datasets where n points (y_i, x_i), $1 \leq i \leq n$ are generated from (4.38) where $p = p^*$, $\beta_0 = \beta_0^*$, $\beta_1 = \beta_1^*$ and $\sigma_y^2 = \sigma_y^{2*}$ are fixed and a fixed percentage of the xs are removed completely at random. The three methods we will compare are: (a) the variational approximation to the Bayesian linear regression model developed in Section 4.5 using only complete cases (CC), (b) the Bayesian missing covariate model fitted using MCMC with `WinBUGs`, (c) the variational posterior approximation to the Bayesian missing covariate model (VPA) developed in Section 4.6.1 and (d) the grid based variational posterior approximation developed in Sections 4.6.2–2.6.4. Note that when missingness is MCAR then the CC analysis is unbiased, although when there are a large number of missing values the loss of efficiency can be substantial (Little & Rubin, 2002).

We compare approximated posterior means using the mean square error (MSE) for the CC, MCMC, VPA and GBVPA approaches over s randomly generated datasets, where for β_0 the MSE is given by

$$\mathbb{E}\left((\widehat{\mu}_0 - \beta_0^*)^2\right) \approx s^{-1}\sum_{i=1}^{s}(\widehat{\mu}_0^{(i)} - \beta_0^*)^2$$

and $\widehat{\mu}_0^{(i)}$ is the ith approximation of posterior mean for β_0. We use a 5,000 point composite trapezoid rule to approximate posterior means for the GBVPA approach.

Table 4.6.3 contains MSE for the approximated posterior means and average times for the CC, MCMC, VPA and GBVPA approaches for $s = 50$ randomly generated datasets with $p^* = 0.5$, $\beta_0^* = 0$, $\beta_1^* = 1$ and $\sigma_y^{2*} = 1$. We used $N = 30$ grid points for GBVPA. We see from this table that the MSE for approximated posterior means for the VPA and

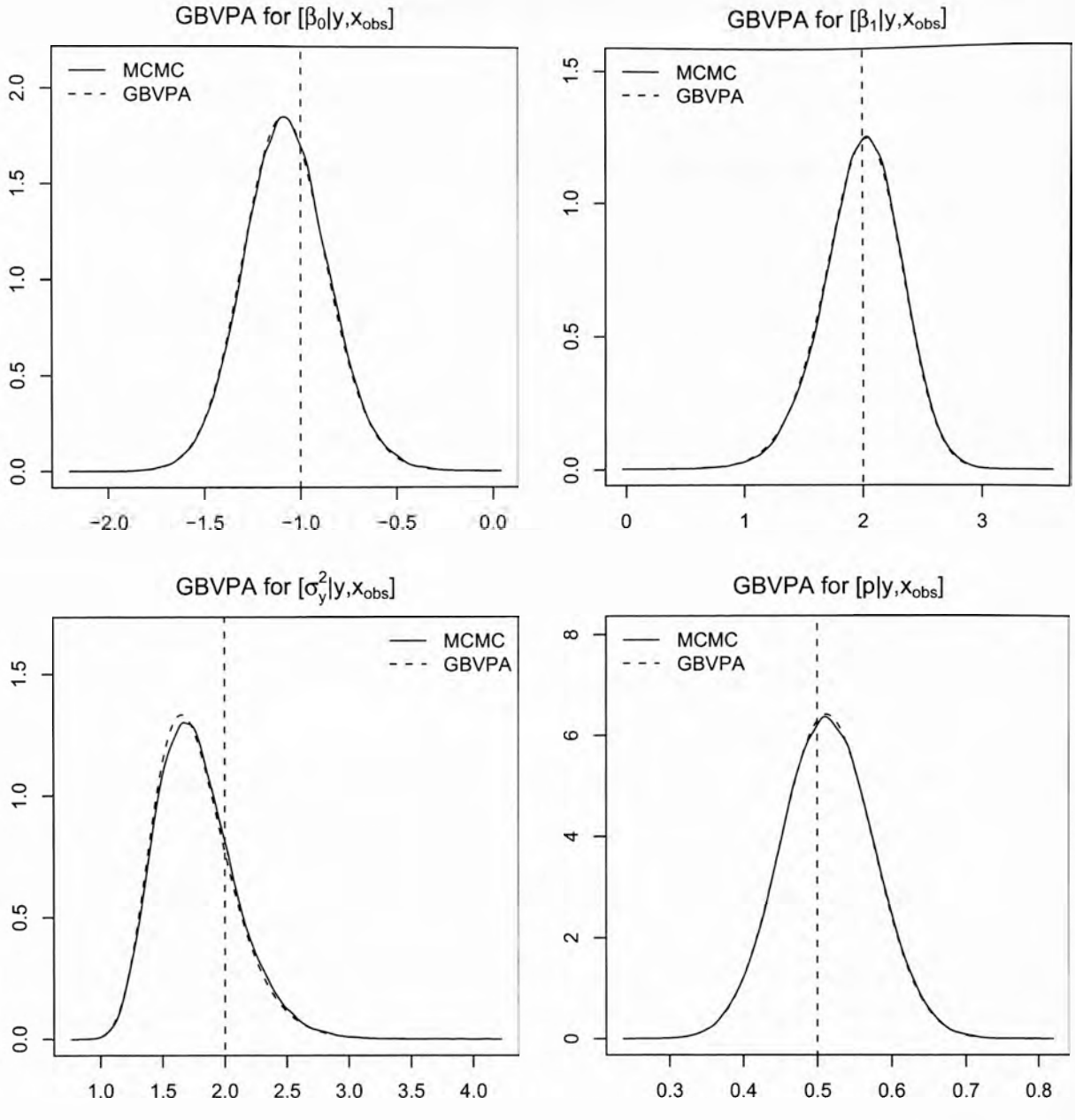


Figure 4.5: *Grid based variational posterior estimates for the Bayesian missing binary covariate model. The dataset for this figure was the same used for Figure 4.4. The dashed vertical lines represent the "true" values used in the simulation.*

GBVPA methods are both comparable with those approximated by MCMC, but that those of GBVPA are closer.

Furthermore we note that the VPA and GBVPA algorithms scale very well to large n. For $n = 10^6$ with 50% of xs randomly removed the VPA algorithm takes about 15 seconds of computing time while GBVPA algorithm takes about 10 minutes and an MCMC approach using `WinBUGS` with using just 10,000 posterior samples took 43 and a half hours of computing time. While it is possible to do much better with custom Markov chain code it is unlikely to take less than 10 minutes for this example.

To compare accuracy of GBVPA and VPA in terms of the ISE, we compared the average ISE for 30 simulated datasets for eight combinations of parameters where $p^* \in \{0.25, 0.75\}$, $\beta_1^* \in \{0, 1\}$, $\sigma_y^{2*} = \{1/2, 2\}$, $n = 200$ and 50% of the xs removed at random.

n	Method	100×MSE for β_0	100×MSE for β_1	100×MSE for σ_y^2	100×MSE for p	Time (s)
100	CC	2.4971	3.6197	2.9929	0.3600	<0.005
	MCMC	1.4549	1.8992	1.0256	0.3733	72.555
	VPA	1.4757	2.1979	2.7703	0.3843	0.03
	GBVPA	1.4317	1.8891	1.1648	0.3760	1.19
200	CC	1.1490	3.4212	1.1509	0.1600	<0.005
	MCMC	0.9058	3.3295	0.4736	0.1230	141.155
	VPA	0.9240	3.3155	1.1560	0.1260	0.03
	GBVPA	0.9089	3.3441	0.4662	0.1240	1.255
400	CC	0.8603	0.9033	0.8245	0.0900	<0.005
	MCMC	0.7909	0.8594	0.3595	0.1203	277.435
	VPA	0.8024	0.8532	0.8187	0.1222	0.03
	GBVPA	0.8011	0.8595	0.3589	0.1213	1.37
800	CC	0.3791	0.2979	0.2125	0.0400	<0.005
	MCMC	0.2384	0.0976	0.1262	0.0427	555.37
	VPA	0.2257	0.1109	0.2139	0.0426	0.03
	GBVPA	0.2342	0.0942	0.1296	0.0428	1.61
1600	CC	0.1192	0.3407	0.0948	0.0077	<0.005
	MCMC	0.0959	0.2728	0.0694	0.0106	1167.44
	VPA	0.0990	0.2834	0.0923	0.0108	0.04
	GBVPA	0.0967	0.2662	0.0694	0.0108	2.09
3200	CC	0.0408	0.1593	0.1145	0.0049	<0.005
	MCMC	0.0295	0.0962	0.0479	0.0042	2932.035
	VPA	0.0296	0.0936	0.0476	0.0042	0.06
	GBVPA	0.0296	0.0946	0.0476	0.0042	3.105

Table 4.6.3: *A comparison of posterior mean square errors and times for the Bayesian binary missing value problem using complete cases (CC), MCMC, VPA and GBVPA. Data* (y_i, x_i), $\leq i \leq n$ *is simulated from where* $x_i \sim Bern(p)$ *and* $y_i \sim N(\beta_0^* + \beta_1^* x_i, \sigma_y^{2*})$ *where the true values are* $p^* = 0.5, \beta_0^* = 0, \beta_1^* = 1, \sigma_y^{2*} = 1$ *and* 50% *of the* x*s are removed completely at random.*

For GBVPA we used $N = 30$ grid points. These results are summarised in Table 4.6.4. From this table we see that in terms of ISE the GBVPA is on average 206.93 time, 363.38, 4.66 and 2944.62 times more accurate for the parameters β_0, β_1, σ_y^2 and p respectively. This represents, for this case, GBVPA offers vast improvement over VPA.

4.7 Conclusion

Efficient and accurate methods for approximation of integrals or summations which are computationally or algebraically intractable are one of the most common problems in statistics. Variational methods are a promising class of new approximations which may be used on a variety of statistical integrals. One such method VEM is a generalisation of the EM algorithm which is typically fast, flexible and may be used to simplify EM calculations.

An important application of these variational methods is the efficient approximation of posteriors in Bayesian analysis. As noted by Humphreys & Titterington (2001), Wang & Titterington (2005) and Consonni & Marin (2007) the covariance matrices corresponding to the variational approximations are typically 'too small' compared with those for

CASE (p, β_1, σ_y^2)	Median $10\times$ISE for VPA of β_0	Median $10\times$ISE for GBVPA of β_0	Ratio β_0	Median $10\times$ISE for VPA of β_1	Median $10\times$ISE for GBVPA of β_1	Ratio β_1	Median $10\times$ISE for VPA of σ_y^2	Median $10\times$ISE for GBVPA of σ_y^2	Ratio σ_y^2	Median $10\times$ISE for VPA of p	Median $10\times$ISE for GBVPA of p	Ratio p	Mean Time (s) VPA	Mean Time (s) GBVPA	Mean Time (s) MCMC
$(0.25, 0, 1/2)$	0.4814	0.0050	96.28	1.7342	0.0030	578.07	0.0676	0.0487	1.39	2.5088	0.0034	737.88	0.02	1.11	137.82
$(0.75, 0, 1/2)$	0.2455	0.0022	111.59	0.8663	0.0014	618.79	0.0190	0.0128	1.48	6.8166	0.0031	2198.90	0.02	1.19	137.61
$(0.25, 1, 1/2)$	0.5597	0.0068	82.31	0.7387	0.0028	263.82	1.0617	0.0175	60.67	3.7004	0.0019	1947.58	0.02	1.16	142.35
$(0.75, 1, 1/2)$	0.2661	0.0026	102.35	0.6751	0.0020	337.55	0.0576	0.0023	25.04	6.8588	0.0031	2212.52	0.03	1.24	138.35
$(0.25, 0, 2)$	1.3778	0.0025	551.12	1.7122	0.0025	684.88	0.0827	0.0525	1.58	10.3969	0.0028	3713.18	0.02	1.07	141.64
$(0.75, 0, 2)$	0.6878	0.0012	573.17	0.8676	0.0014	619.71	0.0184	0.0123	1.50	8.3542	0.0029	2880.76	0.02	1.17	141.62
$(0.25, 1, 2)$	0.8169	0.0093	87.84	0.7056	0.0030	235.20	1.0950	0.0210	52.14	9.2063	0.0018	5114.61	0.02	1.16	147.16
$(0.75, 1, 2)$	0.5704	0.0014	407.43	0.6857	0.0027	253.96	0.0621	0.0023	27.01	8.5316	0.0024	3554.83	0.03	1.23	143.91
COMBINED	0.6001	0.0029	206.93	0.8721	0.0024	363.38	0.0774	0.0166	4.66	7.6560	0.0026	2944.62	0.02	1.16	142.01

Table 4.6.4: *Integrated Square Errors (ISE, see equation (4.34)) and times for variational posterior approximations (VPA) and grid based variational posterior approximations for Bayesian binary missing value model (see Section 4.6). One hundred trials of points* (y_i, x_i), $1 \leq i \leq n$ *were simulated from* $x_i \sim Bern(p^*)$ *and* $y_i \sim N(\beta_0^* + \beta_1^* x_i, \sigma_y^{2*})$ *where* $\beta_0^* = 0$, $n = 200$ *and the values for* $(p^*, \beta_1^*, \sigma_y^{2*})$ *are fixed and given in the first column. Finally 50% of the* x_i*s were removed at random. The final row, COMBINED, contains column values averaged over all* (p, β_1, σ_y^2) *settings.*

the MLE, so that resulting interval estimates for the parameters will be too narrow. We have shown in two examples that the GBVPA algorithm developed in this chapter can improve on the standard VPA algorithm, sometimes dramatically. While we have shown the GBVPA approach to be fast and scalable to large datasets the GBVPA algorithm relies on multiple solutions of algorithms similar to VPA, and improvements may be found which reduce the number of times these algorithms are run.

CHAPTER 5

Variational Approximations for Generalized Linear Mixed Models

5.1 Introduction

The success of linear mixed models (LMMs) in handling complications due to messy data has led to its widespread use in many fields. In longitudinal studies LMMs can be used, for example, to handle the statistical complication of correlation in grouped data leading to simple, hierarchical, crossed and nested random effect models (Verbeke & Molenberghs, 2000; McCulloch & Searle, 2001). Similarly LMMs can be used for function approximation including scatterplot smoothing, random coefficient and kriging models (Ruppert, Wand & Carroll, 2003). The extension of these models to generalised responses, called generalised linear mixed models (GLMMs), are also extremely useful (Zhao, Staudenmayer, Coull & Wand, 2006). Unfortunately, the expression for the marginal likelihood for GLMMs involves an integral with no (known) closed form. The usefulness of GLMMs, along with the inherent difficulties involved, have driven an enormous volume of research in the area over the past several decades.

The appearance of analytically intractable integrals in the marginal likelihood for GLMMs means we need to use approximations to proceed. Approximations include Laplace-like approximations such as penalised quasi-likelihood (PQL, Breslow & Clayton, 1993), Gauss-Hermite quadrature (Naylor & Smith, 1982; Liu & Pierce, 1994) and Monte Carlo methods (Gelman, Carlin, Stern & Rubin, 1995; Clayton, 1996; Robert & Casella, 1999; Gilks, Richardson & Spiegelhalter, 1996). Each of these methods of approximation have computational shortcomings associated with them. Laplace and related approximations do not scale well to higher orders of accuracy, Gauss-Hermite does not scale well to high dimensional integrals and Monte Carlo methods suffer from the problems of the slowness and difficulties accessing convergence (see Section 1.3.1 for a summary). Excellent overviews of existing approximations include McCulloch & Searle (2001, Chapter 10) and Tuerlinckx, Rijmen, Verbeke & de Boeck (2006).

Variational approximations are a class of analytic approximations which offer a fresh alternative for fitting GLMMs and as such, as previously argued, can be useful in a number of contexts. Since analytic approximations are typically faster than numerical approximation alternatives they can be used (i) as a starting point for other more accurate algorithms, (ii) as the basis for a model selection procedure and (iii) when criteria other than the accuracy of approximating the marginal likelihood is of utmost importance. Variational methods have been used to approximate models which give rise to analytically

intractable integrals/summations (Saul, Jaakkola & Jordan, 1996; Jaakkola, 1997; Ghahramani & Jordan, 1997; Ghahramani & Hinton, 2000), and more recently have been used to approximate complicated Bayesian learning models (Hinton & van Camp, 1993; Waterhouse, MacKay & Robinson, 1996; MacKay, 1997; Bishop, 1999; Ghahramani & Beal, 2000).

Unfortunately, currently variational approximations are limited in scope. In all but a few cases the models considered come from the "conjugate-exponential" family (Attias, 2000; Ghahramani & Beal, 2001; Winn & Bishop 2005). Conjugate exponential family distributions include Gaussian and discrete multinomial distributions and conjugacy requires the posterior (up to the normalising constant) to have the same functional form as the prior. A variational approximation package `VIBES` fits models including directed acyclic graphs of multinomial discrete variables (with Dirichlet priors) together with arbitrary subgraphs of linear functions of Gaussian nodes (with gamma/Wishart priors), with mixture nodes providing connections from discrete to the continuous subgraphs (Winn & Bishop, 2005). Special cases include hidden Markov models, Kalman filters, factor analysers, principal component analysers and independent component analysers and robust models stemming from scale mixture Gaussian distributions (Faul & Tipping 2001; Kuss, 2006) and LMMs and Bayesian LMMs (Friston, Glaser, Henson, Kiebel, Phillips & Ashburner, 2002). While this is a fairly general class of models few non "conjugate-exponential" family models have been considered. An important exception is logistic regression, see Jaakkola & Jordan (1997, 2000).

In this chapter we make the following contributions:

1. We derive variational approximations as an alternative method for fitting both GLMMs and Bayesian GLMMs. These jump the "conjugate-exponential" family hurdle for the important case of non-Gaussian response models with Gaussian random effects/prior. These variational approximations find a lower bound for the marginal distribution by approximating the posterior of the random effects by a Gaussian distribution.
2. Derive a new approximation for logistic linear mixed models and compare it with the approximation developed by Jaakkola & Jordan (1997).
3. Develop several algorithms to fitting variational approximations for GLMMs and Bayesian GLMMs.
4. Show that for LMMs the variational approximations considered in this chapter are exact.
5. Show that the variational approximations to Poisson, Gamma and inverse-Gaussian LMMs are better Gaussian approximations in terms of Kullback-Leibler divergence than Laplace's method. They are also are more flexible and have similar form as the Laplace's method.
6. Examine the effectiveness of these approximations via several numerical studies.

5.2 Variational Approximations for Generalised Linear Mixed Models

Suppose we have been given the data $(y_i, \mathbf{x}_i)$, $1 \leq i \leq n$ and wish to predict the ys based on the covariates xs where each $\mathbf{x}_i$ is a row vector of dimension d with $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$. The response vector $\mathbf{y}$ is modelled using the exponential family of distributions given by

$$\begin{aligned} \log[\mathbf{y}|\mathbf{u}] &= \frac{\mathbf{y}^T\theta(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\theta(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}))}{a(\phi)} + \mathbf{1}^T c(\mathbf{y},\phi), \\ \mathbf{u} &\sim N(\mathbf{0}, \mathbf{G}_{\boldsymbol{\sigma}^2}) \end{aligned} \tag{5.1}$$

where $\mathbf{D}_{\boldsymbol{\sigma}^2} = \mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}$ and for simplicity we will assume $\mathbf{D}_{\boldsymbol{\sigma}^2} = \sum_{i=1}^{v} \sigma_i^{-2}\mathbf{D}_i$ where $\mathbf{D}_i = \text{blockdiag}_{1\leq j\leq v}\left(\boldsymbol{\Omega}_j \mathbb{I}_{\{j=i\}}\right)$ for some $q_i \times q_i$ matrices $\boldsymbol{\Omega}_j$, $1 \leq j \leq v$. Table 1.2.1 contains values for $\theta(\eta_i)$, $a(\phi)$, $b(\theta(\eta_i))$ and $c(y_i, \phi)$ for the models we will consider in this chapter. We refer the reader to Section 1.2.1 for fuller details on the specification of GLMMs.

The log-likelihood for this model is

$$\begin{aligned} \ell(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2) &= \log \int [\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \phi][\mathbf{u}; \boldsymbol{\sigma}^2] d\mathbf{u} \\ &= \log \int \exp\left\{ \frac{\mathbf{y}^T\theta(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\theta(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}))}{a(\phi)} - \tfrac{1}{2}\mathbf{u}^T\mathbf{D}_{\boldsymbol{\sigma}^2}\mathbf{u} \right\} d\mathbf{u} \\ &\quad + \mathbf{1}^T c(\mathbf{y}, \phi) + \tfrac{1}{2}\log|\mathbf{D}_{\boldsymbol{\sigma}^2}| - \tfrac{q}{2}\log(2\pi). \end{aligned} \tag{5.2}$$

In general there is no closed form expression for (5.2) except in the case where $\mathbf{y}|\mathbf{u}$ is Gaussian in which case equation (5.1) describes a LMM.

Using the terminology of Section 4.2 we will use *tangent transforms* and *density transforms* to obtain variational lower bounds for ℓ. Tangent transforms use the fact that for any convex differentiable function $f(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^d$ we have

$$f(\mathbf{x}) \geq f(\boldsymbol{\xi}) + (\mathsf{D}_{\mathbf{x}} f(\boldsymbol{\xi}))^T (\mathbf{x} - \boldsymbol{\xi}) \text{ for all } \mathbf{x}, \boldsymbol{\xi} \in \mathbb{R}^n. \tag{5.3}$$

whereas density transforms use the fact that for any density $\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi})$ we have

$$\ell(\boldsymbol{\theta}) = \log \int [\mathbf{y}, \boldsymbol{\vartheta}; \boldsymbol{\theta}] d\boldsymbol{\vartheta} \geq \ell_L(\boldsymbol{\theta}; \boldsymbol{\xi}) = \mathbb{E}_\delta \log[\mathbf{y}, \boldsymbol{\vartheta}; \boldsymbol{\theta}] + \mathcal{H}_\delta \tag{5.4}$$

where $\boldsymbol{\vartheta}$ are variables we want to integrate out, $\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi})$ is a density which approximates the posterior distribution $\boldsymbol{\vartheta}|\mathbf{y}$ with additional parameters $\boldsymbol{\xi}$, $\mathbb{E}_\delta$ denotes expectation with respect to $\delta(\boldsymbol{\vartheta})$, $\mathcal{H}_\delta = -\mathbb{E}_\delta \log(\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi}))$ is the entropy of δ and the subscript L denotes a lower bound. Also, suppose $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_p)$ is some partition of the $\boldsymbol{\vartheta}$ vector, $\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi}) = \prod_{i=1}^{p} \delta_i(\boldsymbol{\vartheta}_i)$ and we are approximating each $\boldsymbol{\vartheta}_i|\mathbf{y}$ by some known distribution F_i, for example a Gaussian or a gamma distribution. Then we denote this by $\boldsymbol{\vartheta}_i|\mathbf{y} \sim_{\delta_i} F_i(\boldsymbol{\vartheta}; \xi)$, $i \in \{1, \ldots, p\}$.

The following result uses density transforms and failing that uses a combination of density and tangent transforms (namely the ξ and log transforms, see Table 4.2.1) to obtain lower bounds to the likelihood ℓ corresponding to the GLMMs described Section 1.2.1.

Result 4.1: *Consider the class of GLMM models defined by (5.1) and Table 1.2.1. Let* $\mathbf{C} = [\mathbf{X}, \mathbf{Z}]$ *and* $\boldsymbol{\nu} = (\boldsymbol{\beta}, \boldsymbol{\mu})$. *Using (5.4) with*

$$\mathbf{u}|\mathbf{y} \sim_\delta N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

a lower bound for the likelihood, denoted ℓ_L, *is given by*

$$\begin{aligned} \ell(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2) &\geq \ell_L(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2; \boldsymbol{\xi}) \\ &= \mathbb{E}_\delta \log[\mathbf{y}, \mathbf{u}] + \mathcal{H}_\delta \\ &= \mathbb{E}_\delta \log[\mathbf{y}|\mathbf{u}] + \mathbb{E}_\delta \log[\mathbf{u}] + \mathcal{H}_\delta \end{aligned} \tag{5.5}$$

which holds for all variational parameters $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *parameters such that* $\boldsymbol{\Sigma}$ *is positive definite. The relevant expectations in (5.5) are*

$$\begin{aligned} \mathbb{E}_\delta \log[\mathbf{y}|\mathbf{u}] &= \frac{\mathbf{y}^T\widehat{\boldsymbol{\theta}} - \mathbf{1}^T\widehat{\mathbf{b}}}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}, \phi) \\ \text{and } \mathbb{E}_\delta \log[\mathbf{u}] &= \tfrac{1}{2}\log|(2\pi)^{-1}\mathbf{D}_{\boldsymbol{\sigma}^2}| - \frac{\boldsymbol{\mu}^T\mathbf{D}_{\boldsymbol{\sigma}^2}\boldsymbol{\mu} + \operatorname{tr}(\boldsymbol{\Sigma}\mathbf{D}_{\boldsymbol{\sigma}^2})}{2} \end{aligned}$$

with $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$, $\widehat{\mathbf{b}} = (\widehat{b}_1, \ldots, \widehat{b}_n)$ *and both* $\widehat{\theta}_i = \mathbb{E}_\delta(\theta(\eta_i))$ *and* $\widehat{b}_i = \mathbb{E}_\delta(b(\theta(\eta_i)))$ *are listed in Table 5.2.1.*

Model	$\widehat{\theta}_i = \mathbb{E}_\delta(\theta(\eta_i))$	$\widehat{b}_i = \mathbb{E}_\delta(b(\theta(\eta_i)))$
Normal	$(\mathbf{C}\boldsymbol{\nu})_i$	$\frac{1}{2}\left((\mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}\right)$
Logistic (with ξ transform)	$(\mathbf{C}\boldsymbol{\nu})_i$	$\leq \frac{(\mathbf{C}\boldsymbol{\nu})_i}{2} + \log(e^{\xi_i/2} + e^{-\xi_i/2}) + \frac{\tanh(\xi_i/2)}{4\xi_i}\left((\mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii} - \xi_i^2\right)$
Logistic (with log transform)	$(\mathbf{C}\boldsymbol{\nu})_i$	$\leq \log\left(1 + e^{(\mathbf{C}\boldsymbol{\nu})_i + \frac{1}{2}(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}\right)$
Poisson	$(\mathbf{C}\boldsymbol{\nu})_i$	$e^{(\mathbf{C}\boldsymbol{\nu})_i + \frac{1}{2}(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}$
Gamma	$-e^{-(\mathbf{C}\boldsymbol{\nu})_i + \frac{1}{2}(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}$	$(\mathbf{C}\boldsymbol{\nu})_i$
Inverse-Gaussian	$-e^{-2(\mathbf{C}\boldsymbol{\nu})_i + 2(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}$	$-e^{-(\mathbf{C}\boldsymbol{\nu})_i + \frac{1}{2}(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}$

Table 5.2.1: *A summary of relevant expectations for the variational approximation (5.5) for the generalised linear mixed models defined by (5.1) and Table 1.2.1. Note that for the logistic model cases exact expressions for* $\mathbb{E}_\delta(b(\theta(\eta_i)))$ *have not been obtained. Instead upper bounds for* $\mathbb{E}_\delta(b(\theta(\eta_i)))$ *obtained via the* ξ *and* log *transforms are listed above.*

Proof: The result follows from applying (5.4) to (5.2) with $\boldsymbol{\vartheta} = \mathbf{u}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2)$. The calculation of $\mathbb{E}_\delta \log[\mathbf{u}]$ follows from the fact that for any random vector $\mathbf{v}$ and constant, appropriately-sized matrix $\mathbf{A}$

$$\mathbb{E}(\mathbf{v}^T\mathbf{A}\mathbf{v}) = \mathbb{E}(\mathbf{v})^T\mathbf{A}\mathbb{E}(\mathbf{v}) + \operatorname{tr}(\mathbf{A}\operatorname{Cov}(\mathbf{v}))$$

with $\mathbb{E}_\delta(\mathbf{u}) = \boldsymbol{\mu}$ and $\operatorname{Cov}_\delta(\mathbf{u}) = \boldsymbol{\Sigma}$.

For the Poisson, gamma and inverse-Gaussian cases $\mathbb{E}_\delta(\theta(\eta_i))$ and $\mathbb{E}_\delta(b(\theta(\eta_i)))$ can be calculated using the fact

$$\mathbb{E}_\delta(e^{\mathbf{t}^T\mathbf{u}}) = \mathrm{Mgf}_\delta(\mathbf{t}) = e^{\mathbf{t}^T\boldsymbol{\mu}+\frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}}$$

where $\mathrm{Mgf}_\delta(\mathbf{t})$ is the moment generating function of δ.

For the logistic case there is no closed form expression for $\mathbb{E}_\delta(b(\theta(\eta_i)))$. Instead, in keeping with variational methodology, we look for a upper bounds for $\mathbb{E}_\delta(b(\theta(\eta_i)))$ (and hence a lower bounds for $-\mathbb{E}_\delta(b(\theta(\eta_i)))$. Examining Table 4.2.1 there are two alternative tangent transforms which we may apply to this end, namely the ξ-transform (as termed by Jaakkola & Jordan, 1997) and the log-transform.

Applying the ξ-transform to $b(\cdot)$ we obtain

$$b(x) \leq \log(e^{\frac{\xi}{2}} + e^{-\frac{\xi}{2}}) + x/2 + \frac{\tanh(\xi/2)}{4\xi}\left(x^2 - \xi^2\right) \tag{5.6}$$

which holds for all (x, ξ). Hence using the ξ-transform

$$\mathbb{E}_\delta(b(\theta(\eta_i))) \leq \frac{(\mathbf{C}\boldsymbol{\nu})_i}{2} + \log(e^{\xi_i/2} + e^{-\xi_i/2}) + \frac{\tanh(\xi_i/2)}{4\xi_i}\left((\mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii} - \xi_i^2\right) \tag{5.7}$$

which holds for all ξ_i.

Alternatively, applying the log-transform to $b(\cdot)$ we obtain

$$b(x) \leq \xi(1 + e^x) - \log(\xi) - 1 \tag{5.8}$$

which holds for all (x, ξ). Hence

$$\begin{aligned}\mathbb{E}_\delta(b(\theta(\eta_i))) &\leq \underset{\xi_i}{\mathrm{argmax}}\left\{\mathbb{E}_\delta(\xi_i(1 + e^{(\mathbf{C}\boldsymbol{\nu})_i + \frac{1}{2}(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}) - \log(\xi_i) - 1)\right\} \\ &= \log\left(1 + e^{(\mathbf{C}\boldsymbol{\nu})_i + \frac{1}{2}(\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}\right).\end{aligned} \tag{5.9}$$

■

We note that the bound for $\ell(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2) \geq \ell_L(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ in (5.5) is *new*, to the best of our knowledge, for the Poisson, gamma and inverse-Gaussian LMM cases. For the logistic LMM case we used a combination of density and tangent transforms to obtain a lower bounds for ℓ. Using (5.9) we obtain *new*, to the best of our knowledge, bounds for ℓ in the context of GLMMs, although the log-transform has been used within the context of graphical models by Saul, Jaakkola & Jordan (1995). The bound (5.7) was first developed in Jaakkola & Jordan (1997) in the context of Bayesian logistic models and was later used for logistic LMMs by Rijmen & Vomlel (2007). Finally, as we will later show, for the Gaussian case $\ell(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\mathrm{argmax}}\, \ell_L(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

5.2.1 *Comparing ξ and* log *transforms for Logistic LMMs*

As a means of comparing the log and ξ tangent transforms we first note that for the ξ-transform we can use Table 4.2.1 to deduce the optimal value $\widehat{\xi}_i$ for ξ_i to be

$$\widehat{\xi}_i = \sqrt{(\mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}. \tag{5.10}$$

Substituting this back into (5.7) we obtain

$$\mathbb{E}_\delta(b(\theta(\eta_i))) \leq \frac{(\mathbf{C}\boldsymbol{\nu})_i}{2} + g\left(\sqrt{(\mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}\right)$$

where $g(x) = \log(e^{\xi/2} + e^{-\xi/2})$.

Hence the variational upper bounds used for $\mathbb{E}_\delta(b(\theta(\eta_i)))$ in the logistic case can be written as

$$b_{U_1}(x) = \frac{x}{2} + \log\left(e^{\sqrt{x^2+y}/2} + e^{-\sqrt{x^2+y}/2}\right) \quad \text{and} \quad b_{U_2}(x) = \log\left(1 + e^{x+y/2}\right) \tag{5.11}$$

where $x = (\mathbf{C}\boldsymbol{\nu})_i$ and $y = (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}$ and $b_{U_1}(x)$ corresponds to the ξ-transform and $b_{U_2}(x)$ corresponds to the log-transform. Plotting over a grid of x and y values we can compare the relative sizes of $b_{U_1}(x)$ and $b_{U_2}(x)$ with smaller values of $b_{U_1}(x)$ and $b_{U_2}(x)$ indicate tighter bounds and hence a better approximation of $b(x)$. Typically $-8 \leq x \leq 8$ and $-8 \leq \log(y) \leq -2$.

Figure 5.1 illustrates the bound between the ξ and log transform approximations of $b(x)$. Figure 5.1 also illustrates the regions of $(x, \log(y))$ space where each of these approximations are better. It roughly appears that if $x = (\mathbf{C}\boldsymbol{\nu})_i$ is greater than about -2 for most values of $y = (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}$ then the ξ-transform is better. Further numerical comparisons will be made in Section 5.4.

5.2.2 *Optimisation*

There are a number of factors which make efficient maximisation of $\ell_L(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ difficult. Consider the first derivatives of ℓ_L

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{\beta}}\ell_L &= \mathbf{X}^T\boldsymbol{\varepsilon}, \\
\mathsf{D}_{\boldsymbol{\mu}}\ell_L &= \mathbf{Z}^T\boldsymbol{\varepsilon} - \mathbf{D}_{\boldsymbol{\sigma}^2}\boldsymbol{\mu}, \\
\mathsf{D}_{\sigma_i^2}\ell_L &= \frac{\boldsymbol{\mu}^T\mathbf{D}_i\boldsymbol{\mu} + \operatorname{tr}(\boldsymbol{\Sigma}\mathbf{D}_i)}{2(\sigma_i^2)^2} - \frac{q_i}{2\sigma_i^2},\ 1 \leq i, \leq v, \\
\mathsf{D}_{\phi}\ell_L &= -(\mathsf{D}_{\phi}a(\phi))\frac{\mathbf{y}^T\widehat{\boldsymbol{\theta}} - \mathbf{1}^T\widehat{\mathbf{b}}}{a(\phi)^2} + (\mathsf{D}_{\phi}\mathbf{1}^T c(\mathbf{y}, \phi)) \\
\text{and } \mathsf{D}_{\Sigma_{ij}}\ell_L &= \operatorname{tr}\left((\boldsymbol{\Sigma}^{-1} - \mathbf{Z}^T\mathbf{W}\mathbf{Z} - \mathbf{D}_{\boldsymbol{\sigma}^2})\mathbf{E}_{ij}\right)/2,\ 1 \leq i, j \leq q
\end{aligned} \tag{5.12}$$

where $\mathsf{D}_{\phi}a(\phi)$ and $\mathsf{D}_{\phi}c(\mathbf{y}, \phi)$ are summarised in Table 1.1, $\boldsymbol{\varepsilon}$ and $\mathbf{W}$ can be obtained from Table 5.2.2 and $\mathbf{E}_{ij}$ are matrices of zeros, except for the (i, j)th entry which is 1, with the same dimensions as $\boldsymbol{\Sigma}$. Assuming v is much smaller than n, p and q, which describes all but some unusual cases, the cost of calculating the first derivatives is $O(n(p+q)^2 + (p+q)^3)$.

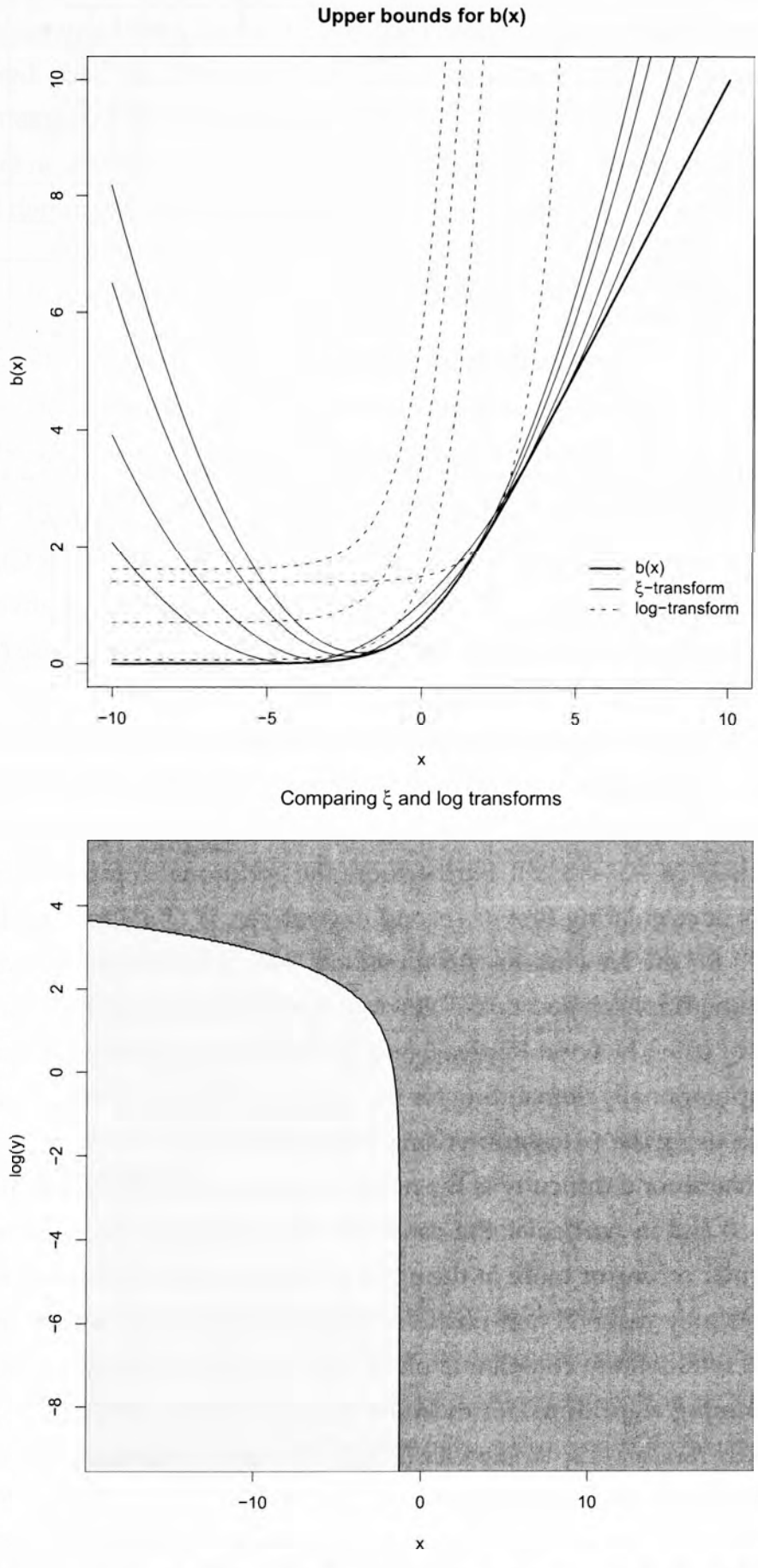


Figure 5.1: *A comparison of the ξ and* log *transforms. The top panel illustrates the upper bounds for $b(x)$ using the ξ-transform of Jaakkola & Jordan (1997) and the* log*-transform. The lower panel compares where each approximation of $\mathbb{E}_\delta(b(\theta(\eta_i)))$. The darker grey region indicates $\frac{x}{2} + \log\left(e^{\sqrt{x^2+y}/2} + e^{-\sqrt{x^2+y}/2}\right) > \log\left(1 + e^{x+y/2}\right)$ where $x = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\mu})_i$ and $y = (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}$. In this region the indicates ξ-transform is a closer than the* log*-transform to $\mathbb{E}_\delta(b(\theta(\eta_i)))$.*

Model	ε_i	$\mathbf{W}_{ii}$	$\mathbf{S}_{ii}$
Normal	$(y_i - \widehat{\theta}_i)/a(\phi)$	$a(\phi)^{-1}$	$a(\phi)^{-1}$
Bernoulli (ξ-trans)	$y_i - 1/2 - \frac{\tanh(\xi_i/2)}{2\xi_i}\widehat{\theta}_i$	$\frac{\tanh(\xi_i/2)}{2\xi_i}$	$\frac{\tanh(\xi_i/2)}{2\xi_i}$
Bernoulli (log-trans)	$y_i - (e^{-\widehat{b}_i} - 1)$	$e^{-\widehat{b}_i} - 1$	$e^{-\widehat{b}_i}(e^{-\widehat{b}_i} - 1)$
Poisson	$y_i - \widehat{b}_i$	$\widehat{b}_i$	$\widehat{b}_i$
Gamma	$-(y_i\widehat{\theta}_i + 1)/a(\phi)$	$-y_i\widehat{\theta}_i/a(\phi)$	$-y_i\widehat{\theta}_i/a(\phi)$
Inverse-Gaussian	$(\widehat{b}_i - 2y_i\widehat{\theta}_i)/a(\phi)$	$(\widehat{b}_i - 4y_i\widehat{\theta}_i)/a(\phi)$	$(\widehat{b}_i - 4y_i\widehat{\theta}_i)/a(\phi)$

Table 5.2.2: *A summary of derivative parameters in (5.12) for (5.5).*

If we combine (5.12), with the values in Tables 1.2.1 and 5.2.1–5.2.2 we can calculate ℓ_L and all the derivatives of ℓ_L with respect to $(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Using this information we maximise ℓ_L using a quasi-Newton method (see Appendix C), for example using the `optim()` function in R. We could also find the second derivatives of ℓ_L with respect to $(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and define Newton-Raphson updates to optimise ℓ_L. There are difficulties with both of these approaches.

The first of these difficulties with efficient maximisation of ℓ_L is that $\boldsymbol{\Sigma}$ represents $q(q+1)/2$ parameters which contributed a large proportion of the total $(p + q + v + q(q+1)/2)$ parameters. The storage costs for Newton-Raphson and quasi-Newton methods is $O((p + v + q^2)^2)$. Furthermore the additional computational costs, excluding the costs of calculating first or second derivatives, is $O(q^4)$ for quasi-Newton methods and $O(q^6)$ for the Newton-Raphson method. Hence both storage and computation costs for Newton-Raphson and quasi-Newton methods are prohibitive for even moderate q. The cost of using Newton-Raphson and quasi-Newton methods directly becomes even more computationally demanding for the logistic LMM case if one considers the ξ_i parameters when using the ξ-transform using these methods.

The second difficulty is the need to take into account the implicit constraints $\boldsymbol{\sigma}^2 > \mathbf{0}$, $\phi > 0$ and in particular the constraint that $\boldsymbol{\Sigma}$ should be positive definite. If $\boldsymbol{\Sigma}$ is near singular or one or more of the σ_i^2 or ϕ are near zero then Newton or quasi-Newton iterations may make $\boldsymbol{\Sigma}$ non-positive definite or one of the σ_i^2 or ϕ negative. Optimisation with semidefinite constraints on $\boldsymbol{\Sigma}$ may be performed by modifying semidefinite programming algorithms, for example Vandenberghe & Boyd (1996) or Kruk, Muramatsu, Rendl, Vanderbei & Wolkowicz (2001). However, we anticipate that this approach would also be computationally expensive.

To avoid these complications we propose to use Newton updates for the $\boldsymbol{\nu}$ variables and fixed point solutions for $\boldsymbol{\Sigma}$ and σ_i^2. These fixed point equations are given by

$$\begin{aligned} \boldsymbol{\nu} &:= \boldsymbol{\nu} + (\mathbf{C}^T\mathbf{S}\mathbf{C} + \mathbf{B}_{\boldsymbol{\sigma}^2})^{-1}(\mathbf{C}\boldsymbol{\varepsilon} - \mathbf{B}_{\boldsymbol{\sigma}^2}\boldsymbol{\nu}) \\ \boldsymbol{\Sigma} &:= (\mathbf{Z}^T\mathbf{W}\mathbf{Z} + \mathbf{D}_{\boldsymbol{\sigma}^2})^{-1} \\ \sigma_i^2 &:= \frac{\boldsymbol{\mu}^T\mathbf{D}_i\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{D}_i)}{q_i},\ 1 \le i \le v \end{aligned} \tag{5.13}$$

where $\mathbf{B}_{\boldsymbol{\sigma}^2} \equiv \text{blockdiag}(\mathbf{0}_p, \mathbf{D}_{\boldsymbol{\sigma}^2})$ and the expressions for $\mathbf{W}$, $\boldsymbol{\varepsilon}$ and $\mathbf{S}$ are case dependent and may be obtained from Tables 5.2.1–5.2.2. For the Gaussian and inverse-Gaussian we can find fixed point equations for the nuisance parameter ϕ as is given in Table 5.2.3. These updates guarantee that ϕ remain positive for these cases.

There are no straightforward fixed point equations for ϕ for gamma LMMs. The first and second derivatives of ℓ_L with respect to ϕ in this case are

$$\begin{aligned} \mathsf{D}_\phi \ell_L &= \mathbf{y}^T \widehat{\boldsymbol{\theta}} - \mathbf{1}^T \widehat{\mathbf{b}} + \log(\phi)\mathbf{1}^T \log(\mathbf{y}) + n - n\psi(\phi) \\ \text{and } \mathsf{H}_\phi \ell_L &= n/\phi - n\psi'(\phi) \end{aligned}$$

where $\psi(\cdot)$ is the digamma function and $\psi'(\cdot)$ is the trigamma function (see Abramowitz & Stegun, 1964, Chapter 6). Unfortunately, Newton-Raphson steps may make $\phi < 0$. Instead we propose to first make the transformation $\phi = e^r$, and then use Newton-Raphson updates on r. The first derivatives of ℓ_L with respect to r are

$$\begin{aligned} \mathsf{D}_r \ell_L &= (\mathsf{D}_r \phi)(\mathsf{D}_\phi \ell_L) &&= \phi(\mathsf{D}_\phi \ell_L), \\ \text{and } \mathsf{H}_r \ell_L &= (\mathsf{H}_r \phi)(\mathsf{D}_\phi \ell_L) + (\mathsf{D}_r \phi)^2 (\mathsf{H}_\phi \ell_L) &&= \phi(\mathsf{D}_\phi \ell_L) + \phi^2(\mathsf{H}_\phi \ell_L). \end{aligned}$$

Using these, the fixed point updates for the nuisance parameters in the Gaussian, gamma and inverse-Gaussian cases are given in Table 5.2.3.

Model	Update
Gaussian	$\phi := \dfrac{2(\mathbf{1}^T \widehat{\mathbf{b}} - \mathbf{y}^T \widehat{\boldsymbol{\theta}}) - \mathbf{1}^T(\mathbf{y}^2)}{n}$
Gamma	$\phi := \phi \exp\Big\{ -(\mathbf{y}^T \widehat{\boldsymbol{\theta}} - \mathbf{1}^T \widehat{\mathbf{b}} + \log(\phi)\mathbf{1}^T \log(\mathbf{y}) + n - n\psi(\phi)) \Big/ (\mathbf{y}^T \widehat{\boldsymbol{\theta}} - \mathbf{1}^T \widehat{\mathbf{b}} + \log(\phi)\mathbf{1}^T \log(\mathbf{y}) + 2n - n\psi(\phi) - n\phi\psi'(\phi)) \Big\}$
Inverse-Gaussian	$\phi := \dfrac{2(\mathbf{1}^T \widehat{\mathbf{b}} - \mathbf{y}^T \widehat{\boldsymbol{\theta}}) - \sum_{i=1}^n y_i^{-1}}{n}$

Table 5.2.3: *A summary fixed point updates for the nuisance parameters in Gaussian, gamma and inverse-Gaussian models.*

The advantage of these updates is that they guarantee that $\boldsymbol{\Sigma}$, $\boldsymbol{\sigma}^2$ and ϕ are positive-definite or positive respectively. Each set of updates has computational cost $O(n(p+q)^2 + (p+q)^3)$, which is smaller than those based on quasi-Newton or Newton-Raphson updates for all parameters. However, the rate of convergence of these updates is unclear. In practice we have found that these updates can sometimes converge quickly, although they usually converge very slowly.

A second problem is revealed when examining the second derivatives of ℓ with respect to σ_i^2 given by

$$\mathsf{H}_{\sigma_i^2} \ell_L = -\frac{\boldsymbol{\mu}^T \mathbf{D}_i \boldsymbol{\mu} + \text{tr}\,(\boldsymbol{\Sigma}\mathbf{D}_i)}{(\sigma_i^2)^3} + \frac{q_i}{2(\sigma_i^2)^2}, \quad 1 \leq i \leq v.$$

Since this is not negative for all $\boldsymbol{\mu}$, σ_i^2 and $\boldsymbol{\Sigma}$ the function ℓ is not concave. In particular $\mathrm{D}_{\sigma_i^2}\ell_L$ is negative if and only if

$$\sigma_i^2 < \frac{2\left[\boldsymbol{\mu}^T\mathbf{D}_i\boldsymbol{\mu} + \mathrm{tr}\left(\boldsymbol{\Sigma}\mathbf{D}_i\right)\right]}{q_i},\ 1 \leq i \leq v.$$

Furthermore $\mathrm{D}_{\sigma_i^2}\ell_L$, is negative for any σ_i^2 satisfying (5.13).

In practice this means that the fixed point updates (5.13) may converge to different points depending on the initial values chosen. We have found that if $\boldsymbol{\mu}$ does not start out "large enough" then the fixed points converge close to $\boldsymbol{\mu} = \mathbf{0}$. To side step this issue we first fix each σ_i^2 at some "large" positive value, and $\boldsymbol{\Sigma}$ as a diagonal matrix with small positive entries. We update $\boldsymbol{\nu}$ until convergence and then update all parameters until convergence. We have found that applying the fixed point updates for $\boldsymbol{\nu}$ more often than for $\boldsymbol{\Sigma}$ and σ_i^2 improves stability. This approach is formalised in Algorithm 6. We use $L = 2$ for most of the numerical experiments in Section 5.4.

Algorithm 6 Fixed Point Maximisation of Variational Approximation to GLMM

1. Initialise $(\boldsymbol{\nu}, \phi, \boldsymbol{\sigma}^2, \boldsymbol{\Sigma})$.

2. Cycle

Apply

$$\boldsymbol{\nu} := \boldsymbol{\nu} + (\mathbf{C}^T\mathbf{S}\mathbf{C} + \mathbf{B}_{\boldsymbol{\sigma}^2})^{-1}(\mathbf{C}\boldsymbol{\varepsilon} - \mathbf{B}_{\boldsymbol{\sigma}^2}\boldsymbol{\nu})$$

and update ϕ using Table 5.2.3.

Until convergence.

3. Cycle

Apply updates for $\boldsymbol{\sigma}^2$ and $\boldsymbol{\Sigma}$ using

$$\begin{aligned}\boldsymbol{\Sigma} &:= (\mathbf{Z}^T\mathbf{W}\mathbf{Z} + \mathbf{D}_{\boldsymbol{\sigma}^2})^{-1} \\ \sigma_i^2 &:= \frac{\boldsymbol{\mu}^T\mathbf{D}_i\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{D}_i)}{q_i},\ 1 \leq i \leq v.\end{aligned}$$

for $iter = 1, \ldots, L$ **do**

Apply

$$\boldsymbol{\nu} := \boldsymbol{\nu} + (\mathbf{C}^T\mathbf{S}\mathbf{C} + \mathbf{B}_{\boldsymbol{\sigma}^2})^{-1}(\mathbf{C}\boldsymbol{\varepsilon} - \mathbf{B}_{\boldsymbol{\sigma}^2}\boldsymbol{\nu})$$

and update ϕ using Table 5.2.3.

end for.

Until convergence.

As an alternative to both these methods, we propose a hybrid between a quasi-Newton and the fixed point approaches. Using quasi-Newton steps to update $\boldsymbol{\Sigma}$ slows down the quasi-Newton because of the $q(q+1)/2$ of parameters to update. To remove this bottleneck we only use quasi-Newton steps to update $(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2, \boldsymbol{\mu})$ and use fixed point iterations to update $\boldsymbol{\Sigma}$. This is described in Algorithm 7. The advantage of this algorithm is that the derivatives for each quasi-Newton update costs at most $O(n(p+q)^2)$ are lower

and each update of $\boldsymbol{\Sigma}$ costs $O(n(p+q)^2+(p+q)^3)$. Again, the rate of convergence of these updates is unclear, but in practice the algorithm usually fits faster than quasi-Newton optimisation or the fixed point updates alone.

We compare each of these methods in Section 5.4.

Algorithm 7 Quasi-Newton/FP Hybrid for Gaussian transform of the GLMM

1. Initialise $(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \phi, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

2. Cycle

2.1. Quasi-Newton Steps

Using equations (5.12) with the values in Tables 1.1 and 5.2.1–5.2.2 to calculate the derivatives of ℓ_L with respect to $(\boldsymbol{\nu}, \boldsymbol{\sigma}^2, \phi)$. Use 5-10 quasi-Newton updates.

2.2. Fixed Point Step for $\boldsymbol{\Sigma}$

Update $\boldsymbol{\Sigma}$ using

$$\boldsymbol{\Sigma} := (\mathbf{Z}^T\mathbf{W}\mathbf{Z} + \mathbf{D}_{\boldsymbol{\sigma}^2})^{-1} \tag{5.14}$$

where $\mathbf{W}$ can be obtained from Table 5.2.1–5.2.2.

Until convergence.

5.2.3 *Comparisons with the Laplace approximation*

Based on the fixed point equations (5.13) for we $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ we may write

$$\begin{aligned}&\underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmax}}\, \ell_L(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\quad = -\tfrac{1}{2}\log|\mathbf{I} + \mathbf{Z}^T\mathbf{W}\mathbf{Z}\mathbf{D}_{\boldsymbol{\sigma}^2}^{-1}| + \frac{\mathbf{y}^T\widehat{\boldsymbol{\theta}} - \mathbf{1}^T\widehat{\mathbf{b}}}{a(\phi)} + \mathbf{1}^T c(\mathbf{y}, \phi) - \tfrac{1}{2}\boldsymbol{\mu}^T\mathbf{D}_{\sigma^2}\boldsymbol{\mu}\end{aligned} \tag{5.15}$$

where the right hand side of (5.15) is subject to the constraints

$$\begin{aligned}\mathbf{Z}^T\boldsymbol{\varepsilon} - \mathbf{D}_{\boldsymbol{\sigma}^2}\boldsymbol{\mu} &= \mathbf{0}, \\ \boldsymbol{\Sigma} - (\mathbf{Z}^T\mathbf{W}\mathbf{Z} + \mathbf{D}_{\boldsymbol{\sigma}^2})^{-1} &= \mathbf{0} \\ \text{and } \boldsymbol{\sigma}^2, \phi &\geq \mathbf{0}.\end{aligned} \tag{5.16}$$

We see that (5.15) together with (5.16) has a similar functional form as the Laplace approximation given by (3.5) and (3.6) except $\widehat{\boldsymbol{\theta}}$, $\widehat{\mathbf{b}}$, $\boldsymbol{\varepsilon}$ and $\mathbf{W}$ are different and there is an additional matrix equality constraint for $\boldsymbol{\Sigma}$.

The variational approximation provides several advantages over the Laplace approximation

- In the case of LMMs we can simplify (5.15) and (5.16) to get the exact marginal likelihood (see Result 5.2 below)
- In the cases of Poisson, gamma and inverse-Gaussian GLMMs, using the Gaussian density transform, i.e. using (5.4) with $\mathbf{u}|\mathbf{y} \sim_\delta N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, can be calculated exactly. Hence, based on (4.12) for fixed $\boldsymbol{\beta}$, ϕ and $\boldsymbol{\sigma}^2$ the equation (5.15) subject to (5.16) is

the optimal Gaussian approximation in terms of the KL-divergence criteria. Furthermore, because these are optimal in the sense of their KL-divergence, they provide better Gaussian approximations than the Laplace approximation in this sense.

- While not considered here, it is possible to impose structure onto the covariance matrix $\boldsymbol{\Sigma}$ by only considering $\boldsymbol{\Sigma}$ with a particular structure, e.g. diagonal $\boldsymbol{\Sigma}$. This might be used to increase computational efficiency of fitting a suitably modified (5.15) subject to (5.16). Alternatively, this fact could be used to enforce a particular structure on $\boldsymbol{\Sigma}$ needed for a particular application.

Although this result should be obvious, based on the discussion in Section 4.2.3, we show this explicitly for illustration.

Result 5.2: *For the Gaussian case maximising (5.5) over the variational parameters* $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *we obtain the marginal likelihood, i.e.*

$$\begin{aligned}\underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmax}}\, \ell_L(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= \ell(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}^2)\end{aligned}$$

where $\phi = \sigma_\varepsilon^2$ *and* $\mathbf{V} = \sigma_\varepsilon^2\mathbf{I} + \mathbf{Z}\mathbf{D}_{\boldsymbol{\sigma}^2}^{-1}\mathbf{Z}^T$.

Proof: Using (5.15) leads to the constraints (5.16). Using (5.16) we may write, for the Gaussian case, $\boldsymbol{\Sigma} = (\sigma_\varepsilon^{-2}\mathbf{Z}^T\mathbf{Z} + \mathbf{D}_{\boldsymbol{\sigma}^2})^{-1}$ and, via simple algebraic manipulations, $\boldsymbol{\mu} = \sigma_\varepsilon^{-2}\boldsymbol{\Sigma}\mathbf{Z}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$. Substituting these back into (5.15) we obtain

$$\begin{aligned}&\underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmax}}\, \ell_L(\boldsymbol{\beta}, \sigma_\varepsilon^{-2}, \boldsymbol{\sigma}^2; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\quad= -\tfrac{n}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\mu}\|^2}{2\sigma_\varepsilon^2} - \frac{\operatorname{tr}(\mathbf{Z}^T\mathbf{Z}\boldsymbol{\Sigma})}{2\sigma_\varepsilon^2} \\ &\qquad+\tfrac{1}{2}\log|\mathbf{D}_{\boldsymbol{\sigma}^2}| - \tfrac{1}{2}\boldsymbol{\mu}^T\mathbf{D}_{\boldsymbol{\sigma}^2}\boldsymbol{\mu} + \tfrac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}\mathbf{D}_{\boldsymbol{\sigma}^2}) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}| \\ &\quad= -\tfrac{n}{2}\log(2\pi\sigma_\varepsilon^2) + \tfrac{1}{2}\log|(\sigma_\varepsilon^{-2}\mathbf{Z}^T\mathbf{Z} + \mathbf{D}_{\boldsymbol{\sigma}^2})^{-1}\mathbf{D}_{\boldsymbol{\sigma}^2}| \\ &\qquad-\tfrac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\left(\sigma_\varepsilon^{-2}\mathbf{I} - \sigma_\varepsilon^{-4}\mathbf{Z}(\sigma_\varepsilon^{-2}\mathbf{Z}^T\mathbf{Z} + \mathbf{D}_{\boldsymbol{\sigma}^2})^{-1}\mathbf{Z}^T\right)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\end{aligned} \tag{5.17}$$

where the result follows from the fact that

$$\begin{aligned}&-\tfrac{1}{2}\log|\sigma_\varepsilon^2\mathbf{I}| + \tfrac{1}{2}\log|(\sigma_\varepsilon^{-2}\mathbf{Z}^T\mathbf{Z} + \mathbf{D}_{\boldsymbol{\sigma}^2})^{-1}\mathbf{D}_{\boldsymbol{\sigma}^2}| \\ &\quad= -\tfrac{1}{2}\log|\sigma_\varepsilon^2\mathbf{I}| - \tfrac{1}{2}\log|\mathbf{I} + \sigma_\varepsilon^{-2}\mathbf{Z}\mathbf{D}_{\boldsymbol{\sigma}^2}^{-1}\mathbf{Z}^T| = -\tfrac{1}{2}\log|\sigma_\varepsilon^2\mathbf{I} + \mathbf{Z}\mathbf{D}_{\boldsymbol{\sigma}^2}^{-1}\mathbf{Z}^T| \\ &\quad= -\tfrac{1}{2}\log|\mathbf{V}|.\end{aligned}$$

Comparing with, for example equation (4.12) from Ruppert *et al.*, (2003) or equation (2.2) from Wand (2003) we see that (5.17) is exact.

■

5.3 Bayesian Generalised Linear Mixed Models

The Bayesian alternative model for GLMMs differs little from the frequentist formulation considered thus far. Here we will consider the Bayesian GLMM where we place the following additional priors on $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$

$$\begin{aligned} \boldsymbol{\beta} &\sim N(\mathbf{0}, \sigma_{\beta}^2 \mathbf{I}) \\ \sigma_i^2 &\sim IG(A_{\sigma^2,i}, B_{\sigma^2,i}),\ 1 \leq i \leq v, \end{aligned} \tag{5.18}$$

with $\sigma_{\boldsymbol{\beta}}^2$ suitably large, for example 10^8, and $A_{\sigma^2,i} = B_{\sigma^2,i}$ are suitably small, for example 10^{-2}, so that the priors are vague. For convenience we write

$$[\boldsymbol{\sigma}^2] = \prod_{i=1}^{v} [\sigma_i^2]$$

and if nuisance parameters are present then we use the prior

$$\phi \sim IG(A_\phi, B_\phi)$$

where $A_\phi = B_\phi$ are again suitably small.

5.3.1 *Marginal Likelihood*

Firstly, we note that even for some of the simplest Bayesian models, some of the various quantities of interest are not known in closed form. For example, consider calculating the marginal distribution $[\mathbf{y}]$ for the Bayesian LMM. The marginal distribution can be used in the context of model selection when calculating the Bayes factor between two models (Gelman *et al.*, 1995; Kass & Raftery, 1995). The marginal distribution for the Bayesian GLMM is given by

$$[\mathbf{y}] = \int [\mathbf{y}|\boldsymbol{\nu}, \phi][\phi][\boldsymbol{\nu}|\boldsymbol{\sigma}^2][\boldsymbol{\sigma}^2] d\boldsymbol{\nu} d\boldsymbol{\sigma}^2 d\phi$$

where we have combined the parameters $\boldsymbol{\beta}$ and $\mathbf{u}$ to be $\boldsymbol{\nu} = (\boldsymbol{\beta}, \mathbf{u})$ so that

$$\boldsymbol{\nu} \sim N\left(\mathbf{0}, \text{blockdiag}\left\{\sigma_{\beta}^2 \mathbf{I}, \mathbf{G}_{\boldsymbol{\sigma}^2}\right\}\right).$$

The most common variational approach to this integral is to select the density transform that "mirrors" the distribution of the priors. Thus we would select

$$\begin{aligned} \boldsymbol{\nu}|\mathbf{y} &\sim_{\delta_\nu} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \sigma_i^2|\mathbf{y} &\sim_{\delta_{\sigma_i^2}} IG(\alpha_{\sigma^2,i}, \beta_{\sigma^2,i}),\ 1 \leq i \leq v, \\ \phi|\mathbf{y} &\sim_{\delta_\phi} IG(\alpha_\phi, \beta_\phi) \end{aligned} \tag{5.19}$$

and $\delta(\boldsymbol{\nu}, \boldsymbol{\sigma}^2, \phi) = \delta_\nu(\boldsymbol{\nu})\delta_\phi(\phi)\prod_{i=1}^v \delta_{\sigma_i^2}(\sigma_i^2)$. Using this density

$$\begin{aligned}
\log[\mathbf{y}] &\geq \log[\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_\phi, \beta_\phi, \boldsymbol{\alpha}_{\sigma^2}, \boldsymbol{\beta}_{\sigma^2}]_L \\
&= \mathbb{E}_\delta \log\left\{[\mathbf{y}|\boldsymbol{\nu}, \phi][\boldsymbol{\nu}|\boldsymbol{\sigma}^2][\boldsymbol{\sigma}^2][\phi]\right\} + \mathcal{H}_\delta \\
&= \mathbb{E}_\delta \log[\mathbf{y}|\boldsymbol{\nu}, \phi] + \mathbb{E}_\delta \log[\boldsymbol{\nu}|\boldsymbol{\sigma}^2] + \mathbb{E}_\delta \log[\phi] + \mathcal{H}_{\delta_\nu} + \mathcal{H}_{\delta_\phi} + \sum_{i=1}^v \mathbb{E}_\delta \log[\sigma_i^2] + \mathcal{H}_{\delta_{\sigma_i^2}}
\end{aligned}$$

where $\boldsymbol{\alpha}_{\sigma^2} = (\boldsymbol{\alpha}_{\sigma^2,1}, \ldots, \boldsymbol{\alpha}_{\sigma^2,v})$, $\boldsymbol{\beta}_{\sigma^2} = (\beta_{\sigma^2,1}, \ldots, \beta_{\sigma^2,v})$ and, ignoring additive constants,

$$\begin{aligned}
\mathbb{E}_\delta \log[\mathbf{y}|\boldsymbol{\nu}] &= \frac{\mathbf{y}^T\widehat{\boldsymbol{\theta}} - \mathbf{1}^T\widehat{\mathbf{b}}}{\widehat{a(\phi)}} + \mathbf{1}^T\widehat{c(\mathbf{y}, \phi)}, \\
\mathbb{E}_\delta \log[\boldsymbol{\nu}|\boldsymbol{\sigma}^2] &= \sum_{i=1}^v \frac{q_i}{2}\left(\psi(\alpha_{\sigma^2,i}) - \log(\beta_{\sigma^2,i})\right) - \frac{\boldsymbol{\nu}^T\mathbf{B}\boldsymbol{\nu} + \operatorname{tr}(\mathbf{B}\boldsymbol{\Sigma})}{2}, \\
\mathbb{E}_\delta \log[\sigma_i^2] &= -(A_{\sigma^2,i} + 1)(\log(\beta_{\sigma^2,i}) - \psi(\alpha_{\sigma^2,i})) - B_{\sigma^2,i}\frac{\alpha_{\sigma^2,i}}{\beta_{\sigma^2,i}},\ 1 \leq i \leq v, \\
\mathbb{E}_\delta \log[\phi] &= -(A_\phi + 1)(\log(\beta_\phi) - \psi(\alpha_\phi)) - B_\phi\frac{\alpha_\phi}{\beta_\phi}, \\
\mathcal{H}_{\delta_\nu} &= \tfrac{1}{2}\log|\boldsymbol{\Sigma}|, \\
\mathcal{H}_{\delta_{\sigma_i^2}} &= \alpha_{\sigma^2,i} + \log(\beta_{\sigma^2,i}) + \log\Gamma(\alpha_{\sigma^2,i}) - (\alpha_{\sigma^2,i} + 1)\psi(\alpha_{\sigma^2,i}),\ 1 \leq i \leq v, \\
\text{and } \mathcal{H}_{\delta_\phi} &= \alpha_\phi + \log(\beta_\phi) + \log\Gamma(\alpha_\phi) - (\alpha_\phi + 1)\psi(\alpha_\phi)
\end{aligned} \tag{5.20}$$

with q_i being the number of rows/columns in $\boldsymbol{\Omega}_i$ and

$$\mathbf{B} = \text{blockdiag}\left\{\sigma_\beta^{-2}\mathbf{I}, \sum_{i=1}^v(\alpha_{\sigma^2,i}/\beta_{\sigma^2,i})\mathbf{D}_i\right\}.$$

Table 5.2.1 contains the values for $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{b}}$ with $\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\nu} = \boldsymbol{\mu}$ and $\widehat{a(\phi)} = 1/\mathbb{E}_\delta(a(\phi)^{-1})$ and $\widehat{c(\mathbf{y}, \phi)} = \mathbb{E}_\delta(c(\mathbf{y}, \phi))$ can be obtained from Table 5.3.4.

The only difficulty with calculating $\widehat{c(\mathbf{y}, \phi)}$ is the term $\mathbb{E}_\delta \log\Gamma(\phi)$ for the gamma LMM case. Obtaining lower bounds is difficult since $-\log\Gamma(x)$ is concave and so we must consider alternatives. Since this term is a one dimensional integral we can evaluate it numerically using Gaussian integration (Abramowitz & Stegun 1964, Chapter 25). Using a change of variables we can write the integral as

$$\mathbb{E}_\delta \log\Gamma(\phi) = \int_0^\infty e^{-t}\frac{t^{\alpha_\phi - 1}}{\Gamma(\alpha_\phi)}\log\Gamma\left(\frac{\beta_\phi}{t}\right)dt. \tag{5.21}$$

Since this integral is of the form $\int_0^\infty e^{-t}g(t)dt$ we can accurately approximate (5.21) using Gauss-Laguerre integration. Using this technique (5.21) is approximated by

$$\mathbb{E}_\delta \log\Gamma(\phi) \simeq \sum_{k=1}^N \frac{c_k t_k^{\alpha_\phi - 1}}{\Gamma(\alpha_\phi)}\log\Gamma\left(\frac{\beta_\phi}{t_k}\right) = \sum_{k=1}^N c_k g(t_k; \alpha_\phi, \beta_\phi)$$

where

$$g(t; \alpha_\phi, \beta_\phi) = \frac{t^{\alpha_\phi - 1}}{\Gamma(\alpha_\phi)}\log\Gamma\left(\frac{\beta_\phi}{t}\right),$$

Model	Gaussian $y_i \sim N(\eta_i, \phi)$	Gamma $y_i \sim \text{Gamma}\,(-e^{\eta_i}, \phi)$	Inverse-Gaussian $y_i \sim IN\,(-e^{\eta_i}, \phi)$
$\widehat{a(\phi)}^{-1}$	$\frac{\alpha_\phi}{\beta_\phi}$	$\frac{\beta_\phi}{\alpha_\phi - 1}$	$\frac{\alpha_\phi}{\beta_\phi}$
$\mathsf{D}_{\alpha_\phi}\widehat{a(\phi)}^{-1}$	$\frac{1}{\beta_\phi}$	$-\frac{\beta_\phi}{(\alpha_\phi - 1)^2}$	$\frac{1}{\beta_\phi}$
$\mathsf{D}_{\beta_\phi}\widehat{a(\phi)}^{-1}$	$-\frac{\alpha_\phi}{\beta_\phi^2}$	$\frac{1}{\alpha_\phi - 1}$	$-\frac{\alpha_\phi}{\beta_\phi^2}$
$\widehat{c(y_i, \phi)}$	$\frac{\psi(\alpha_\phi) - \log(2\pi) - \log(\beta_\phi)}{2} - \frac{y_i^2 \alpha_\phi}{2\beta_\phi}$	$\frac{\beta_\phi}{(\alpha_\phi - 1)}\left(\log(y_i) + \log(\beta_\phi) - \psi(\alpha)\right) + \frac{\beta_\phi}{(\alpha_\phi - 1)^2} - \log(y_i) - \sum_{k=1}^{N} c_k g(t_k; \alpha_\phi, \beta_\phi)$	$\frac{\psi(\alpha_\phi) - \log(2\pi y_i^3) - \log(\beta_\phi)}{2} - \frac{\alpha_\phi}{2y_i\beta_\phi}$
$\mathsf{D}_{\alpha_\phi}\widehat{c(y_i, \phi)}$	$\frac{\psi'(\alpha_\phi)}{2} - \frac{y_i^2}{2\beta_\phi}$	$-\frac{\psi'(\alpha_\phi)\beta_\phi}{\alpha_\phi - 1} + \frac{(\psi(\alpha_\phi) - \log(\beta_\phi) - \log(y_i))\beta_\phi}{(\alpha_\phi - 1)^2} - \frac{2\beta_\phi}{(\alpha_\phi - 1)^3} - \sum_{k=1}^{N}\left(\log(t_k) - \psi(\alpha_\phi)\right) c_k g(t_k; \alpha_\phi, \beta_\phi)$	$\frac{\psi'(\alpha_\phi)}{2} - \frac{1}{2y_i\beta_\phi}$
$\mathsf{D}_{\beta_\phi}\widehat{c(y_i, \phi)}$	$\frac{y_i^2 \alpha_\phi}{2\beta_\phi^2} - \frac{1}{2\beta_\phi}$	$\frac{\log(y_i) + \log(\beta_\phi) - \psi(\alpha) + 1}{(\alpha_\phi - 1)} + \frac{1}{(\alpha_\phi - 1)^2} - \sum_{k=1}^{N} \frac{\frac{1}{t}\psi\left(\frac{t_k}{\beta_\phi}\right)}{\log\Gamma(\alpha_\phi)} c_k g(t_k; \alpha_\phi, \beta_\phi)$	$\frac{\alpha_\phi}{2y_i\beta_\phi^2} - \frac{1}{2\beta_\phi}$

Table 5.3.4: *A summary expectations and derivatives of nuisance parameters. For the gamma LMM case* $g(t; \alpha_\phi, \beta_\phi) = \frac{t^{\alpha_\phi - 1}}{\Gamma(\alpha_\phi)} \log\Gamma\left(\frac{t}{\beta_\phi}\right)$.

the t_ks are zeros of the Nth Laguerre polynomial

$$L_N(t) = e^t \frac{d^N}{dt^N}(e^{-t}t^N),$$

the coefficients c_k are

$$c_k = \frac{1}{L'_N(t_k)} \int_0^\infty \frac{L_N(t)e^{-t}}{t - t_k} dt = \frac{(N!)^2}{t_k[L'_N(t_k)]^2}$$

and has truncation error $\frac{(N!)^2 g^{(2N)}(\xi)}{(2N)!}$ where $\xi = \max_\xi g^{(2N)}(\xi)$. The values of c_k and t_k are available from Abramowitz & Stegun, (1964, Table 25.9) or can be easily calculated using the function `gauss.quad()` from the `R` package `statmod` (Smyth, 2008). The coefficients c_k decay extremely rapidly, for example when $N = 10$ the value for c_{10} is approximately $\times 10^{-12}$.

The first derivatives of $\log[\mathbf{y}]_L$ with respect to $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_\phi, \beta_\phi, \boldsymbol{\alpha}, \boldsymbol{\beta})$ are given by

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{\mu}} \log[\mathbf{y}]_L &= \mathbf{C}^T\boldsymbol{\varepsilon} - \mathbf{B}\boldsymbol{\mu}, \\
\mathsf{D}_{\Sigma_{ij}} \log[\mathbf{y}]_L &= \operatorname{tr}\left((\boldsymbol{\Sigma}^{-1} - \mathbf{C}^T\mathbf{W}\mathbf{C} - \mathbf{B})\mathbf{E}_{ij}\right)/2,\ 1 \le i, j \le q, \\
\mathsf{D}_{\alpha_{\sigma^2,i}} \log[\mathbf{y}]_L &= \left(\frac{q_i}{2} + A_{\sigma^2,i} - \alpha_{\sigma^2,i}\right)\psi'(\alpha_{\sigma^2,i}) + 1 \\
&\quad - \frac{B_{\sigma^2,i} + \boldsymbol{\mu}^T\mathbf{B}_i\boldsymbol{\mu} + \operatorname{tr}(\mathbf{B}_i\boldsymbol{\Sigma})/2}{\beta_{\sigma^2,i}},\ 1 \le i \le v, \\
\mathsf{D}_{\beta_{\sigma^2,i}} \log[\mathbf{y}]_L &= \frac{\alpha_i}{\beta^2_{\sigma^2,i}}\left(B_{\sigma^2,i} + \frac{\boldsymbol{\mu}^T\mathbf{B}_i\boldsymbol{\mu} + \operatorname{tr}(\mathbf{B}_i\boldsymbol{\Sigma})}{2}\right) \\
&\quad - \left(\frac{q_i}{2} + A_{\sigma^2,i}\right)/\beta_{\sigma^2,i},\ 1 \le i \le v, \\
\mathsf{D}_{\alpha_\phi} \log[\mathbf{y}]_L &= (\mathsf{D}_{\alpha_\phi}\widehat{a(\phi)}^{-1})(\mathbf{y}^T\widehat{\boldsymbol{\theta}} - \mathbf{1}^T\widehat{\mathbf{b}}) + (\mathsf{D}_{\alpha_\phi}\mathbf{1}^T\widehat{c(\mathbf{y},\phi)}) \\
&\quad + (A_\phi - \alpha_\phi)\psi'(\alpha_\phi) - \frac{B_\phi}{\beta_\phi} + 1 \\
\text{and } \mathsf{D}_{\beta_\phi} \log[\mathbf{y}]_L &= (\mathsf{D}_{\beta_\phi}\widehat{a(\phi)}^{-1})(\mathbf{y}^T\widehat{\boldsymbol{\theta}} - \mathbf{1}^T\widehat{\mathbf{b}}) + (\mathsf{D}_{\beta_\phi}\mathbf{1}^T\widehat{c(\mathbf{y},\phi)}) - \frac{A_\phi}{\beta_\phi} + \frac{B_\phi\alpha_\phi}{\beta^2_\phi}
\end{aligned} \tag{5.22}$$

where $\mathbf{C} = [\mathbf{X}, \mathbf{Z}]$, the values of $\boldsymbol{\varepsilon}$ and $\mathbf{W}$ can be obtained from Tables 5.2.1–5.2.2,

$$\mathbf{B}_i = \text{blockdiag}\{0 \times \mathbf{I}_p, \mathbf{D}_i\}$$

the derivatives of $\widehat{a(\phi)}^{-1}$ and $\widehat{c(\mathbf{y},\phi)}$ with respect to α_ϕ and β_ϕ are distribution dependent and can be obtained from Table 5.3.4. The calculation of $\log[\mathbf{y}]_L$ and its derivatives are sufficient to fit this model using a quasi-Newton method with the `R` function `optim`. Alternatively we can use

$$\begin{aligned}
\boldsymbol{\mu} &:= \boldsymbol{\mu} + (\mathbf{C}^T\mathbf{S}\mathbf{C} + \mathbf{B})^{-1}(\mathbf{C}\boldsymbol{\varepsilon} - \mathbf{B}\boldsymbol{\mu}), \\
\boldsymbol{\Sigma} &:= (\mathbf{C}^T\mathbf{W}\mathbf{C} + \mathbf{B})^{-1}, \\
\alpha_{\sigma^2,i} &:= A_i + \frac{q_i}{2},\ 1 \le i \le v \\
\text{and } \beta_{\sigma^2,i} &:= B_i + \frac{\boldsymbol{\mu}^T\mathbf{B}_i\boldsymbol{\mu} + \operatorname{tr}(\boldsymbol{\Sigma}\mathbf{B}_i)}{2},\ 1 \le i \le v.
\end{aligned} \tag{5.23}$$

Nuisance parameters need to be handled on a case by case basis. For the LMMs a fixed point updates for $(\alpha_\phi, \beta_\phi)$ are

$$\begin{aligned} \alpha_\phi &:= A_\phi + \tfrac{n}{2} \\ \text{and } \beta_\phi &:= B_\phi + \frac{\|\mathbf{y} - \mathbf{C}\boldsymbol{\mu}\|^2}{2}. \end{aligned} \tag{5.24}$$

For the inverse-Gaussian LMM a fixed point updates for $(\alpha_\phi, \beta_\phi)$ are

$$\begin{aligned} \alpha_\phi &:= A_\phi + \tfrac{n}{2} \\ \text{and } \beta_\phi &:= B_\phi + \tfrac{1}{2}\left(21^T\widehat{\mathbf{b}} - 2\mathbf{y}^T\widehat{\boldsymbol{\theta}} + \textstyle\sum_{i=1}^n \frac{1}{y_i}\right). \end{aligned} \tag{5.25}$$

For gamma LMMs we could use Newton-Raphson iterates for $(\alpha_\phi, \beta_\phi)$, but this would lead to even more complicated expressions than those in Table 5.3.4.

We propose similar fixed point and hybrid fixed point/quasi-Newton approaches to maximising ℓ_L as used for GLMMs in Section 5.2 as Algorithms 8 and 9. Note that, analogous to Algorithms 6 and 7, we set the entries of $\boldsymbol{\beta}_{\sigma^2}$ to be a suitably large constants.

Algorithm 8 Fixed Point Maximisation of Variational Approximation to Bayesian GLMM

1. Initialise $(\boldsymbol{\nu}, \boldsymbol{\beta}_{\sigma^2}, \boldsymbol{\Sigma})$ and set $\alpha_{\sigma^2,i} = A_{\sigma^2,i} + \frac{q_i}{2}$.

2. Cycle

Apply

$$\boldsymbol{\mu} := \boldsymbol{\mu} + (\mathbf{C}^T\mathbf{S}\mathbf{C} + \mathbf{B}_{\boldsymbol{\sigma}^2})^{-1}(\mathbf{C}\boldsymbol{\varepsilon} - \mathbf{B}\boldsymbol{\mu})$$

and update β_ϕ using (5.24) and (5.25).

Until convergence.

3. Cycle

Apply updates for $\boldsymbol{\beta}_{\sigma^2}$ and $\boldsymbol{\Sigma}$ using

$$\begin{aligned} \boldsymbol{\Sigma} &:= (\mathbf{C}^T\mathbf{W}\mathbf{C} + \mathbf{B})^{-1} \\ \text{and } \beta_{\sigma^2,i} &:= \frac{\boldsymbol{\mu}^T\mathbf{B}_i\boldsymbol{\mu} + \text{tr}(\boldsymbol{\Sigma}\mathbf{B}_i)}{q_i},\ 1 \le i \le v. \end{aligned}$$

for $iter = 1, \ldots, L$ **do**

Apply

$$\boldsymbol{\mu} := \boldsymbol{\mu} + (\mathbf{C}^T\mathbf{S}\mathbf{C} + \mathbf{B})^{-1}(\mathbf{C}\boldsymbol{\varepsilon} - \mathbf{B}\boldsymbol{\mu})$$

and update β_ϕ using (5.24) and (5.25).

end for

Until convergence.

Algorithm 9 Quasi-Newton/Fixed Point Hybrid Maximization for Variational Approximation to the Bayesian GLMM

1. Initialise $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}_{\sigma^2}, \boldsymbol{\beta}_{\sigma^2}, \alpha_\phi, \beta_\phi)$.
2. **Cycle**
   - **2.1. Quasi-Newton Steps**
     Using (5.22) with the values in Tables 5.3.4 and 5.2.1–5.2.2 to calculate the derivatives of ℓ_L with respect to $(\boldsymbol{\mu}, \boldsymbol{\alpha}_{\sigma^2}, \boldsymbol{\beta}_{\sigma^2}, \alpha_\phi, \beta_\phi)$. Use 5-10 quasi-Newton updates.
   - **2.2. Fixed Point Step for $\boldsymbol{\Sigma}$**
     Update $\boldsymbol{\Sigma}$ using
     $$\boldsymbol{\Sigma} := (\mathbf{C}^T\mathbf{W}\mathbf{C} + \mathbf{B})^{-1}$$
     where $\mathbf{W}$ can be obtained from Tables 5.2.1–5.2.2.

Until convergence.

Suppose that $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\alpha}}_{\sigma^2}, \widehat{\boldsymbol{\beta}}_{\sigma^2}, \widehat{\alpha}_\phi, \widehat{\beta}_\phi)$ are the values that maximise $[\mathbf{y}; \boldsymbol{\xi}]_L$ then the variational posterior approximations are

$$\begin{aligned} \boldsymbol{\nu}|\mathbf{y} &\sim_{\delta_\nu} N(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}), \\ \sigma_i^2|\mathbf{y} &\sim_{\delta_{\sigma_i^2}} IG(\widehat{\alpha}_{\sigma^2,i}, \widehat{\beta}_{\sigma^2,i}), 1 \le v \le i, \\ \text{and } \phi|\mathbf{y} &\sim_{\delta_\phi} IG(\widehat{\alpha}_\phi, \widehat{\beta}_\phi) \end{aligned} \tag{5.26}$$

and the marginal variational posterior approximations for $\boldsymbol{\nu}$ are

$$\nu_i|\mathbf{y} \sim_{\delta_\nu} N(\widehat{\mu}_i, \widehat{\Sigma}_{ii}). \tag{5.27}$$

5.3.2 *Grid-Based Variational Posterior Approximations*

We will now consider the method of approximating posteriors using the Grid-Based Variational Posterior Approximations (GBVPA) methodology described in Section 4.4 for Bayesian GLMMs. The process for doing this is very similar to approximating the marginal likelihood described in Section 5.3.1.

Grid-Based Variational Posterior Approximation for ν_i

To calculate $\log[\mathbf{y}, \nu_i]_L$ for fixed $\nu_i = \widehat{\nu}_i$ we select $\delta(\boldsymbol{\nu}_{-i}, \boldsymbol{\sigma}^2, \phi|\mathbf{y}) = \delta_{\boldsymbol{\nu}_{-i}}(\boldsymbol{\nu}_{-i})\delta_\phi(\phi)\prod_{i=1}^v \delta_{\sigma_i^2}(\sigma_i^2)$ where the δ-densities are as in (5.19) except that

$$\boldsymbol{\nu}_{-i}|\mathbf{y} \sim_{\delta_{\nu_{-i}}} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Then

$$\begin{aligned} &\log[\mathbf{y}, \nu_i] \ge \log[\mathbf{y}, \nu_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha_\phi, \beta_\phi, \boldsymbol{\alpha}, \boldsymbol{\beta}]_L \\ &\quad = \mathbb{E}_\delta \log[\mathbf{y}|\boldsymbol{\nu}, \phi] + \mathbb{E}_\delta \log[\boldsymbol{\nu}|\boldsymbol{\sigma}^2] + \mathbb{E}_\delta \log[\phi] + \mathcal{H}_{\boldsymbol{\nu}_{-i}} + \mathcal{H}_{\delta_\phi} + \sum_{i=1}^v \mathbb{E}_\delta \log[\sigma_i^2] + \mathcal{H}_{\delta_{\sigma_i^2}} \end{aligned}$$

where the equations (5.20) contain the relevant expectations except, ignoring additive constants,

$$\begin{aligned}\mathbb{E}_\delta \log[\boldsymbol{\nu}|\boldsymbol{\sigma}^2] &= \sum_{i=1}^{v} \frac{q_i}{2}\left(\psi(\alpha_{\sigma^2,i}) - \log(\beta_{\sigma^2,i})\right) - \frac{\widetilde{\boldsymbol{\nu}}^T\mathbf{B}\widetilde{\boldsymbol{\nu}} + \text{tr}(\boldsymbol{\Sigma}\mathbf{B}_{-i,-i})}{2},\\ \text{and } \mathcal{H}_{\boldsymbol{\nu}_{-i}} &= \tfrac{1}{2}\log|\boldsymbol{\Sigma}|,\end{aligned}$$

where $\mathbf{B}_{-i,-i}$ is $\mathbf{B}$ with the ith row and column removed, $\widetilde{\nu}_i = \widehat{\nu}_i$, $\widetilde{\boldsymbol{\nu}}_{-i} = \boldsymbol{\mu}$, Table 5.2.1 contains the values for $\widehat{\boldsymbol{\theta}}$ and $\widehat{\mathbf{b}}$ except that we replace $\boldsymbol{\beta}$ with ν_i, $a(\phi)^{-1}$ with $\widehat{a(\phi)}^{-1}$ and $c(\mathbf{y},\phi)$ with $\widehat{c(\mathbf{y},\phi)}$ which can be obtained from Table 5.3.4. The derivatives with respect to all parameters are given by (5.22) except

$$\begin{aligned}\mathsf{D}_{\boldsymbol{\mu}} \log[\mathbf{y},\nu_i]_L &= \widetilde{\mathbf{C}}^T\boldsymbol{\varepsilon} - \mathbf{B}_{-i}\widetilde{\boldsymbol{\nu}}\\ \mathsf{D}_{\Sigma_{ij}} \log[\mathbf{y},\nu_i]_L &= \text{tr}\left((\boldsymbol{\Sigma}^{-1} - \widetilde{\mathbf{C}}^T\mathbf{W}\widetilde{\mathbf{C}} - \mathbf{B}_{-i,-i})\mathbf{E}_{ij}\right)/2\end{aligned} \tag{5.28}$$

where $\widetilde{\mathbf{C}}$ is the matrix $\mathbf{C}$ with the ith column removed, $\mathbf{B}_{-i}$ is the matrix $\mathbf{B}$ with the ith row removed the values for $\boldsymbol{\varepsilon}$ and $\mathbf{W}$ can be obtained from Table 5.2.2 except $a(\phi)$ is replaced with $\widehat{a(\phi)}$. Updates are the same as in (5.23) except

$$\begin{aligned}\boldsymbol{\mu} &:= \boldsymbol{\mu} + (\widetilde{\mathbf{C}}^T\mathbf{S}\widetilde{\mathbf{C}} + \mathbf{B}_{-i,-i})^{-1}(\widetilde{\mathbf{C}}^T\boldsymbol{\varepsilon} - \mathbf{B}_{-i}\widetilde{\boldsymbol{\nu}})\\ \boldsymbol{\Sigma} &:= (\widetilde{\mathbf{C}}^T\mathbf{W}\widetilde{\mathbf{C}} - \mathbf{B}_{-i,-i})^{-1}\end{aligned}$$

and Algorithms 8 and 9 can be used with a little modification where the value for $\mathbf{S}$ can be obtained from Table 5.2.1-5.2.2.

Grid-Based Variational Posterior Approximation for σ_i^2

To calculate $\log[\mathbf{y},\sigma_i^2]_L$ for fixed $\sigma_i^2 = \widehat{\sigma}_i^2$ we select $\delta(\boldsymbol{\nu},\boldsymbol{\sigma}^2_{-i},\phi|\mathbf{y}) = \delta_{\boldsymbol{\nu}}(\boldsymbol{\nu})\delta_\phi(\phi)\prod_{j\neq i}\delta_{\sigma_j^2}(\sigma_j^2)$ where $\delta_{\boldsymbol{\nu}}$, δ_ϕ and $\delta_{\sigma_j^2}$ are as in (5.19). Then

$$\begin{aligned}\log[\mathbf{y},\sigma_i^2] &\geq \log[\mathbf{y},\sigma_i^2;\boldsymbol{\nu},\boldsymbol{\Sigma},\alpha_\phi,\beta_\phi,\boldsymbol{\alpha}_{\sigma^2,-i},\boldsymbol{\beta}_{\sigma^2,-i}]_L\\ &= \mathbb{E}_\delta \log[\mathbf{y}|\boldsymbol{\nu},\phi] + \mathbb{E}_\delta \log[\boldsymbol{\nu}|\boldsymbol{\sigma}^2] + \log[\sigma_i^2] + \mathbb{E}_\delta \log[\phi] + \mathcal{H}_{\delta_{\boldsymbol{\nu}}} + \mathcal{H}_{\delta_\phi}\\ &\quad + \sum_{j\neq i}\mathbb{E}_\delta \log[\sigma_j^2] + \mathcal{H}_{\delta_{\sigma_j^2}}\end{aligned}$$

where $\boldsymbol{\alpha}_{\sigma^2,-i}$ and $\boldsymbol{\beta}_{\sigma^2,-i}$ are the vectors $\boldsymbol{\alpha}_{\sigma^2}$ and $\boldsymbol{\beta}_{\sigma^2}$ with the ith elements removed, the equations (5.20) contain the relevant expectations with, ignoring additive constants,

$$\begin{aligned}\mathbb{E}_\delta \log[\boldsymbol{\nu}|\boldsymbol{\sigma}^2] &= -\frac{q_i}{2}\log(\widehat{\sigma}_i^2) - \frac{\boldsymbol{\nu}^T\mathbf{B}\boldsymbol{\nu} + \text{tr}\,(\mathbf{B}\boldsymbol{\Sigma})}{2} + \sum_{j\neq i}\frac{q_j}{2}\left(\psi(\alpha_{\sigma^2,j}) - \log(\beta_{\sigma^2,j})\right),\\ \mathbf{B} &= \text{blockdiag}\left\{\sigma_\beta^{-2}\mathbf{I}, \widehat{\sigma}_i^{-2}\mathbf{D}_i + \sum_{j\neq i}(\alpha_{\sigma^2,j}/\beta_{\sigma^2,j})\mathbf{D}_j\right\}\\ \text{and } \log[\sigma_i^2] &= -(A_{\sigma^2,i}+1)\log(\widehat{\sigma}_i^2) - B_{\sigma^2,i}\widehat{\sigma}_i^{-2}.\end{aligned}$$

The derivatives with respect to all parameters are given by (5.22) and the values for $\boldsymbol{\varepsilon}$ and $\mathbf{W}$ can be obtained from Table 5.2.2 except $a(\phi)$ is replaced with $\widehat{a(\phi)}$. Updates are the

same as in (5.23) using the value for $\mathbf{B}$ above and Algorithms 8 and 9 can be used with a little modification.

Grid-Based Variational Posterior Approximation for ϕ

To calculate $\log[\mathbf{y}, \phi]_L$ for fixed $\phi = \widehat{\phi}$ we select $\delta(\boldsymbol{\nu}, \boldsymbol{\sigma}^2) = \delta_{\boldsymbol{\nu}}(\boldsymbol{\nu}) \prod_{j=1}^{v} \delta_{\sigma_j^2}(\sigma_j^2)$ where $\delta_{\boldsymbol{\nu}}$ and $\delta_{\sigma_j^2}$ are as in (5.19). Then

$$\begin{aligned}\log[\mathbf{y}, \phi] &\geq \log[\mathbf{y}, \phi; \widehat{\boldsymbol{\nu}}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}_{\sigma^2}, \boldsymbol{\beta}_{\sigma^2}]_L \\ &= \mathbb{E}_\delta \log[\mathbf{y}|\boldsymbol{\nu}, \phi] + \mathbb{E}_\delta \log[\boldsymbol{\nu}|\boldsymbol{\sigma}^2] + \log[\phi] + \mathcal{H}_{\delta_{\boldsymbol{\nu}}} + \sum_{j=1}^{v} \mathbb{E}_\delta \log[\sigma_j^2] + \mathcal{H}_{\delta_{\sigma_j^2}}\end{aligned}$$

where, ignoring additive constants,

$$\log[\phi] = -(A_\phi + 1)\log(\widehat{\phi}) - A_\phi \widehat{\phi}^{-1}.$$

The equations (5.20) contain the relevant expectations and the derivatives with respect to all parameters are given by (5.22) with the values for $\boldsymbol{\varepsilon}$ and $\mathbf{W}$ obtained from Table 5.2.2, except that the values for $\widehat{a(\phi)}$ and $\widehat{c(\mathbf{y}, \phi)}$ are replaced by the values for $a(\phi)$ and $c(\mathbf{y}, \phi)$ given in Table 1.2.1.

5.4 Numerical Experience

We will consider the accuracy of the above approximations via several studies. For simplicity, in most examples, we will only consider the Poisson and logistic LMM cases. These cases are not only the most common types of GLMM but analysis is simplified by the fact that these cases have no nuisance parameters to deal with. The studies are conducted primarily via simulated data, although one real example will be included. Simulated examples are useful because we can compare computed estimates with the true parameter values, which we cannot do with real examples where "truth" is unknowable. They also enable us to easily examine cases where the underlying means have different levels of roughness, different sample sizes and levels of noise.

We compare various fitting methods for GLMMs including:

- PQL using the `R` function `glmmPQL()` located in the `MASS` package (Venables & Ripley, 2002a, 2002b).
- The variational approximation for the GLMMs as described in Section 5.2, denoted VAR. For the logistic case the use of either ξ or log transforms will be denoted VAR-ξ and VAR-log.
- The variational approximation for the Bayesian GLMMs as described in Section 5.3.1, denoted VB. Again, for the logistic case the use of either ξ or log transforms will be denoted VB-ξ and VB-log.
- Gauss-Hermite Quadrature (GHQ) for one-dimension random effects models (see Section 5.4.2 for details).

- MCMC via the BUGS package in R. We used BUGS to generate chains of length 5,000 after a burn-in of 5,000 and applied a thinning factor of 5, resulting in posterior samples of size 1,000.

In keeping with the recommendations of Crainiceanu, Ruppert & Wand (2005) and Zhao *et al.* (2006) for Bayesian GLMMs we placed diffuse independent $N(0, 10^8)$ priors on the fixed effect parameters and diffuse independent inverse-gamma priors for variance components σ_i^2 with shape and rate parameters both 0.01. Finally, each continuous variable was standardised to improve numerical stability and scale invariance (since the priors are fixed).

As noted in the introduction GLMMs have a number of applications including longitudinal data analysis and function approximation. While it would be an enormous task to cover every application we endeavour to cover some of the more important cases. The situations which we will consider are:

1. **Additive smoothing**. The *trade union* dataset (source: Berndt, 1991) will be fit using VAR-ξ, VAR-log, VB-ξ and VB-log. For the both types of model we will compare running times using the Newton-Raphson/fixed point hybrid (FP), quasi-Newton (QN) and hybrid quasi-Newton/fixed point hybrid (QN/FP). We will also compare the VPA and GBVPA approximations with kernel density estimates of posterior samples obtained via MCMC using the software package BUGS.
2. **Random intercept models.** These are both useful in practice (Diggle, Liang & Zeger, 1994), and because the integrals involved in computing the marginal likelihood are one-dimensional. We can compute the marginal likelihood using relatively simple means. We will compare Poisson and logistic LMMs using PQL, adaptive Gauss-Hermite quadrature and the variational approximation methods VAR-ξ, VAR-log for the logistic case and VAR for the Poisson case.
3. **Scatterplot smoothing.** We compared PQL and the variational approximations presented in this chapter for a variety of scatterplot smoothing settings.

5.4.1 Additive Model Example

In Section 2.5 we considered a penalised spline analysis of the *trade union* dataset which contains trade union membership indicators for a sample of 534 U.S. workers (source: Berndt, 1991) where a subset of the covariates are

$$\mathbf{x}_i = [\texttt{south}_i,\ \texttt{female}_i,\ \texttt{married}_i,\ \texttt{years.educ}_i,\ \texttt{wage}_i,\ \texttt{age}_i].$$

The variables `years.educ`, `years.experience`, `wage` and `age` are continuous and the variables `south`, `female` and `married` are binary. We consider a model of the form

$$\text{logit}\left\{\mathbb{P}\left(\texttt{union.member}_i = 1|\mathbf{x}_i\right)\right\} = f(\mathbf{x}_i)$$

where

$$\begin{aligned} f(\mathbf{x}_i) &= \beta_0 + \mathbb{I}_{\{\texttt{south}_i=1\}}\beta_1 + \mathbb{I}_{\{\texttt{female}_i=1\}}\beta_2 + \mathbb{I}_{\{\texttt{married}_i=1\}}\beta_3 \\ &\quad + f_{\texttt{years.educ}}(\texttt{years.educ}_i) + f_{\texttt{wage}}(\texttt{wage}_i) + f_{\texttt{age}}(\texttt{age}_i) \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \end{aligned}$$

and use the mixed model formulation of cubic O'Sullivan splines, as described in Section 2.5, to model $f_{\texttt{years.educ}}$, $f_{\texttt{wage}}$ and $f_{\texttt{age}}$. We used $K = 25$ inner knots with quantile spacing for each variable. Let $\mathbf{Z}_{\texttt{years.educ}}$, $\mathbf{Z}_{\texttt{wage}}$ and $\mathbf{Z}_{\texttt{age}}$ be the spline matrices for `years.educ`, `wage` and `age` respectively each matrix has $q_i = K + 2$ columns and the matrix $\mathbf{X}$ has 7 columns

$$\mathbf{Z} = \left[\mathbf{Z}_{\texttt{years.educ}}, \mathbf{Z}_{\texttt{wage}}, \mathbf{Z}_{\texttt{age}}\right]$$

and

$$\mathbf{D}_{\sigma^2} = \underset{1\le i\le 3}{\text{blockdiag}} \left\{\sigma_i^{-2}\mathbf{I}_{q_i}\right\}.$$

Comparing Running Times

We compare running times of PQL and MCMC via `BUGS` with VAR-ξ, VAR-log, VB-ξ and VB-log. Each of the variational approximations were fitted using the FP, QP and FP/QN approaches. Each algorithm was run 20 times, except MCMC which was run 5 times, to get an indication of running times of each algorithm. The mean running times are summarised in Table 5.4.5 and each of the fits for the continuous components are illustrated in Figure 5.2.

Approximation	Algorithm	Time (s)
PQL	`glmmPQL`	26.49
VAR-ξ	FP	33.82
	QN	43.38
	QN/FP	3.46
VAR-log	FP	34.89
	QN	46.35
	QN/FP	2.59
VB-ξ	FP	32.70
	QN	252.21
	QN/FP	14.04
VB-log	FP	68.07
	QN	161.74
	QN/FP	21.03
MCMC	`WinBugs`	3729.22

Table 5.4.5: *The mean running times for each method fitting the* `trade union` *model as described in Section 5.4.1.*

Comparing the running times from Table 5.4.5 we see that the FP and QN approaches tended to be slower than the QN/FP hybrid approach. While the cost of each iteration for FP was smaller than QN the FP approach took many hundreds of iterations to converge. This is consistent with the discussion in Section 4.3 and Section 5.2.2.

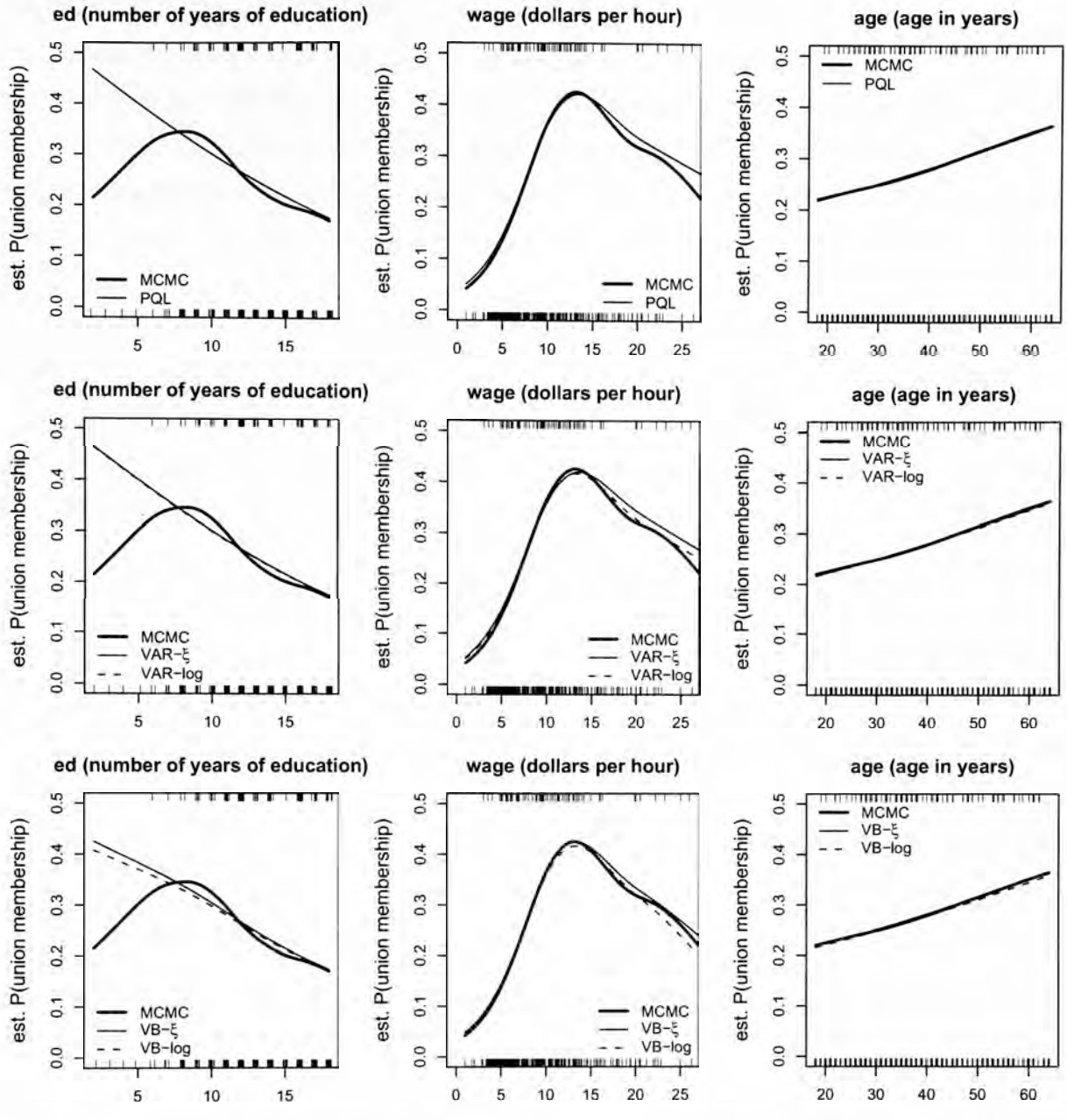


Figure 5.2: *Smooth function fits for the* `Trade Union` *model using the MCMC, PQL, VAR-ξ, VAR-*log*, VB-ξ and VB-*log *approximations.*

From Figure 5.2 we see that the VAR-ξ, VAR-log, VB-ξ and VB-log approximations produce fits which are quite similar to PQL. The approximations for the Bayesian logistic LMM, i.e. VB-ξ and VB-log were, to the eye, slightly closer to the MCMC fit than the other approximations. While the analytic approximations are less accurate than MCMC methods they are extremely fast; taking from seconds to 6 minutes to fit each model depending on the fitting algorithm used. In contrast, for this example, the MCMC approximation using `BUGS` took a little over an hour.

Posterior Density Approximations

Finally, for Bayesian logistic LMMs we can compare posterior density approximations for `south`, `female` and `married` and variance components associated with $f_{\texttt{years.educ}}$, $f_{\texttt{wage}}$

and f_{age}. For the MCMC fit we used BUGS to generate chains of length 50,000 after a burn-in of 5,000 and applied a thinning factor of 5, resulting in posterior samples of size 10,000. and then used the R package KernSmooth to estimate the densities of these posterior samples. We also used VPA as described in Section 5.3.1 and GBVPA approximations as described in Section 5.3.2. These posterior approximations are illustrated in Figure 5.3.

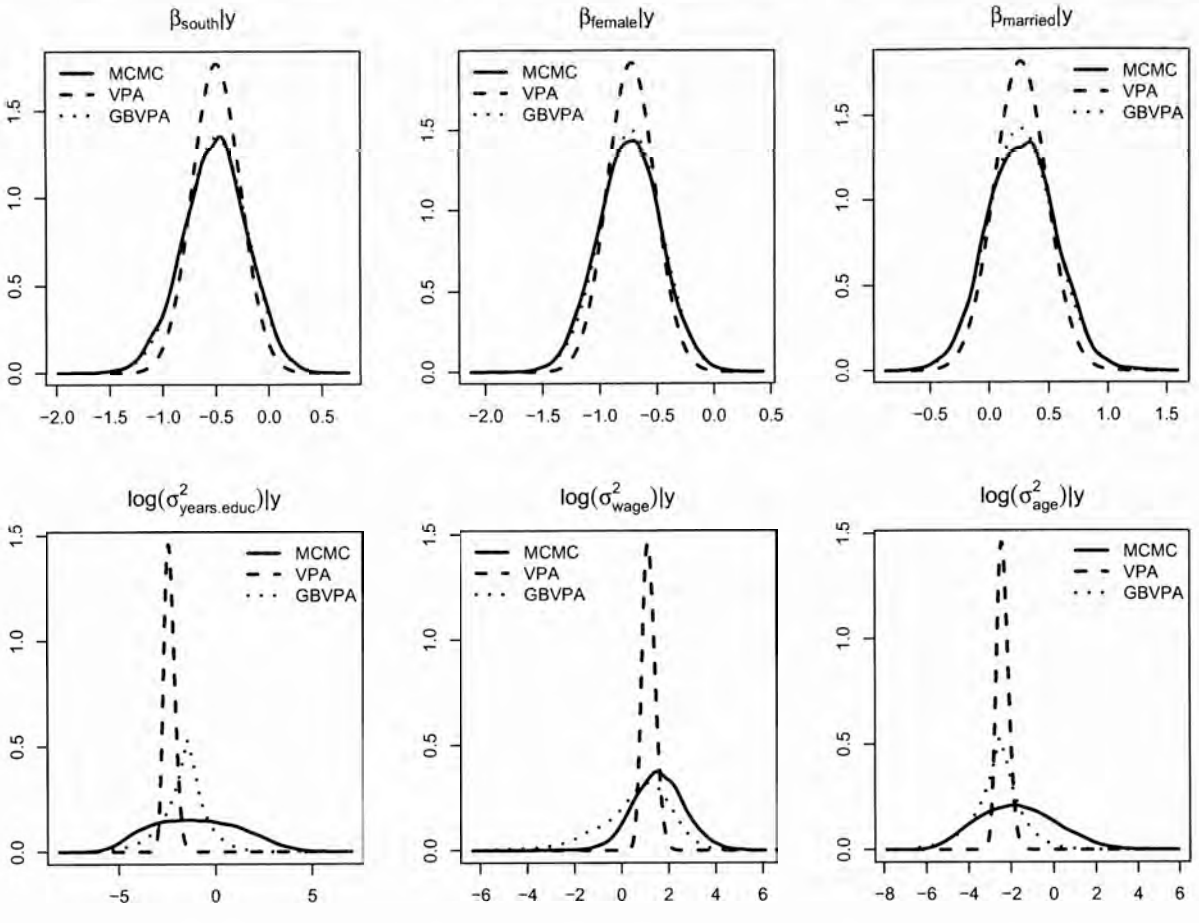


Figure 5.3: *Illustration of the kernel density estimates of MCMC posterior samples, variational posterior approximations (VPA) and grid-based variational posterior approximations (GBVPA) for* `south`, `female` *and* `married` *coefficients and 3 variance components for* `years.educ`, `wage` *and* `age` .

From Figure 5.3 we notice that each of the VPA approximations underestimate the amount of variance of the posteriors, particularly for the variance components. The GB-VPA approximations, on the other hand, were significantly better particularly for the `south`, `female` and `married` coefficients, although the GBVPA approximations for the variance components where clearly not as accurate as for the models examined in Chapter 4. We speculate that this may be because of the use of an additional approximation to obtain marginal posteriors, i.e. the ξ-transform. Each GBVPA approximation took roughly 5 minutes to compute using the FP while the MCMC approach via BUGS took a little over 6 hours.

5.4.2 *Random Intercept Models*

We now consider random intercept models (see McCulloch & Searle, 2001, Section 8.4). Suppose that data are grouped into correlated clusters and are thought to come from either a Poisson or Bernoulli distribution. For example Diggle *et al.* (1994) considered the number of epileptic seizures in patients on a drug or placebo. Alternatively, for the Bernoulli case, we might consider whether an epileptic patients has any seizures while on a drug or placebo.

Let y_{ij} denote the jth count (for the Poisson case) or indicator (for the Bernoulli case) taken on the ith cluster. In both cases we use the canonical link and use the normal distribution to model the random cluster (patient) effects. The random effects u_i are shared among observations within the same cluster and hence those observations are being modelled as correlated. Thus we might consider the model

$$\begin{aligned} \log[y_{ij}|u_i] &= y_{ij}(\mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i) - b(\mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i) + c(y_{ij}); \quad i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n_i; \\ u_i &\sim N(0, \sigma^2) \end{aligned}$$

where b and c can be taken from Table 1.2.1.

For the random intercept model data the log-likelihood can be simplified to

$$\ell(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{m} \log \int [\mathbf{y}_i|u_i; \boldsymbol{\beta}, \phi][u_i|\sigma^2]du_i = \sum_{i=1}^{m} \log \int \exp\left\{f_i(\boldsymbol{\beta}, \sigma^2, u_i)\right\} du_i \qquad (5.29)$$

where, ignoring additive constants,

$$f_i(\boldsymbol{\beta}, \sigma^2, u_i) = \mathbf{y}_i^T(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i u_i) - \mathbf{1}^T b(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i u_i) - \tfrac{1}{2}\log(\sigma^2) - \tfrac{u_i^2}{2\sigma^2}.$$

Since each integral in (5.29) is one dimensional we can use GHQ to accurately approximate $\ell(\boldsymbol{\beta}, \sigma^2)$. Although there is an `R` package `glmmML` which claims to do just this, we have found this package to be unreliable. Instead we implement a similar procedure ourselves.

It should be noted that there may be problems with GHQ if each integral is not shifted and scaled appropriately (Liu & Pierce, 1994; McCulloch & Searle, 2001). Instead we use an adaptive GHQ scheme developed by Liu & Pierce (1994) to perform the appropriate integration.

Let

$$\widehat{u}_i = \underset{u_i}{\text{argmax}} \left\{f_i(\boldsymbol{\beta}, \sigma^2; u_i)\right\} \quad \text{and} \quad \widehat{\sigma}_i^2 = \left[-\frac{\partial f_i}{\partial u_i^2}\right]^{-1}.$$

Using these $\exp\left\{f_i(\boldsymbol{\beta}, \sigma^2, u_i)\right\}$ is "most similar" to a multivariate Gaussian distribution in u_i with mean $\widehat{u}_i$ and variance $\widehat{\sigma}_i^2$. Using the transformation $\sigma^2 = e^\gamma$ the log-likelihood

can then be written as

$$\begin{aligned}\ell(\boldsymbol{\beta},\gamma) &= \sum_{i=1}^{m}\log\int\exp\left\{f_i(\boldsymbol{\beta},\gamma,u_i)+\frac{(u_i-\widehat{u}_i)^2}{2\widehat{\sigma}_i^2}\right\}\exp\left\{-\frac{(u_i-\widehat{u}_i)^2}{2\widehat{\sigma}_i^2}\right\}du_i\\ &= \sum_{i=1}^{m}\log\int\exp\left\{f_i(\boldsymbol{\beta},\gamma,\widehat{u}_i+\sqrt{2}\widehat{\sigma}_i\widetilde{u}_i)+\log(\sqrt{2}\widehat{\sigma}_i)+\widetilde{u}_i^2\right\}e^{-\widetilde{u}_i^2}d\widetilde{u}_i\end{aligned}$$

using the change of variables $u_i = \widehat{u}_i + \sqrt{2}\widehat{\sigma}_i\widetilde{u}_i$. We can now use standard GHQ to approximate $\ell(\boldsymbol{\beta},\gamma)$ by

$$\ell(\boldsymbol{\beta},\gamma)\approx\widehat{\ell}(\boldsymbol{\beta},\gamma)=\sum_{i=1}^{m}\log\left[\sum_{j=1}^{N}w_j\exp\left\{f_{ij}(\boldsymbol{\beta},\sigma^2)\right\}\right]$$

where, ignoring additive constants,

$$\begin{aligned}f_{ij}(\boldsymbol{\beta},\gamma) &= \mathbf{y}_i^T\boldsymbol{\eta}_{ij}-\mathbf{1}^Tb(\boldsymbol{\eta}_{ij})-\frac{\gamma}{2}-\frac{(\widehat{u}_i+\sqrt{2}\widehat{\sigma}_it_j)^2e^{-\gamma}}{2}+\log\widehat{\sigma}_i+t_j^2\\ \text{and}\qquad \boldsymbol{\eta}_{ij} &= \mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i(\widehat{u}_i+\sqrt{2}\widehat{\sigma}_it_j).\end{aligned}$$

The first derivatives of ℓ with respect to $(\boldsymbol{\beta},\sigma^2)$ can be written as

$$\mathsf{D}_{\beta_k}\widehat{\ell}=\sum_{i=1}^{m}\sum_{j=1}^{N}\omega_{ij}\mathsf{D}_{\beta_k}f_{ij}\qquad\text{and}\qquad\mathsf{D}_{\gamma}\widehat{\ell}=\sum_{i=1}^{m}\sum_{j=1}^{N}\omega_{ij}\mathsf{D}_{\gamma}f_{ij}$$

where

$$\begin{aligned}\omega_{ij} &= w_j\exp\left\{f_{ij}(\boldsymbol{\beta},\gamma)\right\}\Big/\sum_{j=1}^{N}w_j\exp\left\{f_{ij}(\boldsymbol{\beta},\gamma)\right\},\\ \mathsf{D}_{\boldsymbol{\beta}}f_{ij} &= \mathbf{X}_i^T\left(\mathbf{y}_i-b'(\boldsymbol{\eta}_{ij})\right)\\ \text{and}\ \mathsf{D}_{\gamma}f_{ij} &= -\tfrac{1}{2}+\frac{(\widehat{u}_i+\sqrt{2}\widehat{\sigma}_i^2t_j)^2e^{-\gamma}}{2}\end{aligned}$$

with $b'(x)=\frac{e^x}{1+e^x}$ for logistic regression models and $b'(x)=e^x$ for Poisson regression models. The $\widehat{u}_i$s satisfy

$$\mathsf{D}_{u_i}f_i(\widehat{u}_i)=\mathbf{Z}_i^T(\mathbf{y}-b'(\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\widehat{u}_i))-\widehat{u}_ie^{-\gamma}=0$$

and

$$\widehat{\sigma}_i^2=\left(\mathbf{Z}_i^T\mathrm{diag}(b''(\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\widehat{u}_i))\mathbf{Z}_i+e^{-\gamma}\right)^{-1}$$

with $b''(x)=e^x/(1+e^x)^2$ for logistic regression models and $b''(x)=e^x$ for Poisson regression models.

At this point we deviate from Liu & Pierce (1994) due to problems in the practical implementation of the above method. While this modification was not stated explicitly in Liu & Pierce (1994) it had doubtless been used elsewhere in practical implementations. For Poisson random intercept models in particular if $\boldsymbol{\beta}$ and/or σ^2 are large then numerical overflow can occur when evaluating $\widehat{\ell}$ or ω_{ij} if care is not taken. For this reason we

instead evaluate $\widehat{\ell}$ and ω_{ij} using the formula

$$\begin{aligned}\widehat{\ell}(\boldsymbol{\beta},\gamma) &= \sum_{i=1}^{m} F_i^* + \log \sum_{j=1}^{N} w_j \exp\left(f_{ij} - F_i^*\right) \\ \text{and } \omega_{ij} &= w_j \exp\left(f_{ij} - F_i^*\right) \Big/ \sum_{j=1}^{N} w_j \exp\left(f_{ij} - F_i^*\right)\end{aligned}$$

where

$$F_i^* = \max_j \left\{\log(w_j) + f_{ij}\right\}.$$

These equations provide sufficient information to maximise $\widehat{\ell}$ using the quasi-Newton optimisation algorithm implemented in the `optim()` function in the standard `R` library.

We will compare this routine with the PQL algorithm implemented in the `R` function `glmmPQL()` in the package `MASS`, the variational approximations VAR-ξ and VAR-log for the logistic case and VAR in the Poisson case.

For the Poisson case each routine we fitted GLMMs for the true β taking 21 values from $-2, \ldots, 2$ for $\sigma^2 \in \{1/10, 1/2, 1\}$ with $n_i = 20$ and $m = 40$ for 100 trails while for the logistic case the true β takes 21 values from $-5, \ldots, 5$ and $\sigma^2 \in \{1, 3\}$ with $n_i = 20$ and $m = 40$. The median absolute bias for β and σ^2 is plotted as a function of β for each case and illustrated in Figures 5.4 and 5.5.

Based on these figures for the Poisson case, the median absolute biases for both PQL and VAR approximations were quite close to those for GHQ. Furthermore, in the Poisson case using the above settings, none of the methods seemed to dominate any of the other methods in terms of median absolute bias. For the logistic case the VAR-ξ and VAR-log approximations did not perform as well as the PQL and GHQ approximations, particularly when the true value for the variance component σ^2 was small. On the other hand in the logistic case for these settings PQL performed reasonably well compared to GHQ.

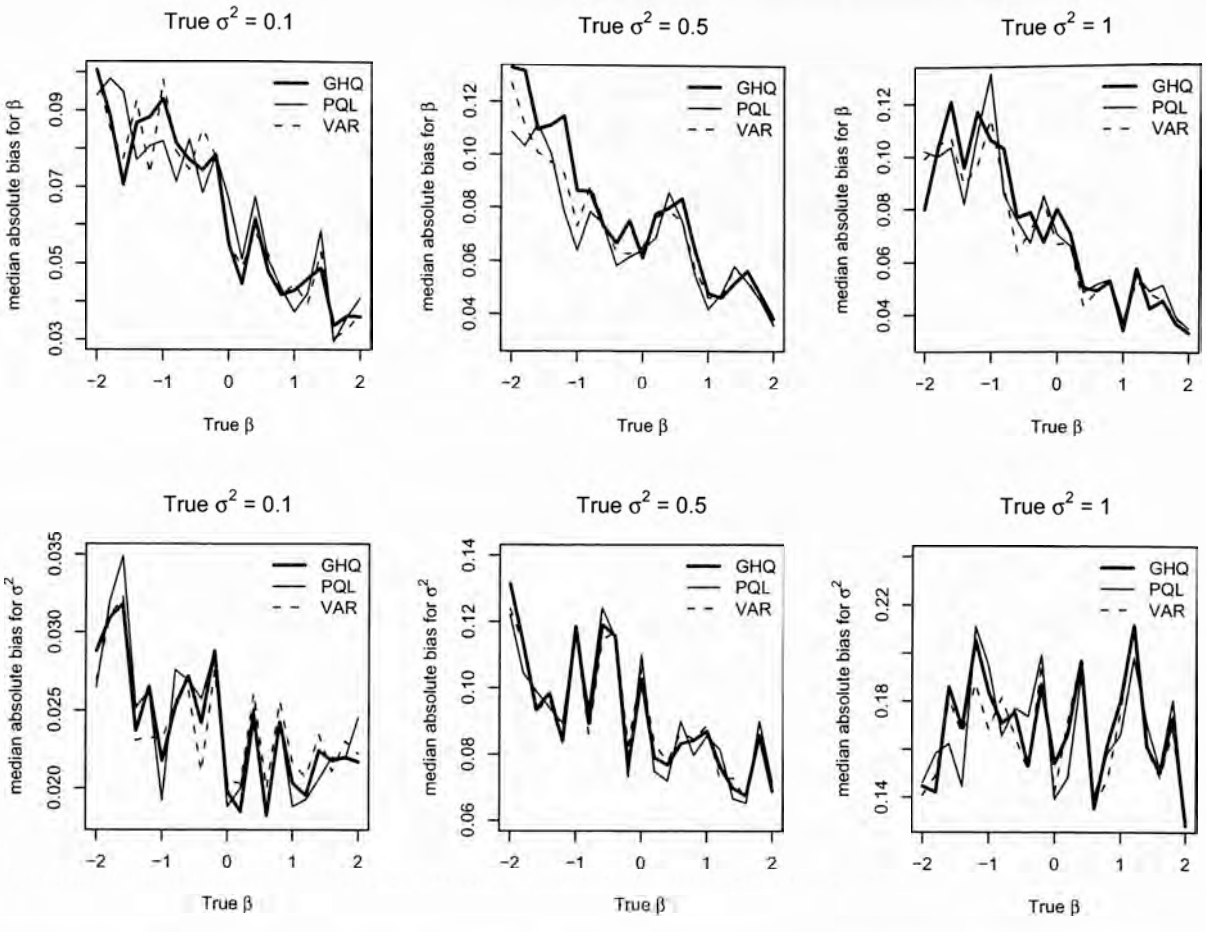


Figure 5.4: *Median absolute biases for Poisson random intercept data. The simulation used* 100 *instances where the true* β *took values from* $-2, \ldots, 2$ *and the true* σ^2 *took values* $\sigma^2 \in \{1/10, 1/2, 1\}$ *with* $n_i = 20$ *observations and* $m = 40$ *groups.*

It has been shown, based on small σ^2 asymptotics, that PQL can have significant biases (Breslow & Lin, 1995; Lin & Breslow, 1996; Sutradhara & Rao, 1998). Breslow & Lin (1995) demonstrated that PQL was particularly biased for analysis random intercept models of paired samples with binomial data. To test the effectiveness for these settings for the Poisson case each routine we fitted GLMMs for the true β taking 11 values from $-2, \ldots, 2$ for $\sigma^2 = 0.1$ with $n_i = 2$ and $m = 200$ for 200 trails while for the logistic case the true β takes 11 values from $-4, \ldots, 4$ and $\sigma^2 = 0.1$ with $n_i = 2$ and $m = 400$. The median absolute bias for β and σ^2 is plotted as a function of β for each case and illustrated in Figures 5.6 and 5.7.

Based on Figure 5.6 VAR performed as well as GHQ for these settings. In particular the median absolute bias for PQL was particularly large for small true β values. Furthermore the median absolute biases for GHQ and VAR were smaller than those for PQL for most values of β.

The results for summarised in Figure 5.7 is far less convincing. While the median absolute biases for GHQ, VAR-ξ and VAR-log were smaller in comparison to PQL it appears that for on the edges of the tested true β values VAR-ξ and VAR-log achieved this by estimating σ^2 close to 0.

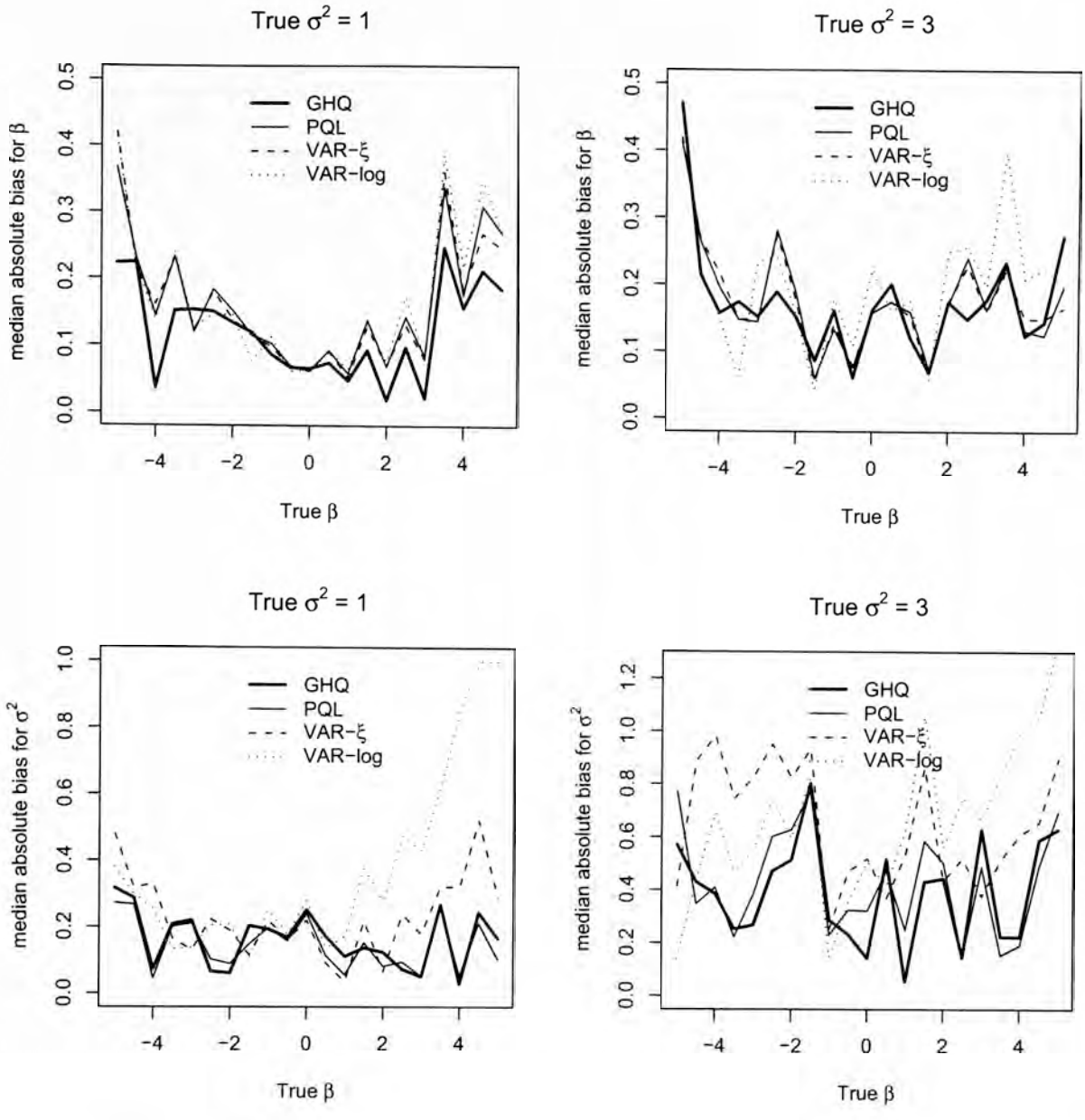


Figure 5.5: *Median absolute biases for logistic random intercept data. The simulation used* 100 *instances where the true* β *took values from* $-5, \ldots, 5$ *and the true* σ^2 *took values* $\sigma^2 \in \{1, 3\}$ *with* $n_i = 20$ *observations and* $m = 40$ *groups.*

Rijmen & Vomlel (2007) performed more a more extensive comparison study between the Laplace approximation (achieved by setting $N = 1$ for GHQ, see Lui & Pearce, 1994) and VAR-ξ. They concluded that shrinkage of the σ^2 estimate was more pronounced for VAR-ξ when the number of observations per group was small. The settings we considered for Figure 5.7 is an extreme cases of this and is consistent with these conclusions.

5.4.3 Scatterplot Smoothing

A number of scatterplot smoothing experiments were conducted to access the speed and accuracy of the variational approximations developed in this chapter. For these experiments we will compare these approximations with the PQL approximation (Breslow & Clayton, 1993). This is the main competitor with these methods when speed is highly

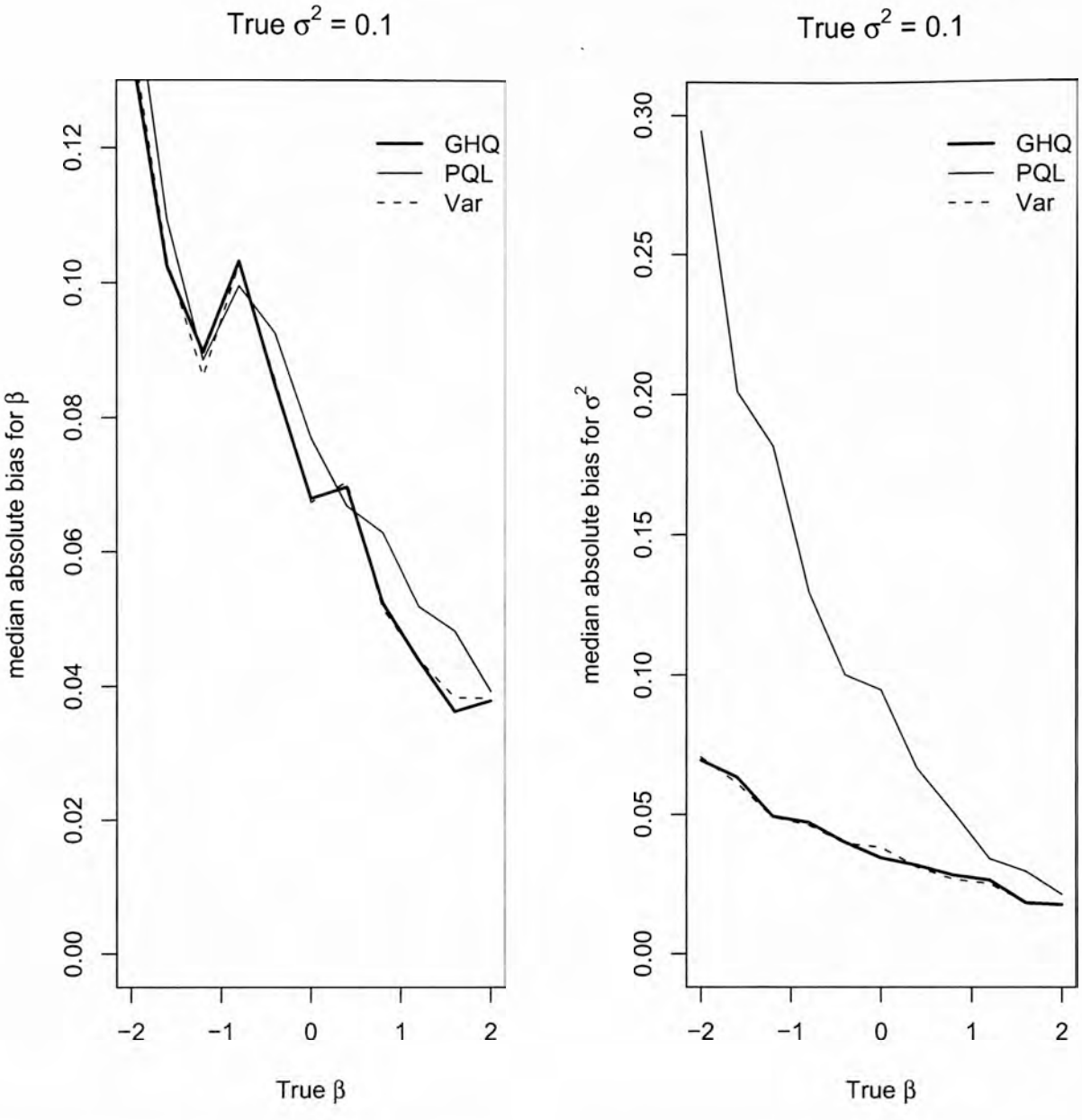


Figure 5.6: *Median absolute biases for Poisson random intercept data. The simulation used* 200 *instances where the true* β *took values from* $-2, \ldots, 2$ *and the true* σ^2 *took the value* $\sigma^2 = 0.1$ *with* $n_i = 2$ *observations and* $m = 200$ *groups.*

desirable. Thus, we will compare are the VAR and VB with PQL for the Poisson case and VAR-ξ, VAR-log, VB-ξ and VB-log with PQL for the logistic case.

For the smoothing and additive model examples we will use the following four functions as the true $\mathbb{E}(y|\eta(x)) = \mu(\theta(\eta(x)))$ where $\mu(\cdot)$ can be obtained from Table 1.1 and:

$$\begin{aligned}
\eta_1(x) &= \sqrt{x(1-x)}\sin\left(\frac{2\pi\left(1+2^{(9-4J)/5}\right)}{x+2^{(9-4J)/5}}\right), \\
\eta_2(x) &= \sin(8(x-0.5)) + 2\exp(-16^2(x-0.5)^2), \\
\eta_3(x) &= 2\sin(2\pi x^{1.3}) \\
\text{and } \eta_4(x) &= \frac{3}{2}\phi\left(\frac{3x-7}{3}\right) - \phi\left(25x-20\right).
\end{aligned}$$

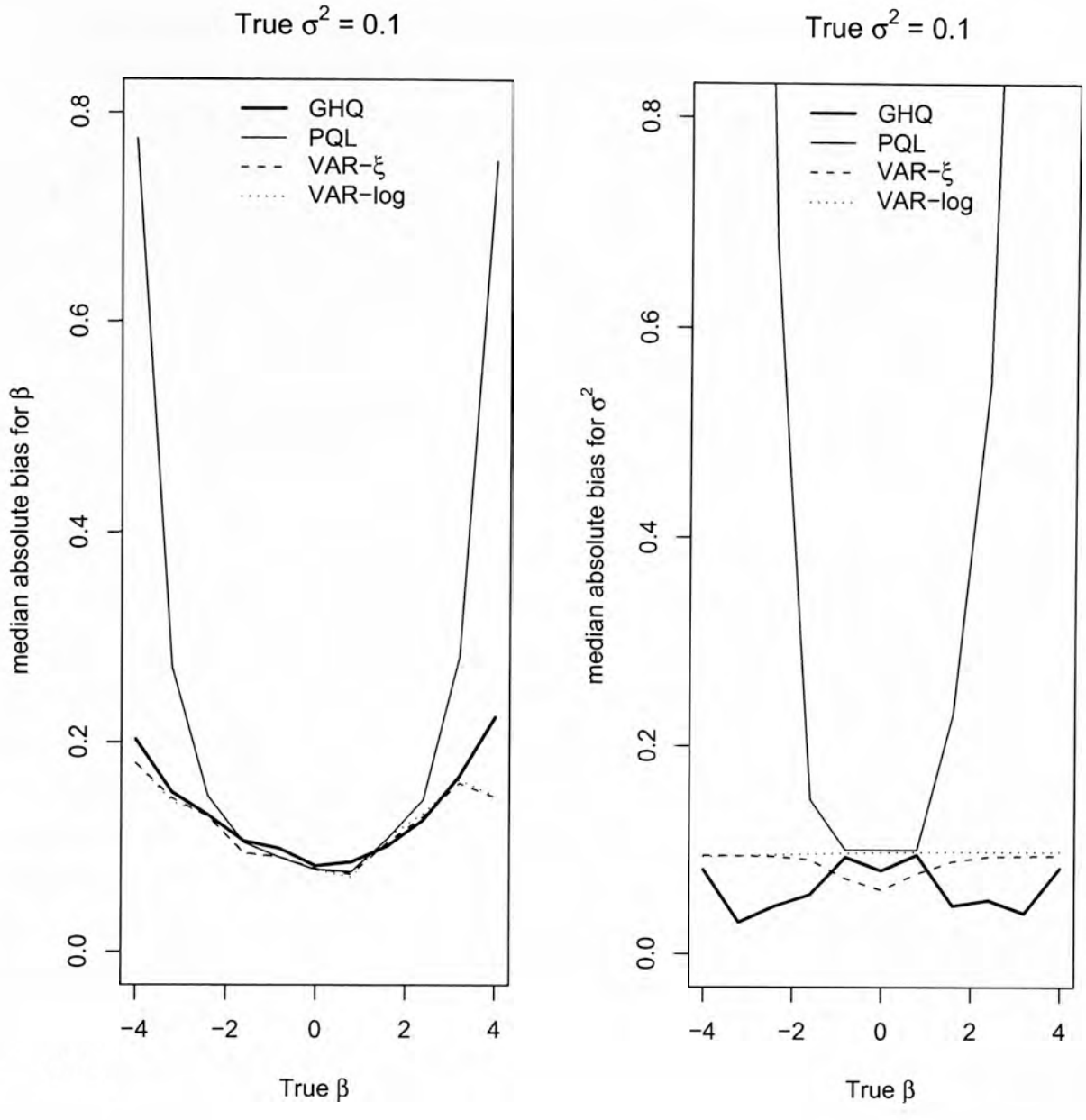


Figure 5.7: *Median absolute biases for logistic random intercept data. The simulation used* 100 *instances where the true* β *took values from* $-4, \ldots, 4$ *and the true* σ^2 *took the value* $\sigma^2 = 0.1$ *with* $n_i = 2$ *observations and* $m = 400$ *groups.*

These represent a variety of spatially inhomogeneous functions and simpler nonlinear functions. The functions $\eta_1, \ldots, \eta_4$ are illustrated in Figure 5.8.

Although there are many aspects of these approximations which we could study we will focus on three: sample size, function complexity and number of knots. While varying these aspects we will compare the above approximations using the mean deviance measure of error

$$\overline{\mathcal{D}}(\boldsymbol{\mu}^*, \widehat{\boldsymbol{\mu}}) \quad = n^{-1} \sum_{i=1}^{n} \mathcal{D}(\mu_i^*, \widehat{\mu}_i)$$

where Table 1.2.2 contains the deviances $\mathcal{D}$ for different generalised linear model families, $\mu_i^* = \mu(\theta(a\widetilde{\eta}(x_i) + b))$ is the true mean and $\widehat{\boldsymbol{\mu}} = \mu(\theta(\mathbf{C}\widehat{\boldsymbol{\nu}}))$ is the estimated mean, $\widetilde{\eta}_i$ for

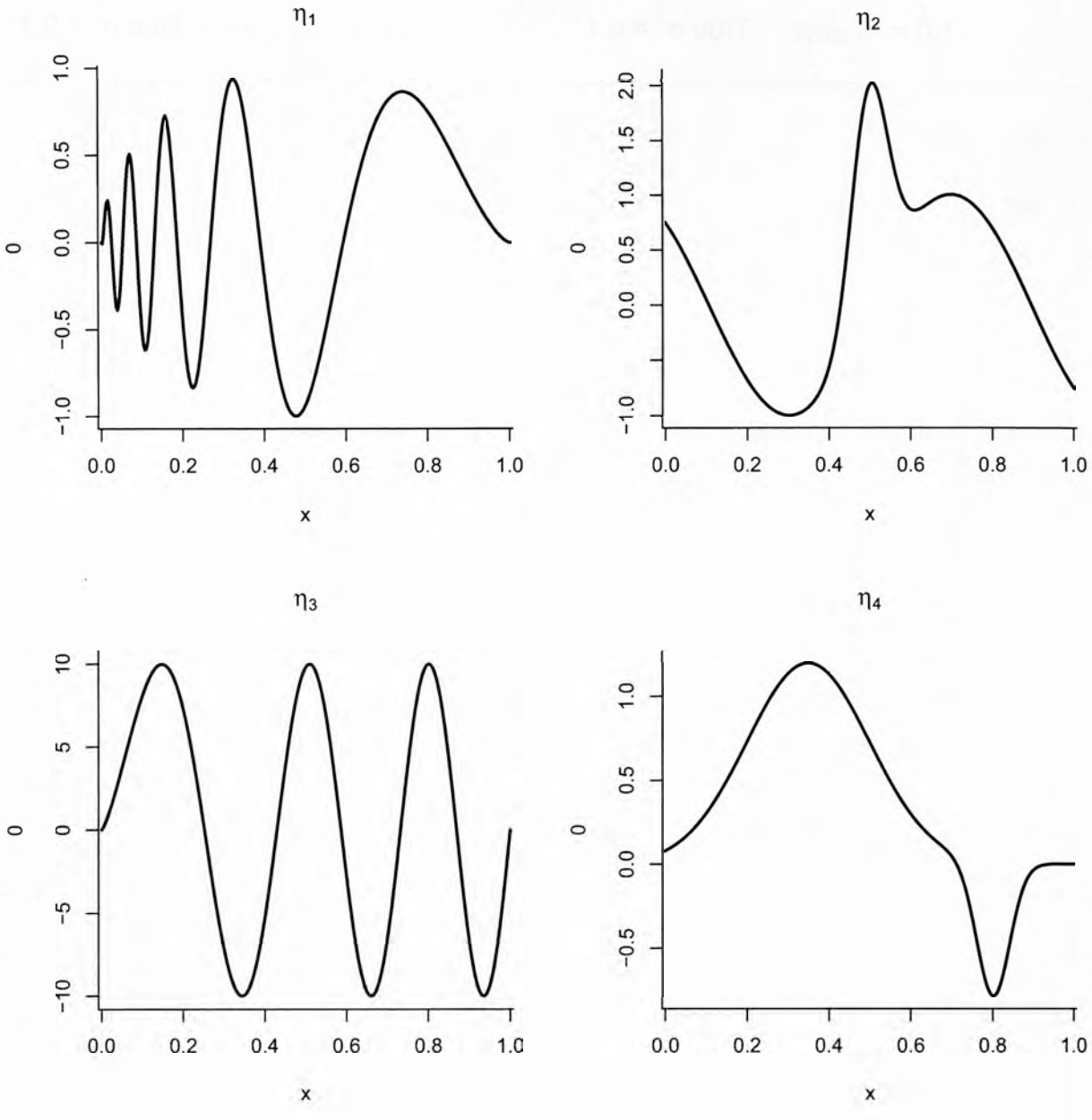


Figure 5.8: *Test functions to be used in scatterplot smoothing simulations.*

$1 \leq i \leq 4$ are the functions $\eta_i(\cdot)$ scaled so that the minimum value is 0 and the maximum value is 1, and (a, b) are shift and scaling constants.

We performed the following simulations:

1. **Sample Size.** For a fixed number of knots $K = 35$, for functions $\eta_1, \ldots, \eta_4$ with $J = 5$ for η_1 for 50 simulations we use for the Poisson case scaling $(a = 3, b = -1)$ and $n = \{250, 500, \ldots, 3750, 4000\}$ while for the logistic case we use scaling $(a = 10, b = -5)$ and $n = \{200, 400, \ldots, 3000, 3200\}$. Using these 50 simulations the median times and median mean deviances over these simulations for these settings are illustrated for in Figure 5.9 for the Poisson case and Figure 5.11 for the logistic case. Differences in median mean deviances for VAR and VB with median mean deviances for PQL are illustrated for in Figure 5.10 for the Poisson case. Differences

in median mean deviances for VAR-ξ, VAR-log, VB-ξ and VB-log with median mean deviances for PQL are illustrated for in Figure 5.12 for the logistic case.

2. **Number of knots and complexity.** For function η_1 with $J = \{1, \ldots, 6\}$ for η_1 we vary the number of knots $K = \{5, 10, \ldots, 45, 50\}$ using scaling $(a = 3, b = -1)$ for the Poisson case and $(a = 10, b = -5)$ for the logistic case. Using these 50 simulations the median times and median mean deviances over these simulations for these settings are illustrated for in Figure 5.13 for the Poisson case and Figure 5.15 for the logistic case. Differences in median mean deviances for VAR and VB with median mean deviances for PQL are illustrated for in Figure 5.14 for the Poisson case. Differences in median mean deviances for VAR-ξ, VAR-log, VB-ξ and VB-log with median mean deviances for PQL are illustrated for in Figure 5.16 for the logistic case.

Based on the Figures 5.9, 5.11, 5.13 and 5.15 we see that the variational approximations have similar accuracy to PQL for the scatterplot smoothing problems considered in this section. In each case optimisation was performed using the FP approach as described in Algorithm 6 and Algorithm 8. For most settings these approaches are faster or have similar running times to the `glmmPQL` function in `R`.

Considering Figure 5.10 and Figure 5.14 we see that in the Poisson case VAR has similar accuracy to PQL but may have a slight advantage for mean functions with more roughness, however these differences are still minor. For Figure 5.12 and Figure 5.16 the log transform might have a slight advantage over the the ξ transform and VB might have a slight advantage over VAR for these settings. However again these differences are relatively minor.

One aspect not apparent in these plots is numerical stability. We found that Algorithm 6 and Algorithm 8 often failed when using the ξ-transform for the logistic case if the problem was not scaled well, i.e. the values for a and b were not chosen favourably. Even for the experiments above Algorithm 6 and Algorithm 8 failed about 5% of the time. The times and deviances were not included in the results.

It is unclear from the above experiments that the variational method using the Gaussian density transform performs better or worse, in terms of mean deviance, than PQL or vice versa, in a general sense. In each of the above experiments there are cases where the variational method performs better than PQL and there are cases where PQL performs better than the variational method.

5.5 Conclusion

We have shown in this chapter that variational approximations are a fresh alternative method for fitting GLMMs. Although, due to the sheer number of potential models we could consider, we have only performed limited empirical studies be believe that variational approximations offer an effective alternative for fitting GLMMs and Bayesian GLMMs.

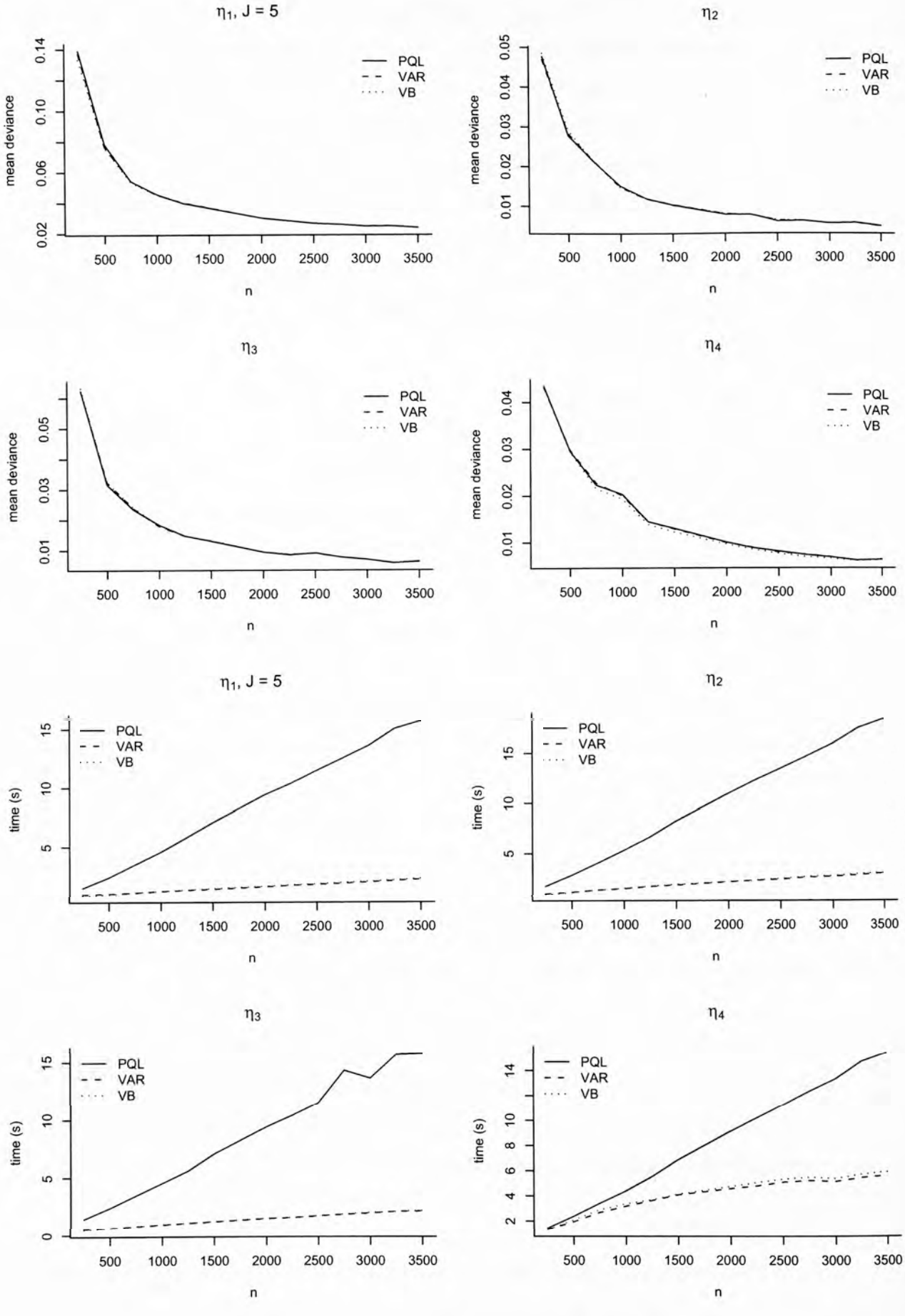


Figure 5.9: *Mean deviances (top six panels) and running times (bottom six panels) for the varying* ***sample size*** *simulations for the Poisson case (see text for details).*

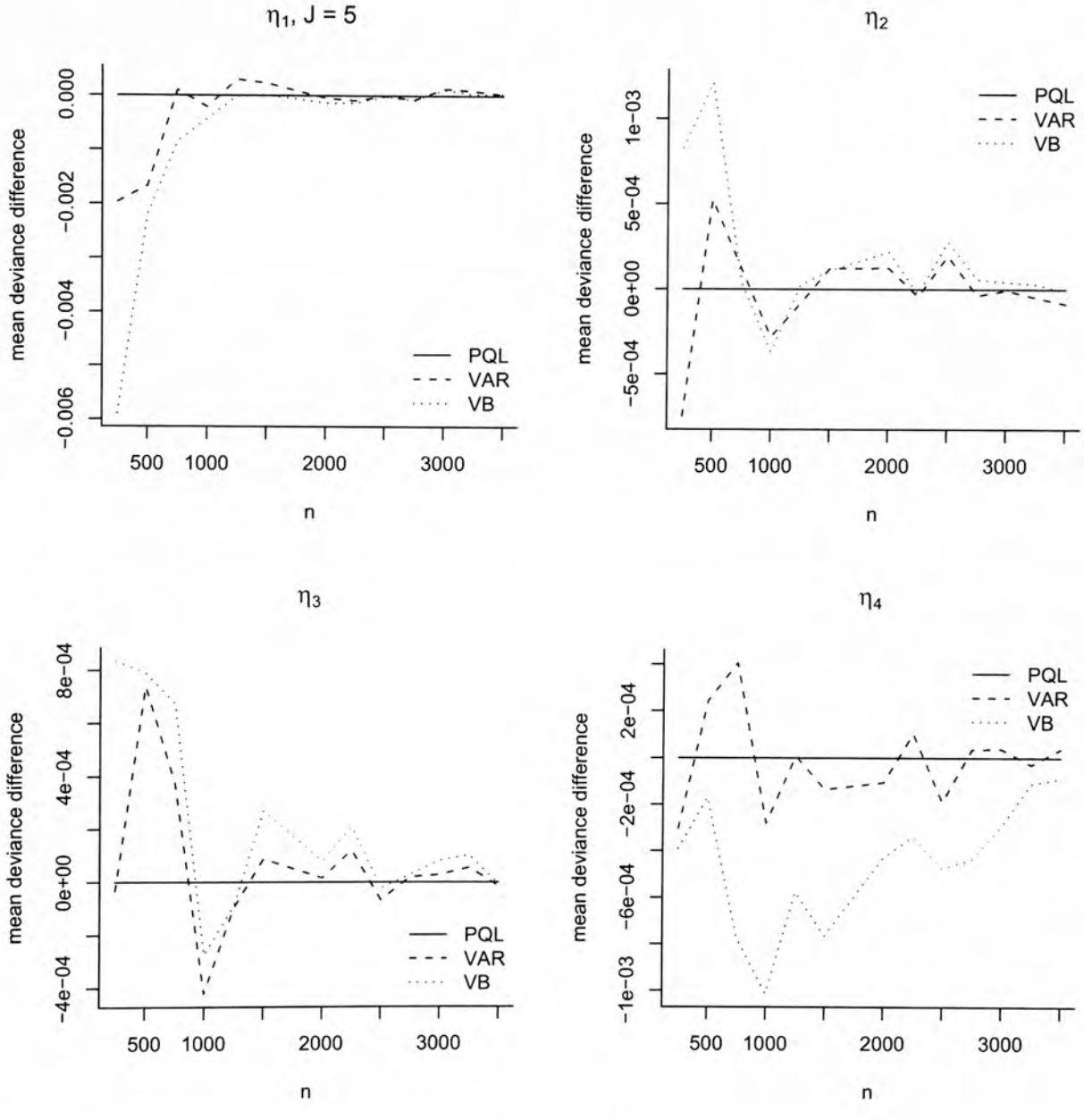


Figure 5.10: *Differences in mean deviances with mean deviances for PQL for the varying* ***sample size*** *simulations for the Poisson case (see text for details).*

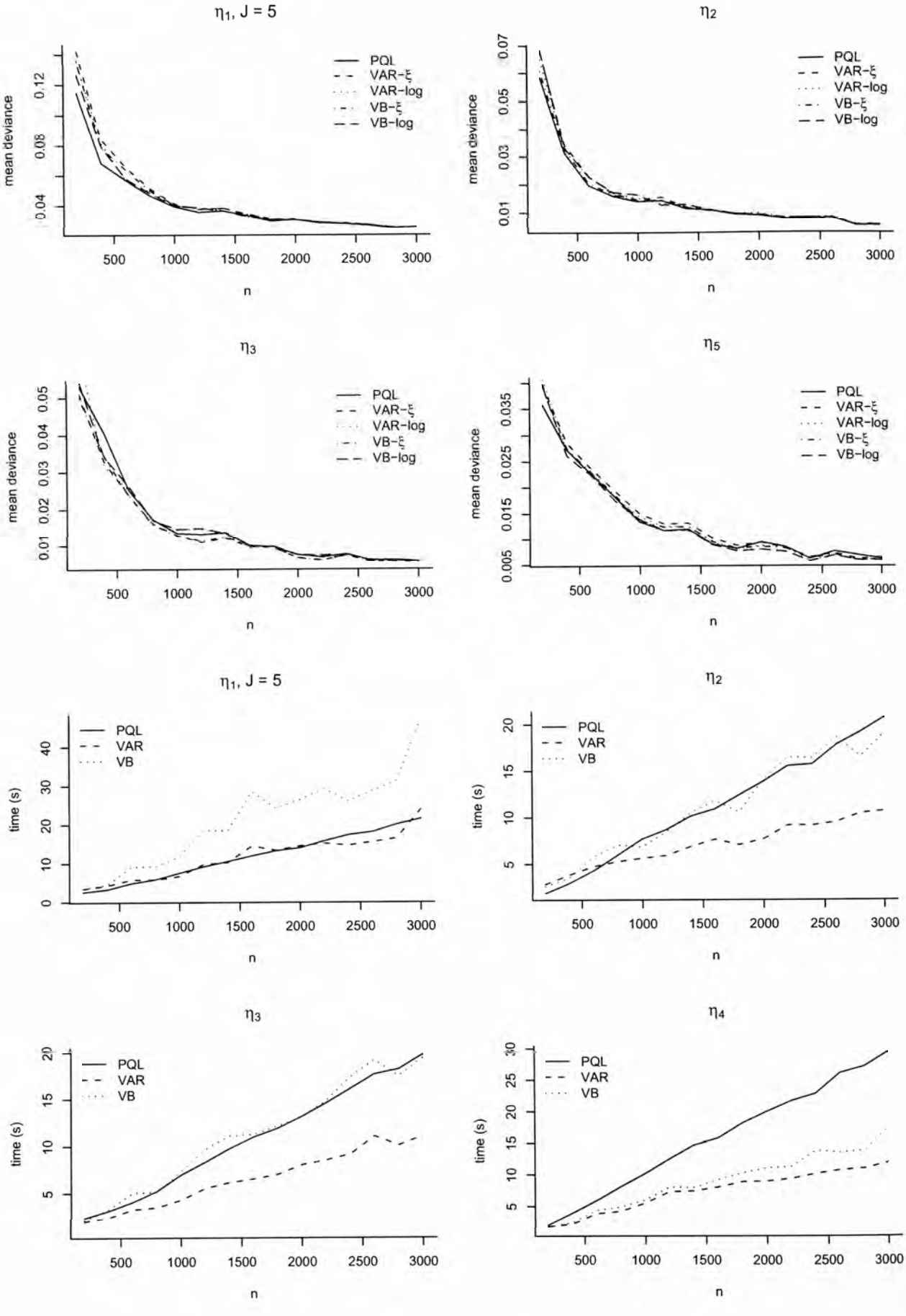


Figure 5.11: *Mean deviances (top six panels) and running times (bottom six panels) for the varying* ***sample size*** *simulations for the logistic case (see text for details). Note running times we averaged for VAR-ξ and VAR-*log *and VB-ξ and VB-*log *for clearer presentation.*

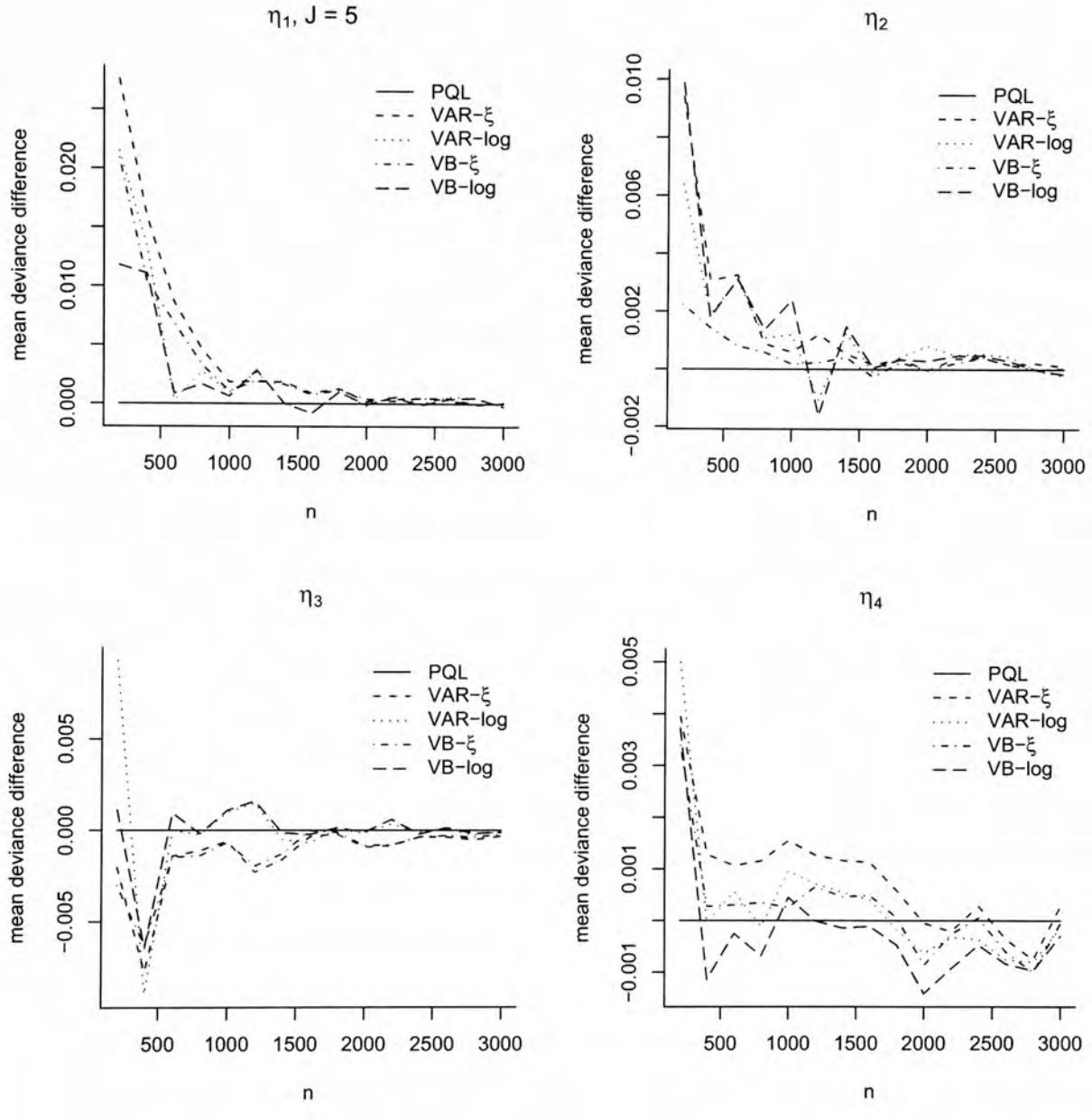


Figure 5.12: *Differences in mean deviances with mean deviances for PQL for the varying* ***sample size*** *simulations for the logistic case (see text for details).*

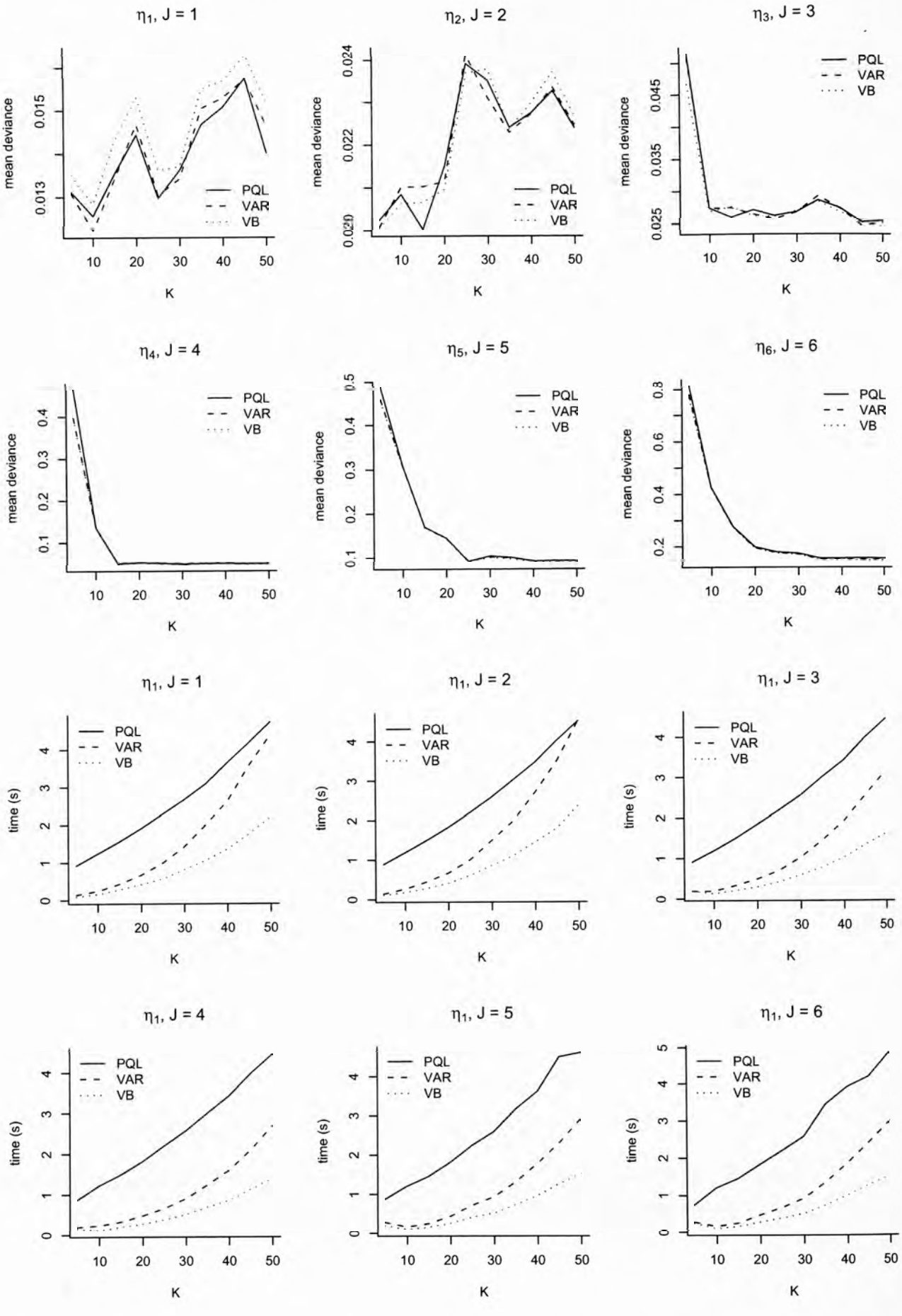


Figure 5.13: *Mean deviances and computational time for the* ***number of knots and complexity*** *simulations for for the Poisson case (see text for details).*

For one particular real dataset we have shown that grid-based variational posterior approximations to Bayesian logistic LMMs did not perform particularly well compared to kernel density estimates based on posterior samples obtained from a MCMC based approach. However, this approach did obtain marginal posterior approximations which were better than those obtained by VPA. Nevertheless, we believe that grid-based variational posterior approximations may still be viable approach for other types of GLMM.

Again, although based only on limited empirical studies, for Poisson random intercept models the variational approximations we considered here did seem to offer a potential improvement over PQL approximations. On the other hand, the variational approximations we considered for logistic random intercept models, did not seem to offer an improvement over PQL, although they appear to be still viable alternatives when the number of observations per group is sufficiently large (Rijmen & Vomlel, 2007). We speculate that the diminished performance for the logistic case is due to the additional use of the ξ or log transforms used to obtain lower bounds on the likelihood.

Finally, for the scatterplot smoothing the variational approximations we have considered have not shown to be clearly better than PQL neither have they been shown to be clearly worse in terms of the mean deviance measure of error. A potential advantage of this approach over PQL is that it is possible to combine it with additional model components to deal with various model complexities.

Variational approximations to GLMMs are yet to be fully explored. There are many alternative lines of research that may be pursued including:

1. In this chapter we used the multivariate Gaussian distribution to approximate the posterior density of the spline coefficients. Similar approximations might be possible using generalisations of the multivariate Gaussian distribution.
2. The logistic LMM case is a sticking point in terms of accuracy because of the need to use ξ and log transforms to obtain lower bounds for the marginal likelihood. A method of combining ξ and log transforms could be developed and still other approximations could be possible.
3. The optimisation problem posed by the "maximisation" step is quite unusual because it can be interpreted as involving a nonlinear matrix equality constraint for the covariance matrix $\mathbf{\Sigma}$. Alternative optimisation procedures could possibly greatly improve numerical stability and running times.
4. In this chapter lower bounds were primarily examined. Upper bounds could also be developed. These could be used in combination in a number of interesting ways including conservative and liberal hypothesis testing via likelihood ratio test and model selection via under/overestimation of the Akaike information criterion.
5. Theoretical properties such as bias and asymptotic consistency are yet to be addressed.

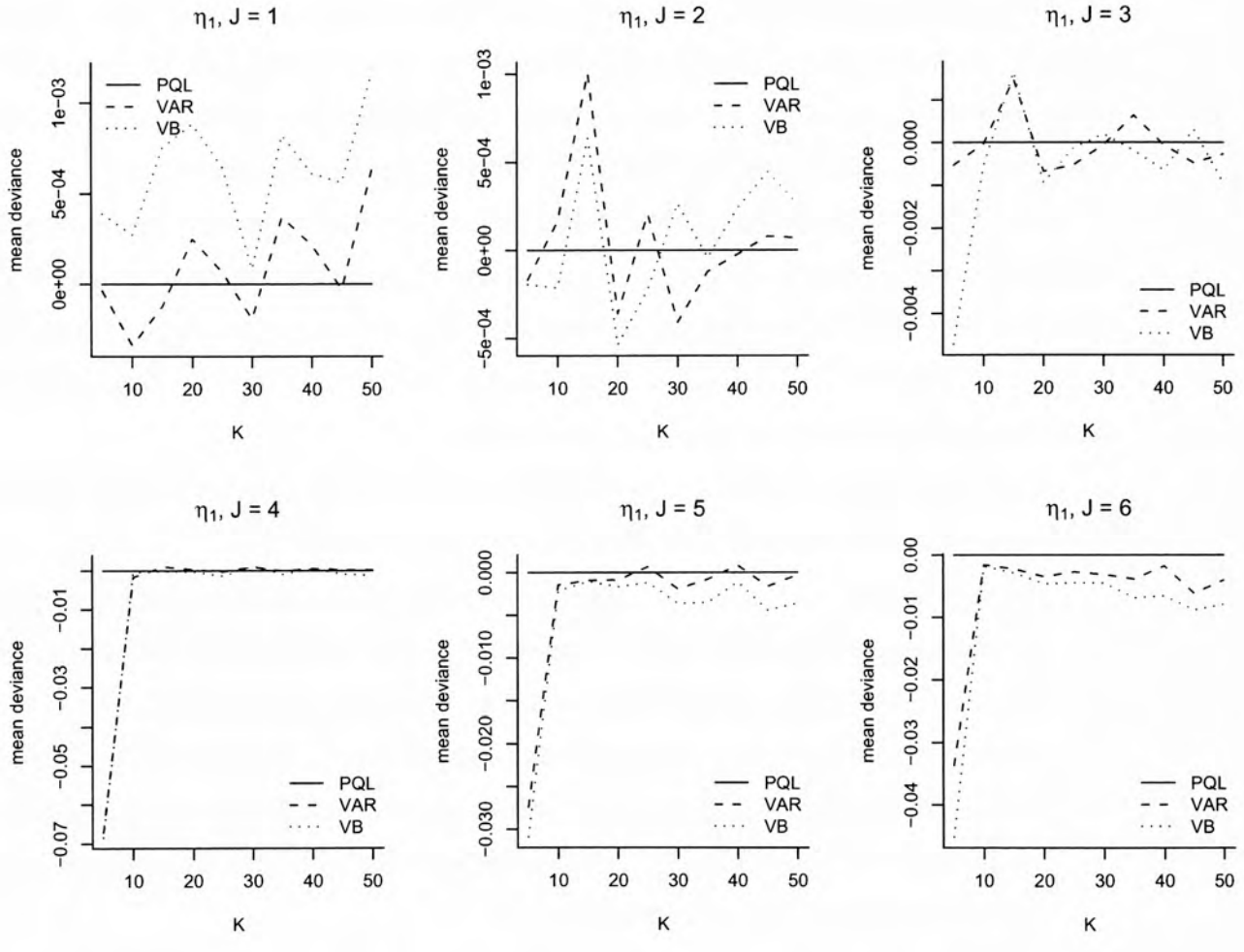


Figure 5.14: *Differences in mean deviances with mean deviances for PQL for the varying* ***number of knots and complexity*** *simulations for the Poisson case (see text for details).*

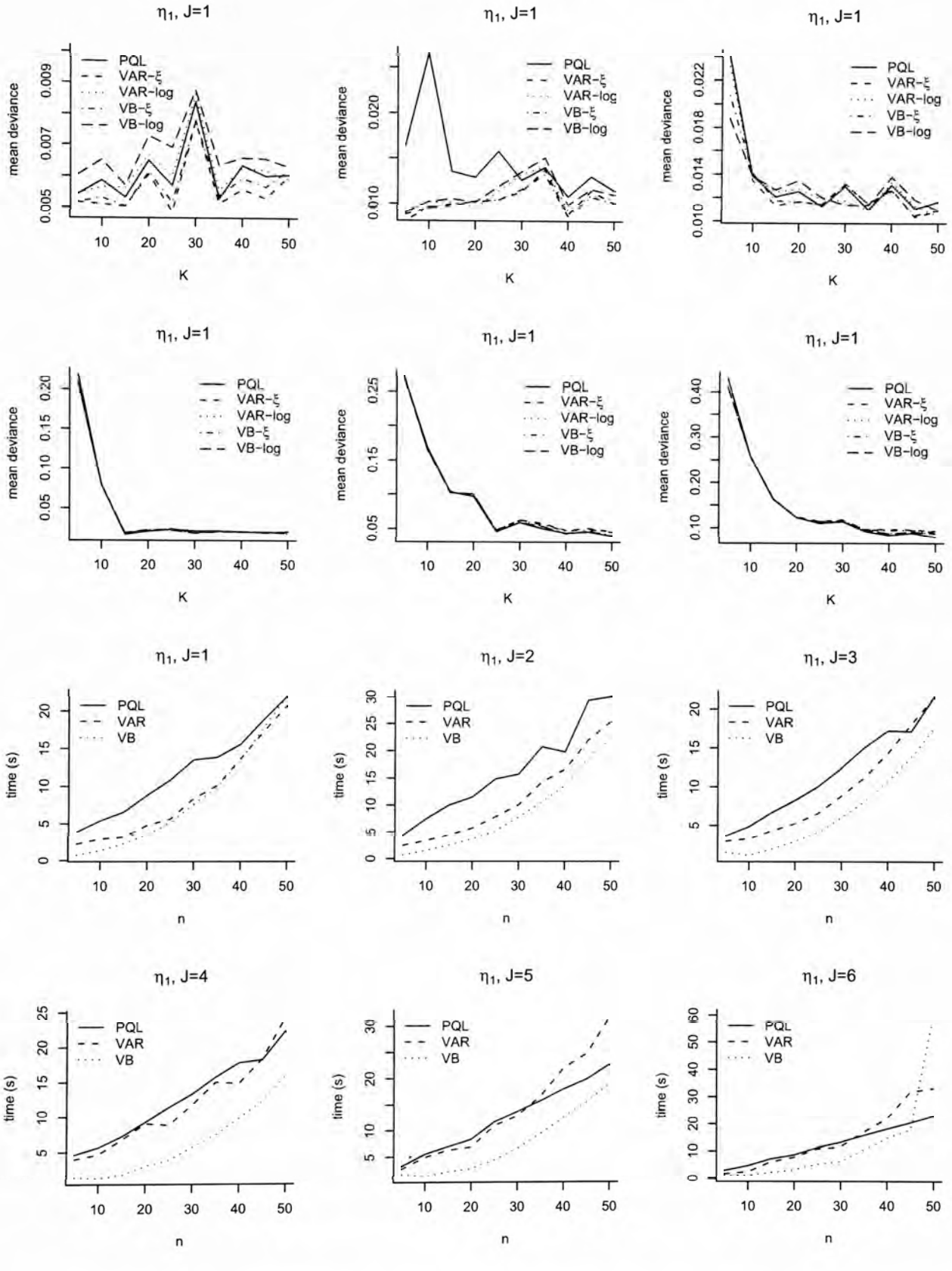


Figure 5.15: *Mean deviances and computational time for the* ***number of knots and complexity*** *simulations for for the logistic case (see text for details).*

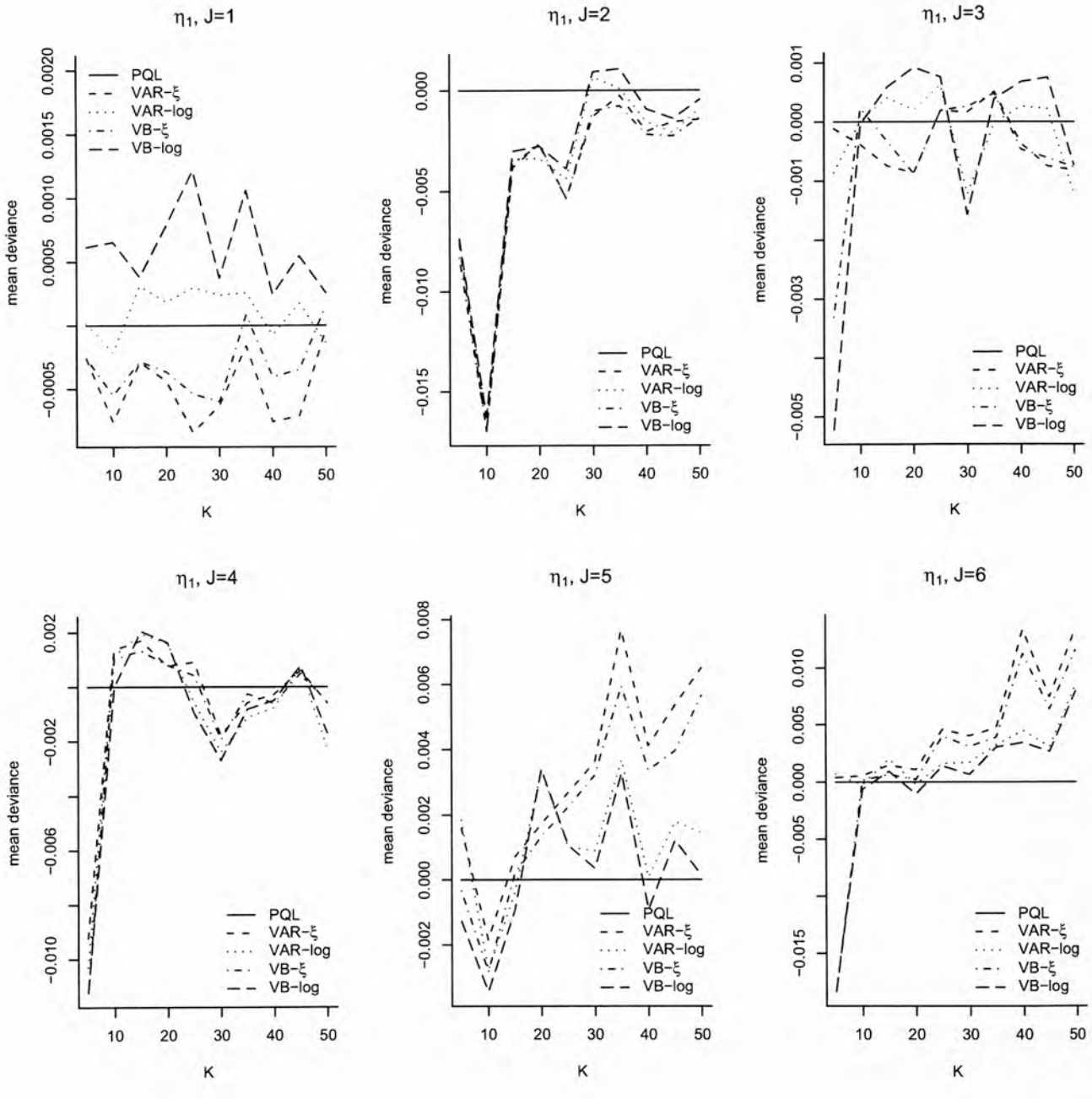


Figure 5.16: *Differences in mean deviances with mean deviances for PQL for the varying* ***number of knots and complexity*** *simulations for the logistic case (see text for details).*

CHAPTER 6

Robust Spatially Adaptive Penalised Splines with Heteroscedastic Errors

6.1 Introduction

The topic of robustness in statistics has been the subject of an enormous amount of research over the past few decades (e.g, Huber, 1981; Hampel, Ronchetti, Rousseeuw & Stahel, 1986; Rousseeuw & Lerow, 1987; Staudte & Sheather, 1990; Wilcox, 1997). When we design a model or procedure for a particular dataset we use a number of working assumptions. In semiparametric regression we might, for example, assume a type of noise and constant variance. When these assumptions are violated the model may not fit the data well. We say a model is robust if the procedure used to fit the model does not perform much worse when the underlying assumptions of the model are violated (Garthwaite, Jolliffe & Jones, 2002).

In practice robust statistics is used to deal with complications arising in the data to be analysed. As discussed in Chapter 1 in data mining applications the behavior of the underlying system might change abruptly and be highly oscillatory and jump, cusp and other change points may occur. The noise in the system under observation may be asymmetric and heteroscedastic. Outliers can occur for a number of reasons including questionable experimental (or methodological) design, measurement error and human error. Furthermore, when designing a model we may make a poor choice on the type of noise which corrupts the data. Due to the high dimensional nature of many real world problems these complications may not be identified from a casual examination of the data. For example, while outliers are by their nature can be easy to spot with the eye in one dimensional scatterplots, outliers in higher dimensional data may not be obvious even from inspection of three dimensional scatterplots using two predictors. Not dealing with these problems can drastically alter the quality of predictions in many applications and so care must be taken to avoid false conclusions about the data.

Huber (1981) in his seminal book on robust statistics developed a class of estimators called M-estimators, which are generalisations of maximum likelihood estimators, via modification to the measure of error or loss used. The Student's t-distribution is an example which deviates from typical normality assumptions in such a way that outliers have reduced influence on the fitted function. The thickness of the tails in the Student's t-distribution can be controlled by the degrees of freedom parameter. In terms of M-estimators this amounts to smaller loss for unusually large deviations from the mean. In

comparison, modelling for Gaussian distributed noise leads to quadratic loss. Thus Student's t-distributed noise will penalise observations large distances from the mean less than the Gaussian distribution would.

In Section 2.4 we considered smoothing via a linear mixed model (LMM) formulation where the noise was assumed to be Gaussian with constant variance. In many real situations this assumption, called *homoscedasticity*, is unrealistic and may lead to false conclusions. Adverse effects of holding this assumption include incorrect confidence intervals, incorrect inferences on particular parameter values and calibration inference (predicting an x based on a y). The converse situation where the variance may change is called *heteroscedasticity* and is examined in, amongst others, Davidian & Carroll (1987), Carroll & Ruppert (1988), Ruppert, Wand & Carroll (2003) and Crainiceanu, Ruppert & Carroll (2007). A model that allows for heteroscedasticity may lead to more robust results we can exploit the heteroscedasticity to obtain better fits in regions where there is less noise corrupting the response.

Finally, spatially adaptive smoothing for spatially inhomogeneous functions may be seen as a form of robustness. A vast number of papers and indeed books have been written on spatially adaptive smoothing. These range from regression spline methods based using both local (Friedman, 1991; Lindstrom, 1999; Zhou & Shen, 2001; Miyata & Shen, 2003; Mao & Zhao, 2003) and global optimisation approaches (Jupp, 1978; Pittman, 2002; Beliakov, 2004), penalised splines (Wahba, 1990; Green & Silverman, 1994; Eilers & Marx, 1996; Eubank 1999; Ruppert & Carroll, 2000; Gu, 2002; Ruppert, 2002; Ruppert *et al.* 2003; Wand & Ormerod, 2008), kernel smoothing (Wand & Jones, 1995; Fan & Gibels, 1996; Loader, 1999), wavelets (Donoho & Johnstone, 1994,1995; Donoho, Johnstone, Kerkyacharian & Picard, 1995), Bayesian (Denison, Mallick & Smith, 1998; DiMatteo, Genovese & Kass, 2001; Denison, Holmes, Mallick & Smith, 2002) and hybrid approaches (Luo & Wahba, 1997). Each of these methods typically work reasonably well in practice but require varying degrees of complexity to implement. In most of the above methods, the quality of the fit typically depends heavily on the time the user is willing to wait for a result. In this chapter we examine the model proposed by Baladandayuthapani, Mallick & Carroll (2005) which, as we will show, bears some similarity to the variance function models of Davidian & Carroll (1987), Carroll & Ruppert (1988), Ruppert *et al.* (2003) and Crainiceanu *et al.* (2007).

In the context of linear mixed model (LMM) smoothing, modifications of the model away from normality typically result in analytically intractable integrals when calculating the likelihood (for example, Staudenmayer, Lake & Wand, 2008). Dealing with these analytically intractable integrals has lead to a great deal of research which has been dominated by Monte Carlo type approaches. These numerical approximations while very accurate are typically slower than analytic approximations such as the Laplace approximation. Unfortunately, in applications where speed is important the Laplace approximation has its limitations. For example, in the context of the models we will consider

the Laplace approximation may not work well because the integrand is not Gaussian in shape. Thus we seek alternative computationally efficient approximations.

Recent developments in a new class of analytic approximations, variational approximations, have lead to computationally efficient estimators for models involving analytically intractable integrals. Several variational approximations have recently been applied to a variety of robustness models (Bishop & Tipping, 2000; Tipping 2001; Faul & Tipping 2001; Tipping & Lawrence, 2003; Kuss, 2006, Chapter 5). We use a similar approach to Bishop & Tipping (2000), Tipping & Lawrence (2003) and Kuss (2006) which exploits the fact that the Student's t-distribution may be written as a Gaussian scale mixture (Andrews & Mallows, 1974).

The methods developed in this chapter show the potential power of variational approximations to fit complex models accurately, efficiently and relatively easily. In this chapter we:

1. Develop variational approximations to Student's t mixed models. This uses a similar approach to the Student's t ideas of Bishop & Tipping (2000), Tipping & Lawrence (2003) and Kuss (2006) except we use a slightly different parameterisation optimise over the degrees of freedom and variance parameters explicitly. We develop a variational approximation for a Student's t mixed model and show heuristically why this approximation provides robustness to outliers.
2. Develop variational approximations for linear mixed models with heteroscedastic noise (robustness to heteroscedasticity).
3. Develop variational approximations for linear mixed models with adaptive variance components (robustness to spatial variation).
4. In a seamless manner we combine any combination of mixed models with Gaussian or Student's t noise, variance function estimation and spatially adaptive variance components.
5. Develop optimisation routines which fits these models in minutes, if not seconds, on a typical 2008 computer.

6.2 Student's t Mixed Models

Let

$$(y_i, \mathbf{x}_i), 1 \leq i \leq n \tag{6.1}$$

be a set of n paired observations with $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^d$ where the response variable $\mathbf{y} = (y_1, \ldots, y_n)$ is a continuous variable. Consider modelling the y_is using the univariate Student's t-distribution with mean $\mu_i = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i$, constant variance σ_y^2 and degrees of freedom parameter ν_y where $\mathbf{X}$ and $\mathbf{Z}$ are $n \times p$ and $n \times q$ matrices respectively with associated basis functions $\{X_i(\cdot)\}_{i=1}^p$ and $\{Z_i(\cdot)\}_{i=1}^q$ as described in Chapter 1 & 2, $\boldsymbol{\beta}$ is a vector of fixed effects and $\mathbf{u}$ is a vector of random effects. Thus we have

$$\begin{aligned} y_i|\mathbf{u}; \boldsymbol{\beta}, \sigma_y^2, \nu_y &\overset{ind.}{\sim} t((\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i, \sigma_y^2, \nu_y) \\ \mathbf{u}; \boldsymbol{\sigma}^2 &\sim N(\mathbf{0}, \mathbf{G}_{\boldsymbol{\sigma}^2}) \end{aligned}$$

where $\mathbf{G}_{\boldsymbol{\sigma}^2}$ is the covariance matrix for $\mathbf{u}$ parameterized by $\boldsymbol{\sigma}^2$ as described in Chapter 1, the density for the univariate Student's t-distribution is given by

$$S(y_i|\mu_i, \sigma_y^2, \nu_y) = \frac{\Gamma\left(\frac{1+\nu_y}{2}\right)}{\Gamma\left(\frac{\nu_y}{2}\right)\left(\pi\nu_y\sigma_y^2\right)^{\frac{1}{2}}}\left(1 + \frac{(y_i - \mu_i)^2}{\nu_y\sigma_y^2}\right)^{-\frac{1+\nu_y}{2}}$$

and $\Gamma(\cdot)$ denotes the gamma function (see Abramowitz & Stegun, 1964, Chapter 6). We note that as $\nu_y \to \infty$ the Student's t-distribution approaches that of the univariate Gaussian distribution.

The log-likelihood for the Student's t mixed model (STMM) model may be written as

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_y^2, \nu_y) &= \log \int [\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \sigma_y^2, \nu_y][\mathbf{u}; \boldsymbol{\sigma}^2] d\mathbf{u} \\ &= \log \int \frac{\Gamma\left(\frac{1+\nu_y}{2}\right)^n}{\Gamma\left(\frac{\nu_y}{2}\right)^n \left(\pi\nu_y\sigma_y^2\right)^{\frac{n}{2}}} \prod_{i=1}^{n}\left(1 + \frac{(y_i - (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})_i)^2}{\nu_y\sigma_y^2}\right)^{-\frac{1+\nu_y}{2}} \\ &\quad \times \frac{|\mathbf{D}_{\boldsymbol{\sigma}^2}|^{\frac{1}{2}}}{(2\pi)^{\frac{q}{2}}} e^{-\frac{\mathbf{u}^T \mathbf{D}_{\boldsymbol{\sigma}^2} \mathbf{u}}{2}} d\mathbf{u}\end{aligned}$$

which involves an integral with no (known) closed form.

Bishop & Tipping (2000) and Tipping & Lawrence (2003) advocated a particularly elegant variational approximation to the integral. We adopt different approach to the variational approximation proposed by Tipping & Lawrence (2003) but use a parameterisation explicitly in terms of σ_y^2 and ν_y. Noting from Chapter 4 that for any density $\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi})$ we have

$$\ell(\boldsymbol{\theta}) = \log \int [\mathbf{y}, \boldsymbol{\vartheta}; \boldsymbol{\theta}] d\boldsymbol{\vartheta} \geq \ell_L(\boldsymbol{\theta}; \boldsymbol{\xi}) = \mathbb{E}_\delta \log[\mathbf{y}, \boldsymbol{\vartheta}; \boldsymbol{\theta}] + \mathcal{H}_\delta \tag{6.2}$$

where $\boldsymbol{\vartheta}$ are variables we want to integrate out, $\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi})$ is a density, parameterized by $\boldsymbol{\xi}$, which approximates the posterior distribution $\boldsymbol{\vartheta}|\mathbf{y}$, $\mathbb{E}_\delta$ denotes expectation with respect to δ, $\mathcal{H}_\delta = -\mathbb{E}_\delta \log(\delta(\boldsymbol{\vartheta}; \boldsymbol{\xi}))$ is the entropy of δ and the subscript L denotes a lower bound. Using similar terminology to that of Jaakkola & Jordan (2000), who might call (6.2) a δ-transform, we call a density transform.

To apply this approximation first note that the Student's t-distribution can be derived as a Gaussian scale mixture of gamma random variables (for example, Andrews & Mallows, 1974; Liu & Rubin 1995), i.e we can write

$$S(y_i|\mu_i, \sigma_y^2, \nu_y) = \int \phi_{\gamma^{-1}\sigma_y^2}(y_i - \mu_i) g(\gamma; \nu_y/2, \nu_y/2) d\gamma$$

where

$$\begin{aligned}\phi_{\gamma^{-1}\sigma_y^2}(y_i - \mu_i) &= \frac{\gamma^{\frac{1}{2}}}{(2\pi\sigma_y^2)^{\frac{1}{2}}} \exp\left\{-\frac{\gamma}{2\sigma_y^2}(y_i - \mu_i)^2\right\} \text{ and} \\ g(\gamma; \nu_y/2, \nu_y/2) &= \frac{\left(\frac{\nu_y}{2}\right)^{\frac{\nu_y}{2}}}{\Gamma\left(\frac{\nu_y}{2}\right)} \gamma^{\frac{\nu_y}{2}-1} \exp\left(-\frac{\nu_y\gamma}{2}\right).\end{aligned}$$

Thus we can also write the likelihood as

$$\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_y^2, \nu_y) = \log \int [\mathbf{u}; \boldsymbol{\sigma}^2] \left\{ \prod_{i=1}^{n} \int_0^{\infty} [y_i|\mathbf{u}; \boldsymbol{\beta}, \gamma_{y,i}^{-1}\sigma_y^2][\gamma_{y,i}; \nu] d\gamma_{y,i} \right\} d\mathbf{u}.$$

It is easy to integrate out either $\mathbf{u}$ or $\boldsymbol{\gamma}_y = (\gamma_{y,1}, \ldots, \gamma_{y,n})$ but difficult to integrate out both.

Tipping & Lawrence (2003) chose the density transform mirroring the priors on $\mathbf{u}$ and $\boldsymbol{\gamma}_y$, i.e. $\delta(\mathbf{u}, \boldsymbol{\gamma}_y) = \delta_{\mathbf{u}}(\mathbf{u}) \prod_{i=1}^{n} \delta_{\gamma_{y,i}}(\gamma_{y,i})$ where

$$\begin{aligned} \mathbf{u}|\mathbf{y} &\sim_{\delta_{\mathbf{u}}} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \gamma_{y,i}|\mathbf{y} &\sim_{\delta_{\gamma_{y,i}}} \text{Gamma}(A_{y,i}, B_{y,i}) \end{aligned}$$

with $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, A_{y,1}, \ldots, A_{y,n}, B_{y,1}, \ldots, B_{y,n})$. Here $\mathbf{u}|\mathbf{y} \sim_{\delta_{\mathbf{u}}} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes approximating $\mathbf{u}|\mathbf{y}$ by the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and similarly for $\gamma_{y,i}$. Using this density a lower bound for ℓ is given by

$$\begin{aligned} \ell_L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_y^2, \nu_y; \boldsymbol{\xi}) &= \mathbb{E}_\delta \log[\mathbf{u}; \boldsymbol{\sigma}^2] + \mathcal{H}_{\delta_{\mathbf{u}}} \\ &\quad + \sum_{i=1}^{n} \mathbb{E}_\delta \log[y_i|\mathbf{u}; \boldsymbol{\beta}, \gamma_{y,i}^{-1}\sigma_y^2] + \mathbb{E}_\delta \log[\gamma_{y,i}; \nu_y] + \mathcal{H}_{\gamma_{y,i}} \end{aligned}$$

where, ignoring additive constants, $1 \le i \le n$

$$\begin{aligned} \mathbb{E}_\delta \log[\mathbf{u}; \boldsymbol{\sigma}^2] &= \tfrac{1}{2} \log |\mathbf{D}_{\boldsymbol{\sigma}^2}| - \frac{\boldsymbol{\mu}^T \mathbf{D}_{\boldsymbol{\sigma}^2} \boldsymbol{\mu} + \text{tr}(\boldsymbol{\Sigma} \mathbf{D}_{\boldsymbol{\sigma}^2})}{2}, \\ \mathcal{H}_{\delta_{\mathbf{u}}} &= \tfrac{1}{2} \log |\boldsymbol{\Sigma}|, \\ \mathbb{E}_\delta \log[y_i|\mathbf{u}; \boldsymbol{\beta}, \gamma_{y,i}^{-1}\sigma_y^2] &= \frac{\psi(A_{y,i}) - \log(B_{y,i}) - \log(\sigma_y^2)}{2} \\ &\quad - \frac{A_{y,i}}{B_{y,i}} \cdot \frac{(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}{2\sigma_y^2}, \\ \mathbb{E}_\delta \log[\gamma_{y,i}; \nu_y] &= \frac{\nu_y}{2} \log\left(\frac{\nu_y}{2}\right) - \log \Gamma\left(\frac{\nu_y}{2}\right) \\ &\quad + \left(\frac{\nu}{2} - 1\right)(\psi(A_{y,i}) - \log(B_{y,i})) - \frac{\nu_y A_{y,i}}{2B_{y,i}} \\ \text{and } \mathcal{H}_{\gamma_{y,i}} &= A_{y,i} - \log(B_{y,i}) + \log \Gamma(A_{y,i}) + (1 - A_{y,i})\psi(A_{y,i}) \end{aligned}$$

with $\mathbf{C} \equiv [\mathbf{X}, \mathbf{Z}]$ and $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}, \boldsymbol{\mu})$. Here we have used the facts that $\mathbb{E}_\delta(\gamma_{y,i}) = A_{y,i}/B_{y,i}$, $\mathbb{E}_\delta(\log \gamma_{y,i}) = \psi(A_{y,i}) - \log(B_{y,i})$, $\mathbb{E}_\delta(\mathbf{x}^T \mathbf{A} \mathbf{x}) = \mathbb{E}_\delta(\mathbf{x})^T \mathbf{A} \mathbb{E}_\delta(\mathbf{x}) + \text{tr}(\mathbf{A}\text{Cov}_\delta(\mathbf{x}))$ for any random vector $\mathbf{x}$ and appropriately-sized matrix $\mathbf{A}$ and $\psi(\cdot)$ is the digamma function (see Abramowitz & Stegun, 1964, Chapter 6). These may be verified by direct integration, integration by parts or by using a symbolic computing package, for example `Maple` or `Mathematica`.

The first derivatives of ℓ_L with respect to $\boldsymbol{\nu}$ and $A_{y,i}, B_{y,i}$ and $1 \leq i \leq n$ are

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{\nu}}\ell_L &= \mathbf{C}^T\text{diag}(\mathbf{w})(\mathbf{y} - \mathbf{C}\boldsymbol{\nu}) - \mathbf{B}_{\boldsymbol{\sigma}^2}\boldsymbol{\nu} \\
\mathsf{D}_{A_{y,i}}\ell_L &= -\left(\frac{\nu_y}{2} + \frac{(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}{2\sigma_y^2}\right) \cdot \frac{1}{B_{y,i}} + \left(\frac{1+\nu_y}{2} - A_{y,i}\right)\psi'(A_{y,i}) + 1 \\
\mathsf{D}_{B_{y,i}}\ell_L &= \left(\frac{\nu_y}{2} + \frac{(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}{2\sigma_y^2}\right) \cdot \frac{A_{y,i}}{B_{y,i}^2} - \frac{1+\nu_y}{2} \cdot \frac{1}{B_{y,i}}
\end{aligned} \tag{6.3}$$

where $w_i = A_{y,i}/B_{y,i}\sigma_y^2$ and $\mathbf{B}_{\boldsymbol{\sigma}^2} \equiv \text{blockdiag}(\mathbf{0}_p, \mathbf{D}_{\boldsymbol{\sigma}^2})$.

Thus, the first order optimality conditions imply that

$$\begin{aligned}
\boldsymbol{\nu} &:= \left(\mathbf{C}^T\text{diag}(\mathbf{w})\mathbf{C} + \mathbf{B}_{\sigma^2}\right)^{-1}\mathbf{C}^T\text{diag}(\mathbf{w})\mathbf{y}, \\
A_{y,i} &:= \frac{\nu_y + 1}{2} \\
\text{and } B_{y,i} &:= \frac{\nu_y}{2} + \frac{(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}{2\sigma_y^2}.
\end{aligned} \tag{6.4}$$

Note that the solution for $\boldsymbol{\nu}$ is a weighted least squares solution with weight vector $\mathbf{w} = (w_1, \ldots, w_n)$ and

$$w_i = \frac{1 + \nu_y}{\nu_y\sigma_y^2 + (\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii}}. \tag{6.5}$$

From equation (6.5) we can see how the value of ν_y and the size of the residuals $(\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i^2$ affect the fit. As $\nu_y \to \infty$ we have $w_i \to \sigma_y^{-2}$, which is the w_i value for linear mixed models (see Section 5.2 for example). For small ν_y the size of w_i decreases as the size of the residuals decrease. Thus, for small ν_y points with larger residuals there is a reduced effect on the fit, so we should expect that maximisation of ℓ_L with respect to $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \sigma_y^2, \nu_y)$ and $\boldsymbol{\xi}$ should produce fits which have some resistance to outliers.

In the following sections we will combine this model for STMMs with variance function and adaptive variance component ideas. Instead of describing how to optimise ℓ_L with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ now we defer discussion of this to Section 6.5.1.

Suppose that $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\sigma}_y^2, \widehat{\nu}_y)$ and $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}, \widehat{A}_{y,1}, \ldots, \widehat{A}_{y,n}, \widehat{B}_{y,1}, \ldots, \widehat{B}_{y,n})$ are the values of the likelihood parameters and variational parameters which maximise ℓ_L. As usual predictions for the mean function use

$$\widehat{f}(\mathbf{x}) = \sum_{i=1}^{p} X_i(\mathbf{x})\widehat{\beta}_i + \sum_{j=1}^{q} Z_i(\mathbf{x})\widehat{\mu}_i. \tag{6.6}$$

The fitted y_is are given by $\widehat{\mathbf{f}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\boldsymbol{\mu}} = (\widehat{f}_1, \ldots, \widehat{f}_n)$ where $\widehat{f}_i = \widehat{f}(\mathbf{x}_i)$.

6.2.1 *Numerical Experience*

To test the effectiveness of the above variational approximation for fitting STMMs we will use the following functions (and corresponding number of data points n)

$$\begin{aligned} f_1(x) &= 4x - 2 && (n = 50) \\ f_2(x) &= \sin(8(x - 0.5)) + 2\exp(-16^2(x - 0.5)^2) && (n = 400) \\ f_3(x) &= 1/(0.1 + x) + 8\exp(-400(x - 0.5)^2) && (n = 800) \\ f_4(x) &= 3\phi\left((3x - 7)/3\right)/2 - \phi\left(25x - 20\right) && (n = 200) \end{aligned} \tag{6.7}$$

where the x values will be equally spaced between 0 and 1. These represent a linear function and a variety of nonlinear functions for the true mean. We will also consider a variety of noise types

1. Gaussian noise, $y_i \sim N(f_j(x), \sigma_y^2)$
2. Student's t noise, $y_i \sim t(f_j(x), \sigma_y^2, \nu_y)$
3. Gaussian mixture (GM) noise, $y_i \sim (1 - W_i)N(f_j(x), \sigma_{y,1}^2) + W_i N(f_j(x), \sigma_{y,2}^2)$ with $W_i \sim \text{Bern}(\omega)$.

where σ_y^2, ν_y, $\sigma_{y,1}^2$, $\sigma_{y,2}^2$ and ω are fixed constants. Here we will use the noise settings

1. Gaussian noise ($\sigma_y^2 = 0.25$),
2. Student's t noise with 1 degree of freedom (Cauchy noise, $\sigma_y^2 = 0.25, \nu_y = 1$),
3. Student's t noise with 3 degrees of freedom ($\sigma_y^2 = 0.25, \nu_y = 3$),
4. Student's t noise with 5 degrees of freedom ($\sigma_y^2 = 0.25, \nu_y = 5$) and
5. Gaussian mixture noise ($\omega = 0.05, \sigma_{y,1}^2 = 0.25, \sigma_{y,2}^2 = 1$).

We will use thin plate splines (see Section 1.2.4) for these experiments with $m = 3$, $K_1 = 25$ knots for the construction of the $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{D}_i$ matrices. Note that we standardised the xs to have zero mean and unit variance which typically improves numerical stability. These knots are spaced using the quantities of the unique x_is, i.e. κ_i satisfies

$$\kappa_k = \left(\frac{k}{K+1}\right) \text{th sample quantile of the unique } x_i\text{s}, 1 \leq k \leq K_1 \tag{6.8}$$

and the we will measure the error of each fit by the sample mean square error

$$\text{MSE}(f_j, \widehat{f}) = n^{-1}\sum_{i=1}^{n}(f_j(x_i) - \widehat{f}(x_i))^2. \tag{6.9}$$

Each of these settings were fit using LMM smoothing (see for example Section 2.4) and the STMM approximation for 100 trials. The median MSE, standard error in brackets and estimates for σ_y^2 and ν_y for each setting is summarised in Table 6.2.1. From this we see that the MSEs and the STMM estimates for σ_y^2 is better for just about every case. For the Gaussian noise cases the STMM approximation estimates for $\widehat{\nu}_y$ are large, noting that for ν_y larger than 40 the univariate Student's t-distribution is extremely close to the univariate Gaussian Student's t-distribution. Also in particular MSEs for the STMM

approximation are not obviously worse for the Gaussian noise cases. Also the STMM approximation estimates for ν_y are reasonable for the cases with Student's t noise.

Finally, Figure 6.1 illustrates come exemplar plots and absolute residuals for each of the mean functions used with student noise with variance $\sigma_y^2 = 0.25$ and degrees of freedom $\nu_y = 3$. Note each of the examples has a fair proportion of outliers and that STMM gives slightly better fits in each case.

6.3 Variance Function Estimation

Suppose that we model the relationship between the ys and the xs in (6.1) using the Gaussian distribution

$$y_i|\mathbf{u};\boldsymbol{\beta},\sigma_y^2,\nu_y \overset{ind.}{\sim} N(f(\mathbf{x}_i),\sigma_y^2)$$

where $f(\cdot)$ is the mean function. It is often implicitly assumed that

$$\text{Var}(\mathbf{y}|\mathbf{x}) = \sigma_y^2$$

where σ_y^2 is a constant parameter to be estimated, i.e. homoscedasticity. Instead we will consider a variance function model

$$\begin{aligned} y_i|\mathbf{x}_i &\overset{ind.}{\sim} N(f(\mathbf{x}_i),\sigma_y^2(\mathbf{x}_i)) \\ \log\left(\sigma_y^2(\mathbf{x}_i)\right) &= g(\mathbf{x}_i) \end{aligned}$$

where $g(\cdot)$ is the function associated with the variance of $y_i|\mathbf{x}_i$. This model has been considered in, amongst others, Ruppert *et al.* (2003, Chapter 14).

We model $f(\cdot)$ and $g(\cdot)$ using the spline methodology developed in Chapters 1 & 2 so that

$$\begin{aligned} f(x_i) &= (\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})_i, & \mathbf{u} &\sim N(\mathbf{0},\mathbf{G}_{\boldsymbol{\sigma}^2}) \\ g(x_i) &= (\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}+\widetilde{\mathbf{Z}}\widetilde{\mathbf{u}})_i \quad \text{and} & \widetilde{\mathbf{u}} &\sim N(\mathbf{0},\widetilde{\mathbf{G}}_{\widetilde{\boldsymbol{\sigma}}^2}). \end{aligned}$$

Here $\mathbf{X}$ and $\mathbf{Z}$ are $n\times p$ and $n\times q$ basis matrices for the mean function with associated basis functions $\{X_i(\cdot)\}_{i=1}^p$ and $\{Z_i(\cdot)\}_{i=1}^q$ respectively, $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Z}}$ are $n\times\widetilde{p}$ and $n\times\widetilde{q}$ basis matrices for the variance function with associated basis functions $\{\widetilde{X}_i(\cdot)\}_{i=1}^{\widetilde{p}}$ and $\{\widetilde{Z}_i(\cdot)\}_{i=1}^{\widetilde{q}}$ respectively and the associated penalty matrices are $\mathbf{G}_{\boldsymbol{\sigma}^2}^{-1}=\mathbf{D}_{\boldsymbol{\sigma}^2}$ and $\widetilde{\mathbf{G}}_{\widetilde{\boldsymbol{\sigma}}^2}^{-1}=\widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2}$.

Note that we allow for the possibility that $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Z}}$ can be possibly different from $\mathbf{X}$ and $\mathbf{Z}$. The motivation for this is situations where the mean may depend on one set of variables while the noise many depend on another set of variables. For compactness we will write $\sigma_y^2(x_i)=\sigma_{y,i}^2$ and $\boldsymbol{\sigma}_y^2=(\sigma_{y,1}^2,\ldots,\sigma_{y,n}^2)$.

f_j	Noise Type	Noise Setting	Median MSE LMM	Median MSE STMM	$\widehat{\sigma}_y^2$ LMM	$\widehat{\sigma}_y^2$ STMM	$\widehat{\nu}_y$ STMM
$j=1$	N	$1.(\nu_y \to \infty)$	0.0096 (0.0008)	0.0095 (0.0009)	0.228 (0.0036)	0.208 (0.0041)	125.25 (13.06)
	S	$2.(\nu_y = 1)$	0.6659 (0.5388)	0.0202 (0.0017)	11.519 (6.5762)	0.217 (0.0098)	0.97 (0.02)
	S	$3.(\nu_y = 3)$	0.0310 (0.0027)	0.0186 (0.0017)	0.582 (0.0261)	0.220 (0.0086)	2.88 (0.60)
	S	$4.(\nu_y = 5)$	0.0165 (0.0013)	0.0157 (0.0011)	0.366 (0.0087)	0.228 (0.0068)	4.79 (4.02)
	GM	5.	0.0120 (0.0009)	0.0113 (0.0008)	0.265 (0.0052)	0.235 (0.0056)	44.84 (11.71)
$j=2$	N	$1.(\nu_y \to \infty)$	0.0112 (0.0003)	0.0113 (0.0003)	0.250 (0.0014)	0.243 (0.0014)	138.05 (5.86)
	S	$2.(\nu_y = 1)$	1.3315 (0.6833)	0.0335 (0.0022)	136.515 (71.5807)	0.285 (0.0044)	1.05 (0.01)
	S	$3.(\nu_y = 3)$	0.0347 (0.0015)	0.0199 (0.0005)	0.649 (0.0114)	0.270 (0.0033)	3.35 (0.05)
	S	$4.(\nu_y = 5)$	0.0204 (0.0006)	0.0170 (0.0005)	0.410 (0.0039)	0.262 (0.0028)	5.49 (0.15)
	GM	5.	0.0132 (0.0004)	0.0130 (0.0004)	0.283 (0.0021)	0.253 (0.0021)	17.44 (3.23)
$j=3$	N	$1.(\nu_y \to \infty)$	0.0083 (0.0002)	0.0082 (0.0002)	0.252 (0.0011)	0.245 (0.0012)	109.54 (4.59)
	S	$2.(\nu_y = 1)$	4.3262 (0.5962)	0.0159 (0.0004)	255.309 (107.9578)	0.265 (0.0027)	1.03 (0.01)
	S	$3.(\nu_y = 3)$	0.0189 (0.0006)	0.0106 (0.0002)	0.661 (0.0128)	0.267 (0.0019)	3.20 (0.03)
	S	$4.(\nu_y = 5)$	0.0120 (0.0003)	0.0105 (0.0003)	0.406 (0.0030)	0.261 (0.0020)	5.65 (0.11)
	GM	5.	0.0086 (0.0002)	0.0083 (0.0002)	0.289 (0.0015)	0.254 (0.0016)	18.87 (1.52)
$j=4$	N	$1.(\nu_y \to \infty)$	0.0246 (0.0005)	0.0247 (0.0005)	0.264 (0.0020)	0.255 (0.0023)	114.62 (6.9738)
	S	$2.(\nu_y = 1)$	0.9885 (0.3775)	0.0372 (0.0009)	46.637 (31.3852)	0.277 (0.0048)	1.03 (0.0107)
	S	$3.(\nu_y = 3)$	0.0375 (0.0010)	0.0297 (0.0006)	0.662 (0.0121)	0.274 (0.0040)	3.20 (0.0810)
	S	$4.(\nu_y = 5)$	0.0291 (0.0006)	0.0271 (0.0005)	0.405 (0.0054)	0.279 (0.0047)	6.29 (0.3730)
	GM	5.	0.0261 (0.0004)	0.0257 (0.0004)	0.297 (0.0031)	0.269 (0.0030)	28.64 (7.0489)

Table 6.2.1: *Mean square errors (MSE), standard errors (in brackets) and noise parameter estimates for linear mixed model (LMM) and Student's t mixed model (STMM). The noise types include Gaussian (N), Student's t (S) and Gaussian mixture (GM), see the text for details.*

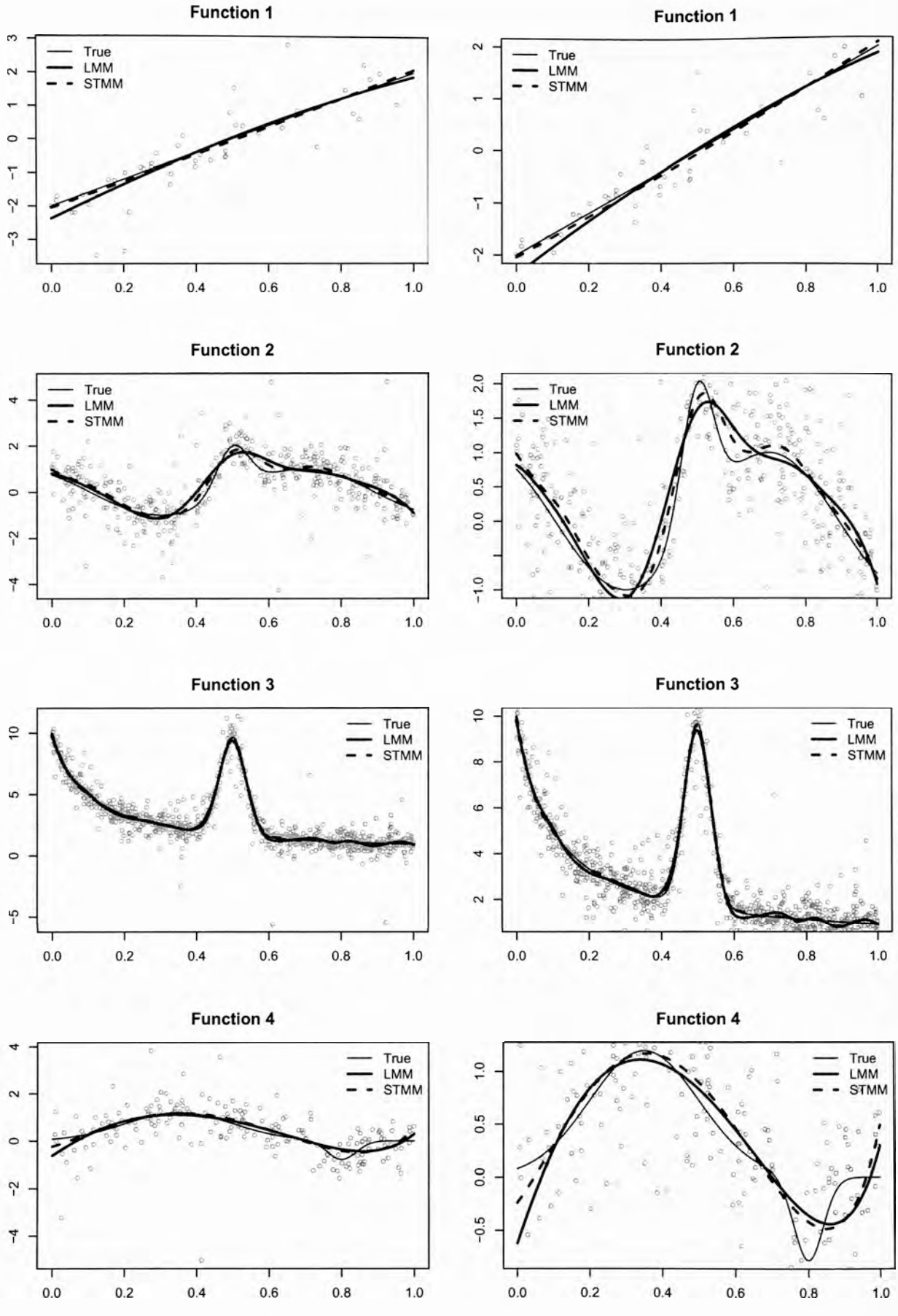


Figure 6.1: *Exemplar plots using a linear mixed model (LMM) smoother and Student's t mixed model (STMM) for $f_1, \ldots, f_4$ with student t noise with variance $\sigma_y^2 = 0.25$ and degrees of freedom $\nu_y = 3$. Left panels are limited to the range of the data and the right panels are limited to the range of the fitted functions.*

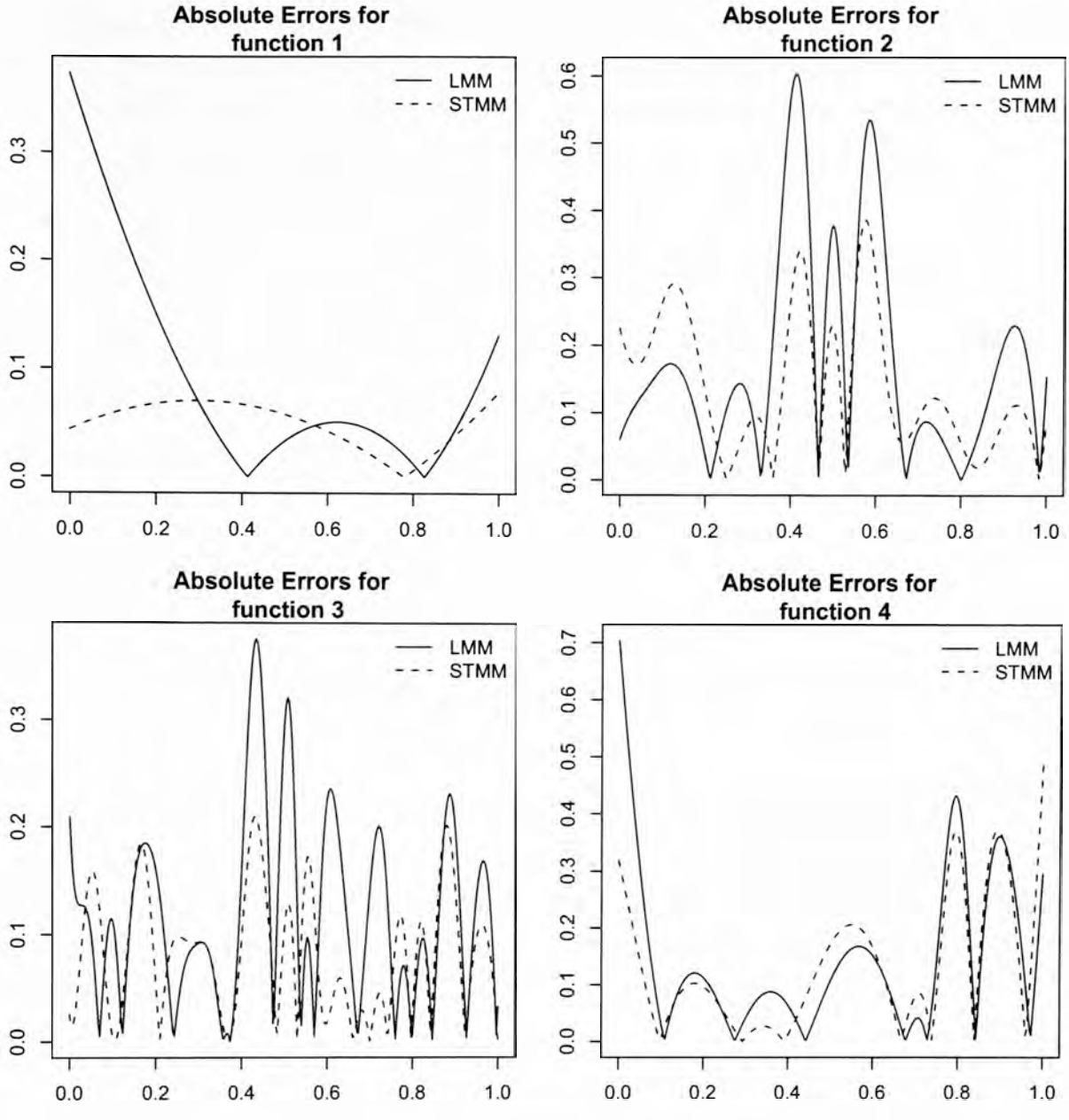


Figure 6.2: *Absolute errors for fits in Figure 6.1.*

The log-likelihood for this linear mixed model with non-constant variance function (LMMVF) can be written as

$$\begin{aligned}
&\ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\sigma}}^2) \\
&\quad = \log \int \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2} \mathbf{1}^T (\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{Z}}\widetilde{\mathbf{u}}) \right. \\
&\qquad \left. -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T \operatorname{diag}\left(e^{-\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} - \widetilde{\mathbf{Z}}\widetilde{\mathbf{u}}} \right) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) \right) \\
&\qquad \times \frac{|\mathbf{D}_{\boldsymbol{\sigma}^2}|^{\frac{1}{2}}}{(2\pi)^{\frac{q}{2}}} \exp\left(-\frac{\mathbf{u}^T \mathbf{D}_{\boldsymbol{\sigma}^2} \mathbf{u}}{2} \right) \frac{|\widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2}|^{\frac{1}{2}}}{(2\pi)^{\frac{\widetilde{q}}{2}}} \exp\left(-\frac{\widetilde{\mathbf{u}}^T \widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2} \widetilde{\mathbf{u}}}{2} \right) d\mathbf{u}\, d\widetilde{\mathbf{u}}.
\end{aligned}$$

The difficulty with calculating ℓ comes with trying to integrate out the vector $\tilde{\mathbf{u}}$. Using the density transform (6.2) with $\delta(\mathbf{u}, \tilde{\mathbf{u}}) = \delta_{\mathbf{u}}(\mathbf{u})\delta_{\tilde{\mathbf{u}}}(\tilde{\mathbf{u}})$ where

$$\begin{aligned} \mathbf{u}|\mathbf{y} &\sim_{\delta_{\mathbf{u}}} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \tilde{\mathbf{u}}|\mathbf{y} &\sim_{\delta_{\tilde{\mathbf{u}}}} N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}) \end{aligned}$$

we obtain a lower bound for ℓ is given by

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\sigma}}^2) &\geq \ell_L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\sigma}}^2; \boldsymbol{\xi}) \\ &= \mathbb{E}_\delta \log[\mathbf{y}|\mathbf{u}, \tilde{\mathbf{u}}] + \mathbb{E}_\delta \log[\mathbf{u}] + \mathbb{E}_\delta \log[\tilde{\mathbf{u}}] + \mathcal{H}_{\delta_{\mathbf{u}}} + \mathcal{H}_{\delta_{\tilde{\mathbf{u}}}}. \end{aligned}$$

Here $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ and, ignoring additive constants,

$$\begin{aligned} \mathbb{E}_\delta \log[\mathbf{y}|\mathbf{u}, \tilde{\mathbf{u}}] &= -\frac{1}{2}\mathbf{1}^T\tilde{\mathbf{C}}\tilde{\boldsymbol{\nu}} - \frac{1}{2}\tilde{\mathbf{y}}^T \exp\left(-\tilde{\mathbf{C}}\tilde{\boldsymbol{\nu}} + \frac{1}{2}\mathrm{dg}\left(\tilde{\mathbf{Z}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{Z}}^T\right)\right), \\ \mathbb{E}_\delta \log[\mathbf{u}] &= \frac{1}{2}\log(\mathbf{D}_{\boldsymbol{\sigma}^2}) - \frac{\boldsymbol{\mu}^T\mathbf{D}_{\boldsymbol{\sigma}^2}\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{D}_{\boldsymbol{\sigma}^2})}{2}, \\ \mathbb{E}_\delta \log[\tilde{\mathbf{u}}] &= \frac{1}{2}\log(\tilde{\mathbf{D}}_{\tilde{\boldsymbol{\sigma}}^2}) - \frac{\tilde{\boldsymbol{\mu}}^T\tilde{\mathbf{D}}_{\tilde{\boldsymbol{\sigma}}^2}\tilde{\boldsymbol{\mu}} + \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{D}}_{\tilde{\boldsymbol{\sigma}}^2})}{2}, \\ \mathcal{H}_{\delta_{\mathbf{u}}} &= \frac{1}{2}\log(\boldsymbol{\Sigma}) \\ \text{and } \mathcal{H}_{\delta_{\tilde{\mathbf{u}}}} &= \frac{1}{2}\log(\tilde{\boldsymbol{\Sigma}}). \end{aligned}$$

Here $\mathbf{C} \equiv [\mathbf{X}, \mathbf{Z}]$, $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}, \boldsymbol{\mu})$, $\tilde{\mathbf{C}} \equiv [\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}]$, $\tilde{\boldsymbol{\nu}} \equiv (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\mu}})$,

$$\tilde{y}_i = (\mathbf{y} - \mathbf{C}\boldsymbol{\nu})_i^2 + (\mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^T)_{ii},$$

$\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)$, $\mathrm{dg}(\mathbf{A})$ is the vector corresponding to the diagonal elements of $\mathbf{A}$ and $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ are additional variational parameters. Again we have used the fact $\mathbb{E}_\delta(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathbb{E}_\delta(\mathbf{x})^T\mathbf{A}\mathbb{E}_\delta(\mathbf{x}) + \mathrm{tr}(\mathbf{A}\mathrm{Cov}_\delta(\mathbf{x}))$ for any random vector $\mathbf{x}$ appropriately and sized matrix $\mathbf{A}$ but also that $\mathbb{E}(e^{\mathbf{t}^T\mathbf{x}}) = e^{\mathbf{t}^T\tilde{\boldsymbol{\mu}} + \frac{1}{2}\mathbf{t}^T\tilde{\boldsymbol{\Sigma}}\mathbf{t}}$ for any Gaussian vector $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and appropriately-sized vector $\mathbf{t}$.

The first derivatives of ℓ_L with respect to $\tilde{\boldsymbol{\nu}}$ are

$$\mathsf{D}_{\tilde{\boldsymbol{\nu}}}\ell_L = \frac{1}{2}\tilde{\mathbf{C}}^T\left(\tilde{\mathbf{y}} \odot e^{-\tilde{\mathbf{C}}\tilde{\boldsymbol{\nu}} + \frac{1}{2}\mathrm{dg}(\tilde{\mathbf{Z}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{Z}}^T)} - \mathbf{1}\right) - \tilde{\mathbf{B}}_{\tilde{\boldsymbol{\sigma}}^2}\tilde{\boldsymbol{\nu}} \tag{6.10}$$

where $\tilde{\mathbf{B}}_{\tilde{\boldsymbol{\sigma}}^2} \equiv \mathrm{blockdiag}\{\mathbf{0}, \tilde{\mathbf{D}}_{\tilde{\boldsymbol{\sigma}}^2}\}$.

The equation (6.10) correspond to fitting a gamma LMM for fixed $\tilde{\mathbf{y}}$ using the variational method described in Chapter 5 with

$$\tilde{\mathbf{y}}|\tilde{\mathbf{u}} \sim \mathrm{Gamma}\left(2e^{\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{Z}}\tilde{\mathbf{u}}}, \frac{1}{2}\right)$$

whereas in Ruppert *et al.* (2003) it was noted that, for fixed $(\boldsymbol{\beta}, \mathbf{u})$,

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^2|\tilde{\mathbf{u}} \sim \mathrm{Gamma}\left(2e^{\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{Z}}\tilde{\mathbf{u}}}, \frac{1}{2}\right).$$

Again we will defer discussion of the maximisation of ℓ_L with respect to $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\sigma}}^2)$ and $\boldsymbol{\xi}$ now we defer discussion of this to Section 6.5.1. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_y^2, \widehat{\widetilde{\boldsymbol{\beta}}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\widetilde{\boldsymbol{\sigma}}}^2)$ and $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\widetilde{\boldsymbol{\mu}}}, \widehat{\widetilde{\boldsymbol{\Sigma}}})$ be the values which maximise ℓ_L. The predictions for the mean are made using (6.6) and the variance function is estimated by

$$\widehat{g}(x) = \sum_{i=1}^{\widetilde{p}} \widetilde{X}_i(x)\widehat{\widetilde{\beta}}_i + \sum_{j=1}^{\widetilde{q}} \widetilde{Z}_i(x)\widehat{\widetilde{\mu}}_i.$$

6.3.1 *Numerical Experience*

To test the effectiveness of the above algorithm for fitting variance functions we will use the same functions as (6.7). We will also consider a variety of variance functions (with corresponding n values):

$$\begin{aligned}
g_1(x) &= \log(0.25) && (n = 200)\\
g_2(x) &= \log(0.5 - 0.8x + 1.6(x - 0.5)_+) && (n = 800)\\
g_3(x) &= \log\left(\frac{r}{32} + \frac{3r}{32}x^2\right) && (n = 400)\\
g_4(x) &= -3.9 + 1.7\exp(\sin(5\pi x)) && (n = 1600)
\end{aligned}$$

The x_is will be equally spaced between 0 and 1. We will use thin plate splines (see Chapter 1) to construct the $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{D}_i$ matrices for these experiments with $m = 3$, $K_1 = 25$ knots for the mean function and $\widetilde{K}_1 = 10$ knots for the variance function. Note that we standardised the xs to have zero mean and unit variance which typically improves numerical stability. The knots are spaced using the quantities of the unique xs as per equation (6.8).

We will measure the error for the estimated mean function by the sample mean square error (6.9) and we will measure the error for the variance functions by the mean deviance for gamma generalised linear models

$$\begin{aligned}
\overline{\mathcal{D}}(\mathbf{g}_j, \widehat{\mathbf{g}}) &= \frac{2}{n}\sum_{i=1}^{n} -\log\left(\frac{\exp(g_j(x_i))}{\exp(\widehat{g}(x_i))}\right) + \frac{\exp(g_j(x_i)) - \exp(\widehat{g}(x_i))}{\exp(\widehat{g}(x_i))}\\
&= \frac{2}{n}\sum_{i=1}^{n}(\widehat{g}_i - g_j(x_i)) + \exp(g_j(x_i) - \widehat{g}_i) - 1
\end{aligned}$$

noting that for the constant variance case, i.e. a LMM we use $\exp(\widehat{g}(x_i)) = \widehat{\sigma}_y^2$.

Each of these settings will be fit using a linear mixed model smoothing and the Variance Function (LMMVF) variational approximation for 100 trials. The median MSE and $\overline{\mathcal{D}}$ with standard error in brackets for each setting is summarised in Table 6.2.3.

From Table 6.2.3 we see that for mean functions f_1, f_2 and f_3 the LMMVF have a relatively minor impact on the accuracy of estimates for the mean function. Sometimes the MSEs for the means were a little better for LMMVF, sometimes a little worse. However the LMMVF approach reduced $\overline{\mathcal{D}}$ for the variance function compared with the LMM fit. This is important because in some applications the variance function is itself of intrinsic interest.

Finally Figure 6.3 illustrates some exemplar fits using a LMM and LMMVF for f_4 using the non-constant variance functions, and fits of the variance functions using the mean residuals. Note that in each case the fitted variance function for LMMVF has some resemblance with the true variance function and is particularly accurate for g_2. However, based on the absolute errors in Figure 6.3, the LMMVF fits in do not appear to improved the estimation of the mean function over the LMM fits.

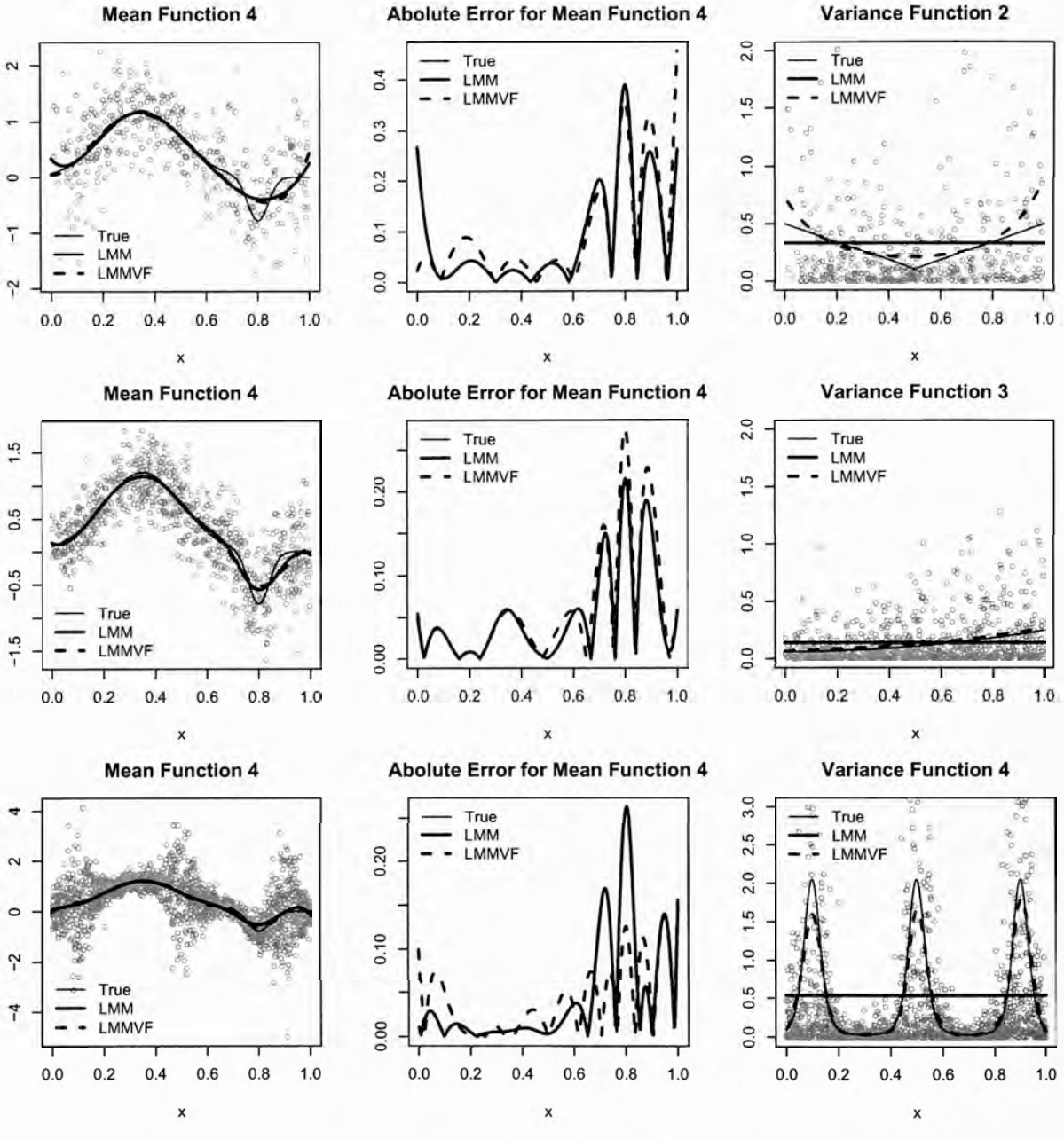


Figure 6.3: *Exemplar plots (left panels) and estimated variance, absolute errors (middle panels) and functions (right panels) for f_4 and variance function g_2, g_3 and g_4.*

f_j	g_k	Noise MSE LMM	Median MSE LMMVF	Median $\overline{\mathcal{D}}$ LMM	Median $\overline{\mathcal{D}}$ LMMVF
$j=1$	$k=1$	0.0026 (0.0003)	0.0032 (0.0002)	0.3693 (0.0046)	0.1699 (0.0131)
	$k=2$	0.0010 (0.0001)	0.0009 (0.0001)	0.4703 (0.0027)	0.0347 (0.0016)
	$k=3$	0.0014 (0.0002)	0.0014 (0.0002)	0.5801 (0.0030)	0.0568 (0.0043)
	$k=4$	0.0009 (0.0001)	0.0003 (<0.00005)	1.8551 (0.0032)	0.0221 (0.0008)
$j=2$	$k=1$	0.0252 (0.0010)	0.0260 (0.0012)	0.4044 (0.0050)	0.0438 (0.0044)
	$k=2$	0.0071 (0.0002)	0.0060 (0.0002)	0.4747 (0.0026)	0.0278 (0.0011)
	$k=3$	0.0095 (0.0003)	0.0091 (0.0003)	0.7288 (0.0048)	0.0176 (0.0015)
	$k=4$	0.0060 (0.0003)	0.0076 (0.0004)	1.8593 (0.0035)	0.0256 (0.0010)
$j=3$	$k=1$	0.0294 (0.0008)	0.0302 (0.0008)	0.3852 (0.0049)	0.0301 (0.0027)
	$k=2$	0.0101 (0.0003)	0.0098 (0.0003)	0.4734 (0.0025)	0.0213 (0.0008)
	$k=3$	0.0352 (0.0010)	0.0335 (0.0010)	0.2646 (0.0024)	0.0181 (0.0014)
	$k=4$	0.0080 (0.0003)	0.0086 (0.0003)	1.8530 (0.0030)	0.0256 (0.0008)
$j=4$	$k=1$	0.0247 (0.0005)	0.0245 (0.0005)	0.4138 (0.0041)	0.0305 (0.0030)
	$k=2$	0.0143 (0.0005)	0.0169 (0.0005)	0.4904 (0.0028)	0.0330 (0.0011)
	$k=3$	0.0092 (0.0004)	0.0158 (0.0004)	1.0163 (0.0062)	0.0231 (0.0018)
	$k=4$	0.0089 (0.0004)	0.0038 (0.0002)	1.8594 (0.0036)	0.0259 (0.0009)

Table 6.3.2: *Mean square errors (MSE), variance function mean deviances $\overline{\mathcal{D}}$ and standard errors (in brackets), for linear mixed model (LMM) and variational approximation of the variance function model (LMMVF).*

6.4 Spatially Adaptive Variance Components

Similar to the variance function model described in the previous section is the spatially adaptive variance component scheme first proposed by Baladandayuthapani *et al.* (2005) which was based on earlier work by Ruppert & Carroll (2000). Suppose that now we model the relationship between the ys and the xs in (6.1) using the Gaussian distribution

$$y_i|\mathbf{u};\boldsymbol{\beta},\sigma_y^2,\nu_y \sim N(f(\mathbf{x}_i),\sigma_y^2)$$

where σ_y^2 is a constant parameter to be estimated and $f(\cdot)$ is the mean function. Consider the additive model for the mean

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{v} f_i(\mathbf{x}_{I_i}).$$

Here $\mathcal{I} = \{I_1,\ldots,I_v\}$ form a partition of a subset of the indices $\{1,\ldots,d\}$, (see Section 1.2.1 for examples) and

$$f_i(\mathbf{x}_{I_i}) = \sum_{j=1}^{p_i} \beta_{ij} X_{ij}(\mathbf{x}_{I_i}) + \sum_{j=1}^{K_i} u_{ij} Z_{ij}(\mathbf{x}_{I_i};\boldsymbol{\kappa}_{ij})$$

where $\boldsymbol{\beta}_i = (\beta_{i1},\ldots,\beta_{ip_i})$ are coefficients for $\{X_i(\cdot)\}_{i=1}^{p_i}$ and $\mathbf{u}_i = (u_{i1},\ldots,u_{iK_i})$ are the coefficients for the spline functions $\mathcal{Z} = \{Z_{ij}(\cdot;\boldsymbol{\kappa}_{ij})\}_{j=1}^{K_i}$. Here, unlike Chapter 1, we explicitly assume that each of the spline functions $\mathcal{Z}$ depends depend on $\mathbf{x}_{I_i}$ locally around knots $\{\boldsymbol{\kappa}_{ij}\}_{1\leq j\leq K_i}$ (with dimension equal to the dimension of $\mathbf{x}_{I_i}$) and that the number of knots is equal to the number of basis functions.

Unfortunately, the usual splines used in previous chapters, i.e. the mixed model O'Sullivan splines described in Chapter 2, do not work with this model. The problem that arises is that the number of splines in the B-spline basis is not equal to the number of knots. Example of such splines which do satisfy these assumptions are truncated power splines and thin plate splines (see Sections 1.2.3–1.2.4). These splines were used in Baladandayuthapani *et al.* (2005) and Krivobokova *et al.* (2007).

Thus far we have assumed constant variance components, i.e.

$$\mathbf{u}_i \sim N(\mathbf{0},\sigma_i^2\mathbf{I})$$

where σ_i^2 are model parameters to be estimated. Since the spline functions depend on the xs locally around the knot locations we can make the penalty spatially adaptive by allowing variance components to depend on the $\boldsymbol{\kappa}_{ij}$s. Thus the idea behind adaptive variance components, similar to the various function model, is to have

$$\log(\sigma_i^2(\boldsymbol{\kappa})) = h_i(\boldsymbol{\kappa}).$$

Again, using the same penalised spline methodology used throughout this thesis, we model $h_i(\boldsymbol{\kappa})$ using

$$h_i(\boldsymbol{\kappa}) = \sum_{j=1}^{\overline{p}} \overline{\beta}_{ij} \overline{X}_{ij}(\boldsymbol{\kappa}) + \sum_{j=1}^{\overline{K}_i} \overline{u}_{ij} \overline{Z}_{ij}(\boldsymbol{\kappa}; \boldsymbol{\tau}_{ik})$$

where $\{\overline{X}_{ij}(\cdot)\}_{j=1}^{\overline{p}_i}$ and $\{\overline{Z}_{ij}(\cdot;\boldsymbol{\tau}_{ij})\}_{j=1}^{\overline{K}_i}$ are basis functions truncated power or thin plate spline basis matrices with knots $(\boldsymbol{\tau}_{ij}, \ldots, \boldsymbol{\tau}_{i\overline{K}_i})$ and $\overline{K}_i \leq K_i$.

Thus we consider the model

$$\begin{aligned}
\mathbf{y}|\mathbf{u}, \sigma_y^2 &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_y^2\mathbf{I}) \\
\mathbf{u}_i|\boldsymbol{\sigma}_i^2 &\sim N(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_i^2)) \\
\log[\boldsymbol{\sigma}_i^2|\overline{\mathbf{u}}_i] &= \overline{\mathbf{X}}_i\overline{\boldsymbol{\beta}}_i + \overline{\mathbf{Z}}_i\overline{\mathbf{u}}_i \\
\overline{\mathbf{u}}_i|\overline{\sigma}_i^2 &\sim N(\mathbf{0}, \overline{\sigma}_i^2\mathbf{I})
\end{aligned}$$

for $1 \leq i \leq v$ where $\mathbf{X} = [\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_v]$, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_v)$, $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_v]$, $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_v)$, $\boldsymbol{\sigma}_i^2 = (\sigma_{i1}^2, \ldots, \sigma_{iK_i}^2)$ with $\sigma_{ij}^2 = \sigma_i^2(\boldsymbol{\kappa}_j)$ and

$$\begin{aligned}
[\mathbf{X}_i]_{jk} &= X_{ik}(\mathbf{x}_{jI_i}) && \text{for } 1 \leq j \leq n, 1 \leq k \leq p_i \\
[\mathbf{Z}_i]_{jk} &= Z_{ik}(\mathbf{x}_{jI_i}; \boldsymbol{\kappa}_{ik}) && \text{for } 1 \leq j \leq n, 1 \leq k \leq K_i \\
[\overline{\mathbf{X}}_i]_{jk} &= \overline{X}_{ik}(\boldsymbol{\kappa}_{ij}) && \text{for } 1 \leq j \leq K_i, 1 \leq k \leq \overline{p}_i \\
[\overline{\mathbf{Z}}_i]_{jk} &= \overline{Z}_{ik}(\boldsymbol{\kappa}_{ij}; \boldsymbol{\tau}_{ik}) && \text{for } 1 \leq j \leq K_i, 1 \leq k \leq \overline{K}_i
\end{aligned}$$

so that $\mathbf{X}_i$ is a $n \times p$ matrix, $\mathbf{Z}_i$ is a $n \times K_i$ matrix, $\overline{\mathbf{X}}_i$ is a $K_i \times \overline{p}_i$ matrix and $\overline{\mathbf{Z}}_i$ is a $K_i \times \overline{K}_i$ matrix.

With some additional priors Baladandayuthapani *et al.* (2005) fitted this model using an MCMC approach. Krivobokova *et al.* (2007) fit this model using the Laplace's method and Crainiceanu *et al.* (2007) combine this idea with variance component fitting with yet another MCMC scheme. The MCMC schemes, particularly in the last reference, are quite complicated and too slow for some contexts. Krivobokova *et al.* (2007) show that approximations can be developed for the above adaptive penalty model and, while not quite as accurate as the current best smoothing techniques, are much faster and easier to implement.

The log-likelihood for this model is given by

$$\begin{aligned}
&\ell(\boldsymbol{\beta}, \phi, \overline{\boldsymbol{\beta}}_1, \ldots, \overline{\boldsymbol{\beta}}_v, \overline{\boldsymbol{\sigma}}^2) \\
&\quad = \log \int [\mathbf{y}|\mathbf{u}, \sigma_y^2][\mathbf{u}|\overline{\mathbf{u}}_1, \ldots, \overline{\mathbf{u}}_v] \prod_{i=1}^{v} [\overline{\mathbf{u}}_i|\overline{\sigma}_i^2] d\mathbf{u} d\overline{\mathbf{u}}_1 \ldots d\overline{\mathbf{u}}_v \\
&\quad = \log \int \frac{1}{(2\pi\sigma_y^2)^{\frac{n}{2}}} \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2}{2\sigma_y^2}\right) \\
&\qquad \times \prod_{i=1}^{v} \frac{1}{(2\pi)^{\frac{K_i}{2}}} \exp\left(-\frac{1}{2}\mathbf{1}^T\left(\overline{\mathbf{X}}_i\overline{\boldsymbol{\beta}}_i + \overline{\mathbf{Z}}_i\overline{\mathbf{u}}_i\right) - \frac{\mathbf{u}_i^T \mathrm{diag}\left(e^{-\overline{\mathbf{X}}_i\overline{\boldsymbol{\beta}}_i - \overline{\mathbf{Z}}_i\overline{\mathbf{u}}_i}\right)\mathbf{u}_i}{2}\right) \\
&\qquad \times \prod_{i=1}^{v} \frac{1}{(2\pi\overline{\sigma}_i^2)^{\frac{\overline{K}_i}{2}}} \exp\left(-\frac{\|\overline{\mathbf{u}}_i\|^2}{2\overline{\sigma}_i^2}\right) d\mathbf{u} d\overline{\mathbf{u}}_1 \ldots d\overline{\mathbf{u}}_v.
\end{aligned}$$

It is clear that the difficulty in fitting the above model stems from integrating out the vectors $\mathbf{u}, \overline{\mathbf{u}}_1, \ldots, \overline{\mathbf{u}}_v$. We propose the following variational approximation based on the density transform (6.2) with $\delta(\mathbf{u}, \overline{\mathbf{u}}_1, \ldots, \overline{\mathbf{u}}_v) = \delta_{\mathbf{u}}(\mathbf{u}) \prod_{i=1}^{v} \delta_{\overline{\mathbf{u}}_i}(\overline{\mathbf{u}}_i)$ where

$$\begin{aligned} \mathbf{u}|\mathbf{y} &\sim_{\delta_{\mathbf{u}}} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \overline{\mathbf{u}}_i|\mathbf{y} &\sim_{\delta_{\overline{\mathbf{u}}_i}} N(\overline{\boldsymbol{\mu}}_i, \overline{\boldsymbol{\Sigma}}_i) \end{aligned}$$

to obtain a lower bound on ℓ given by

$$\begin{aligned} &\ell(\boldsymbol{\beta}, \sigma_y^2, \overline{\boldsymbol{\beta}}_1, \ldots, \overline{\boldsymbol{\beta}}_v, \overline{\boldsymbol{\sigma}}^2) \geq \ell_L(\boldsymbol{\beta}, \phi, \overline{\boldsymbol{\beta}}_1, \ldots, \overline{\boldsymbol{\beta}}_v, \overline{\boldsymbol{\sigma}}^2; \boldsymbol{\xi}) \\ &= \mathbb{E}_\delta \log[\mathbf{y}|\mathbf{u}, \sigma_y^2] + \mathbb{E}_\delta \log[\mathbf{u}|\overline{\mathbf{u}}_1, \ldots, \overline{\mathbf{u}}_v] + \mathcal{H}_{\delta_{\mathbf{u}}} + \sum_{i=1}^{v} \mathbb{E}_\delta \log[\overline{\mathbf{u}}_i|\overline{\sigma}_i^2] + \mathcal{H}_{\delta_{\overline{\mathbf{u}}_i}}. \end{aligned}$$

Here $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \overline{\boldsymbol{\mu}}_1, \ldots, \overline{\boldsymbol{\mu}}_v, \overline{\boldsymbol{\Sigma}}_1, \ldots, \overline{\boldsymbol{\Sigma}}_v)$ and (ignoring additive constants)

$$\begin{aligned} \mathbb{E}_\delta \log[\mathbf{y}|\mathbf{u}, \sigma_y^2] &= -\frac{n}{n} \log(\sigma_y^2) - \frac{\|\mathbf{y} - \mathbf{C}\boldsymbol{\nu}\|^2 + \operatorname{tr}(\boldsymbol{\Sigma}\mathbf{D})}{2\sigma_y^2}, \\ \mathbb{E}_\delta \log[\mathbf{u}|\overline{\mathbf{u}}_1, \ldots, \overline{\mathbf{u}}_v] &= \sum_{i=1}^{v} -\frac{1}{2}\mathbf{1}^T \overline{\mathbf{C}}_i \overline{\boldsymbol{\nu}}_i - \frac{1}{2}\overline{\mathbf{y}}_i^T \exp\left(-\overline{\mathbf{C}}_i \overline{\boldsymbol{\nu}}_i + \frac{1}{2}\operatorname{dg}\left(\overline{\mathbf{Z}}_i \overline{\boldsymbol{\Sigma}}_i \overline{\mathbf{Z}}_i^T\right)\right), \\ \mathbb{E}_\delta \log[\overline{\mathbf{u}}_i|\overline{\sigma}_i^2] &= -\frac{K_i}{2} \log(\overline{\sigma}_i^2) - \frac{\|\overline{\boldsymbol{\mu}}_i\|^2 + \operatorname{tr}(\overline{\boldsymbol{\Sigma}}_i)}{2\overline{\sigma}_i^2}, \\ \mathcal{H}_{\delta_{\mathbf{u}}} &= \frac{1}{2} \log|\boldsymbol{\Sigma}|, \\ \text{and } \mathcal{H}_{\delta_{\overline{\mathbf{u}}_i}} &= \frac{1}{2} \log|\overline{\boldsymbol{\Sigma}}_i| \end{aligned}$$

where $\mathbf{C} \equiv [\mathbf{X}, \mathbf{Z}]$, $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}, \boldsymbol{\mu})$, $\overline{\mathbf{C}}_i \equiv [\overline{\mathbf{X}}_i, \overline{\mathbf{Z}}_i]$, $\overline{\boldsymbol{\nu}}_i \equiv (\overline{\boldsymbol{\beta}}_i, \overline{\boldsymbol{\mu}}_i)$ for $1 \leq i \leq v$,

$$\mathbf{D} = \operatorname{diag}\left(e^{-\overline{\mathbf{C}}_1\overline{\boldsymbol{\nu}}_1 + \frac{1}{2}\operatorname{dg}\left(\overline{\mathbf{Z}}_1\overline{\boldsymbol{\Sigma}}_1\overline{\mathbf{Z}}_1^T\right)}, \ldots, e^{-\overline{\mathbf{C}}_v\overline{\boldsymbol{\nu}}_v + \frac{1}{2}\operatorname{dg}\left(\overline{\mathbf{Z}}_v\overline{\boldsymbol{\Sigma}}_v\overline{\mathbf{Z}}_v^T\right)}\right).$$

Note that we have used the same indexing for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as $\mathbf{u}$, i.e.

$$\begin{aligned} \boldsymbol{\mu}_i &= (\mu_{i1}, \ldots, \mu_{iK_i}) \\ [\boldsymbol{\Sigma}_i]_{jk} &= [\boldsymbol{\Sigma}]_{s(i,j),s(i,k)} \end{aligned}$$

for $1 \leq i \leq v$, $1 \leq j \leq K_i$, $1 \leq k \leq K_i$ with $s(i,j) = j + \sum_{k=1}^{i-1} K_k$ so that

$$\overline{\mathbf{y}}_i = \boldsymbol{\mu}_i^2 + \operatorname{diag}(\boldsymbol{\Sigma}_i),\ 1 \leq i \leq n.$$

The first derivatives of ℓ_L with respect to $\overline{\boldsymbol{\nu}}_i$ are

$$\mathsf{D}_{\overline{\boldsymbol{\nu}}_i}\ell_L = \frac{1}{2}\overline{\mathbf{C}}_i^T\left(\overline{\mathbf{y}}_i \odot e^{-\overline{\mathbf{C}}_i\overline{\boldsymbol{\nu}}_i + \frac{1}{2}\operatorname{dg}\left(\overline{\mathbf{Z}}_i\overline{\boldsymbol{\Sigma}}_i\overline{\mathbf{Z}}_i^T\right)} - \mathbf{1}\right) - \overline{\sigma}_i^{-2}\overline{\boldsymbol{\nu}}_i. \tag{6.11}$$

This equation correspond to fitting a gamma GLMM for fixed $\overline{\mathbf{y}}_i$ using the variational method described in Chapter 5 with

$$\overline{\mathbf{y}}_i|\overline{\mathbf{u}}_i \sim \text{Gamma}\left(2e^{\overline{\mathbf{X}}_i\overline{\boldsymbol{\beta}}_i + \overline{\mathbf{Z}}_i\overline{\mathbf{u}}_i}, \frac{1}{2}\right)$$

whereas for fixed $\mathbf{u}_i$ we should have

$$\mathbf{u}_i^2|\overline{\mathbf{u}}_i \sim \text{Gamma}\left(2e^{\overline{\mathbf{X}}_i\overline{\boldsymbol{\beta}}_i+\overline{\mathbf{Z}}_i\overline{\mathbf{u}}_i}, \frac{1}{2}\right).$$

Again we will defer discussion of the maximization of ℓ_L with respect to $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_y^2, \overline{\boldsymbol{\beta}}_1, \ldots, \overline{\boldsymbol{\beta}}_v, \overline{\boldsymbol{\sigma}}^2)$ and $\boldsymbol{\xi}$ now we defer discussion of this to Section 6.5.1.

6.4.1 *Numerical Experience*

To test the effectiveness of the above algorithm for fitting our variational approximation of the adaptive variance component (AVC) model we will compare this method with some of the latest methods for spatially adaptive smoothing. It is difficult to make extensive comparisons given the fact that many papers use different functions to test the effectiveness of various methods. We will restrict our method with the methods:

- Spatially-adaptive penalties for spline fitting method (RC) of Ruppert & Carroll (2000),
- Bayesian Adaptive Regression Splines (BARS) of DiMatteo *et al.* (2001),
- Bayesian P-splines (BPS) of Baladandayuthapani *et al.* (2005),
- Spatially adaptive Bayesian P-Splines with heteroscedastic errors (CRC) Crainiceanu *et al.* (2007) and
- `AdaptFit` of Krivobokova *et al.* (2007)

which compare some of the same problems. We note that the methods `AdaptFit` and AVC correspond to Laplace's and a variational approximation of the model proposed in BPS. Furthermore BPS is itself a similar to the local penalty method developed in RC. The CRC method is again a similar model to BPS but includes components for variance function estimation. Finally, BARS uses free-knots splines with the random number and location of knots, using reversible jump MCMC for estimation which is completely different from other methods considered here.

We will use the following functions and settings for each method

$$\begin{aligned}
f_5(x) &= \sqrt{x(1-x)}\sin\left(\frac{2\pi\left(1+2^{(9-4J)/5}\right)}{x+2^{(9-4J)/5}}\right) && n = 400, K = 6, \sigma_y^2 = 0.04 \\
f_6(x) &= \sqrt{x(1-x)}\sin\left(\frac{2\pi\left(1+2^{(9-4J)/5}\right)}{x+2^{(9-4J)/5}}\right) && n = 400, K = 3, \sigma_y^2 = 0.04 \\
f_7(x) &= \exp\left(-400(x-0.6)^2\right) + \frac{5}{3}\exp\left(-500(x-0.75)^2\right) \\
&\quad +2\exp\left(-500(x-0.9)^2\right) && n = 1000, \sigma_y^2 = 0.25.
\end{aligned} \tag{6.12}$$

The xs were equally spaced between 0 and 1. We will use thin plate splines (see Chapter 1) for these experiments with $K_1 = 80$ knots for the mean function and $\overline{K}_1 = 20$ for the construction of the matrices $\mathbf{X}$, $\mathbf{Z}$ and $(\overline{\mathbf{X}}_i, \overline{\mathbf{Z}}_i)$, $1 \leq i \leq v$. These knots are spaced using the quantities of the unique xs as per equation (6.8). Note that we standardised the xs to have zero mean and unit variance which typically improves numerical stability.

We will measure the error of each fit by the sample mean square error (6.9) which we will average over 100 repeated simulations of each dataset. The MSEs for the AVC method and the reported MSEs for each other method is summarised in Table 6.4.3. From this table we see that the MSEs for AVC is similar to the other methods.

Method	f_5	f_6	f_7
RC	0.0026	0.0007	0.0065
BARS			0.0043
BPS	0.0027	0.0006	0.0061
CRC			0.0054
AdaptFit	0.0034		0.0048
AVC	0.0033	0.0008	0.0047

Table 6.4.3: *Mean square errors for functions f_5, f_6 and f_7 using the methods: Spatially-adaptive penalties for spline fitting method (RC) of Ruppert & Carroll (2000), BARS (DiMatteo* et al., *2001), Bayesian P-splines (BPS, Baladandayuthapani* et al.*,2005), Spatially adaptive Bayesian P-Splines with heteroscedastic errors (CRC, Crainiceanu* et al., *2007),* `AdaptFit` *(Krivobokova* et al., *2007) and the variational approximation of the adaptive variance component model (AVC).*

The reported times taken for each of the methods are remarkably different. The fits for AVC each took between on average 66 seconds for f_5 and f_6 and 132 seconds for f_7. In comparison the reported time for the RC method was about 10 seconds, for the `AdaptFit` method half a minute, the BARS method takes as long as 4 hours to fit one model and finally while no time reported by Crainiceanu *et al.* (2007) for CRC to fit a single model they did report the total time for all simulations being over 1000 hours.

Finally, Figure 6.4 illustrates some exemplar fits along with coefficient "responses" $\overline{\mathbf{y}}_1$ and fits

$$\boldsymbol{\sigma}_1^2(\kappa) = \exp\left(\overline{\mathbf{X}}_1\widehat{\overline{\boldsymbol{\beta}}}_1 + \overline{\mathbf{Z}}_1\widehat{\overline{\mathbf{u}}}_1\right). \tag{6.13}$$

Note that for these cases all of the variance component functions using the variational approximation are linear.

6.5 Optimisation, Alternatives and Extensions

In this chapter we have thus far seen how to derive variational approximations for Student's t noise, variance function and spatially adaptive variance components. If we wished we could also, with relative ease, combine each of these.

Consider the model with student's response model with non-constant variance function and spatially adaptive variance components.

$$\begin{aligned}
y_i|\mathbf{u}, \boldsymbol{\sigma}_y^2, \boldsymbol{\gamma}_y &\sim N\left((\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})_i^2, \gamma_{y,i}^{-1}\sigma_{y,i}^2\right), \\
u_{ij}|\sigma_{ij}^2 &\sim N\left(0, \sigma_{ij}^2\right), \\
\gamma_{y,i} &\sim \text{Gamma}\left(\frac{\nu_y}{2}, \frac{\nu_y}{2}\right), \\
\sigma_{y,i}^2 &= \exp\left((\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{Z}}\widetilde{\mathbf{u}})_i\right), \\
\widetilde{\mathbf{u}} &\sim N(\mathbf{0}, \widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2}), \\
\sigma_{ij}^2 &= \exp\left((\overline{\mathbf{X}}_i\overline{\boldsymbol{\beta}}_i + \overline{\mathbf{Z}}_i\overline{\mathbf{u}}_i)_j\right) \\
\text{and } \overline{\mathbf{u}}_i &\sim N(\mathbf{0}, \overline{\sigma}_i^2\mathbf{I})
\end{aligned}$$

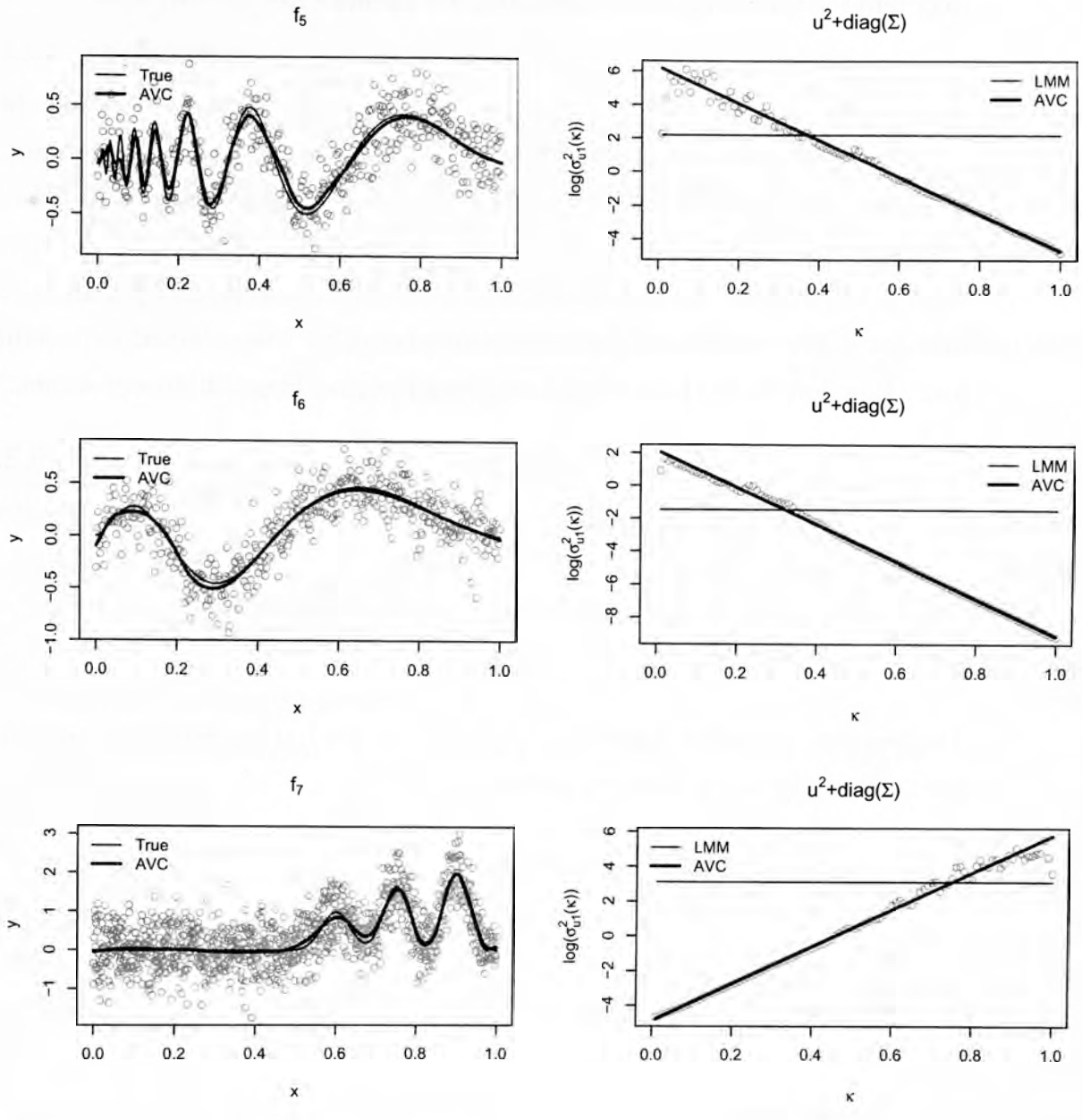


Figure 6.4: *Exemplar plots (left panels) for f_5, f_6 and f_7 using a LMM an the variational approximation to the adaptive variance component model (AVC) and fitted variance component functions (right panels) for coefficient "response" values.*

where we have used the same notation as specified in this chapter. Using a variational approach we use the density transform (6.2) with $\delta(\mathbf{u}, \widetilde{\mathbf{u}}, \overline{\mathbf{u}}, \boldsymbol{\gamma}_y) = \delta_{\mathbf{u}}(\mathbf{u})\delta_{\widetilde{\mathbf{u}}}(\widetilde{\mathbf{u}})\delta_{\boldsymbol{\gamma}_y}(\boldsymbol{\gamma}_y)\prod_{i=1}^{v}\delta_{\overline{\mathbf{u}}_i}(\overline{\mathbf{u}}_i)$ where

$$\begin{aligned}
\mathbf{u}|\mathbf{y} &\sim_{\delta_{\mathbf{u}}} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\widetilde{\mathbf{u}}|\mathbf{y} &\sim_{\delta_{\widetilde{\mathbf{u}}}} N(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}) \\
\overline{\mathbf{u}}_i|\mathbf{y} &\sim_{\delta_{\boldsymbol{\gamma}_y}} N(\overline{\boldsymbol{\mu}}_i, \overline{\boldsymbol{\Sigma}}_i) \\
\gamma_{y,i}|\mathbf{y} &\sim_{\delta_{\overline{\mathbf{u}}_i}} \text{Gamma}\,(A_{y,i}, B_{y,i})
\end{aligned}$$

to obtain the following lower bound on the likelihood

$$\begin{aligned}\ell(\boldsymbol{\theta}) &\geq \ell_L(\boldsymbol{\theta};\boldsymbol{\xi}) \\ &= \mathbb{E}_\delta \log[\mathbf{y}|\mathbf{u},\boldsymbol{\sigma}_y^2,\boldsymbol{\gamma}_y] + \mathbb{E}_\delta \log[\widetilde{\mathbf{u}}] + \mathbb{E}_\delta \log[\boldsymbol{\gamma}_y] \\ &\quad + \mathbb{E}_\delta \log[\mathbf{u}|\boldsymbol{\sigma}^2] + \mathcal{H}_{\delta_{\boldsymbol{\gamma}_y}} + \mathcal{H}_{\delta_{\mathbf{u}}} + \mathcal{H}_{\delta_{\widetilde{\mathbf{u}}}} + \sum_{i=1}^{v} \mathbb{E}_\delta(\log[\overline{\mathbf{u}}_i]) + \mathcal{H}_{\delta_{\overline{\mathbf{u}}_i}}.\end{aligned} \tag{6.14}$$

Here $\boldsymbol{\theta} = (\boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\sigma}}^2, \overline{\boldsymbol{\beta}}_1, \ldots, \overline{\boldsymbol{\beta}}_v, \overline{\boldsymbol{\sigma}}^2)$ and $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}, \overline{\boldsymbol{\mu}}_1, \ldots, \overline{\boldsymbol{\mu}}_v, \overline{\boldsymbol{\Sigma}}_1, \ldots, \overline{\boldsymbol{\Sigma}}_v)$ are the likelihood and variational parameters respectively. The relevant expectations and entropy function for the base model are given by, ignoring additive constants,

$$\begin{aligned}\mathbb{E}_\delta \log[\mathbf{y}|\mathbf{u},\boldsymbol{\sigma}_y^2,\boldsymbol{\gamma}_y] &= \sum_{i=1}^{n} \frac{\psi(A_{y,i}) - \log(B_{y,i}) - \mathbb{E}_q(\log(\sigma_{y,i}^{-2}))}{2} - \frac{A_{y,i}\widetilde{y}_i \mathbb{E}_q(\sigma_{y,i}^{-2})}{2B_{y,i}}, \\ \mathbb{E}_\delta \log[\mathbf{u}|\boldsymbol{\sigma}^2] &= \sum_{i=1}^{v}\sum_{j=1}^{K_i} -\frac{\mathbb{E}_q(\log(\sigma_{ij}^{-2}))}{2} - \frac{\overline{y}_{ij}\mathbb{E}_q(\sigma_{ij}^{-2})}{2} \\ \text{and } \mathcal{H}_{\delta_{\mathbf{u}}} &= \frac{1}{2}\log|\boldsymbol{\Sigma}|.\end{aligned}$$

The relevant expectation and entropy function for having Student's t-distributed noise are given by, ignoring additive constants,

$$\begin{aligned}\mathbb{E}_\delta \log[\boldsymbol{\gamma}_y] &= \frac{n\nu_y}{2}\log\left(\frac{\nu_y}{2}\right) - n\log\Gamma\left(\frac{\nu_y}{2}\right) \\ &\quad + \left(\frac{\nu_y}{2} - 1\right)\sum_{i=1}^{n}(\psi(A_{y,i}) - \log(B_{y,i})) - \frac{A_{y,i}\nu_y}{2B_{y,i}} \\ \text{and } \mathcal{H}_{\delta_{\boldsymbol{\gamma}_y}} &= \sum_{i=1}^{n} A_{y,i} - \log(B_{y,i}) + \log\Gamma(A_{y,i}) + (1 - A_{y,i})\psi(A_{y,i}).\end{aligned}$$

The relevant expectations and entropy function for having a non-constant variance function are given by, ignoring additive constants,

$$\begin{aligned}\mathbb{E}_\delta\left(\log(\sigma_{y,i}^2)\right) &= \widetilde{\mathbf{C}}\widetilde{\boldsymbol{\nu}}, \\ \mathbb{E}_\delta\left(\sigma_{y,i}^{-2}\right) &= \exp\left(-(\widetilde{\mathbf{C}}\widetilde{\boldsymbol{\nu}})_i + \frac{1}{2}(\widetilde{\mathbf{Z}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{Z}}^T)_{ii}\right), \\ \mathbb{E}_\delta \log[\widetilde{\mathbf{u}}] &= \frac{1}{2}\log|\widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2}| - \frac{\widetilde{\boldsymbol{\mu}}^T\widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2}\widetilde{\boldsymbol{\mu}} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2})}{2} \\ \text{and } \mathcal{H}_{\delta_{\widetilde{\mathbf{u}}}} &= \frac{1}{2}\log|\widetilde{\boldsymbol{\Sigma}}|\end{aligned}$$

where $\widetilde{\mathbf{C}} \equiv [\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}}]$ and $\widetilde{\boldsymbol{\nu}} \equiv (\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\mu}})$. Finally, the relevant expectations and entropy function for having adaptive variance components are given by, ignoring additive constants,

$$\begin{aligned}\mathbb{E}_\delta\left(\log(\sigma_{ij}^2)\right) &= \overline{\mathbf{C}}_i\overline{\boldsymbol{\nu}}_i, \\ \mathbb{E}_\delta\left(\sigma_{ij}^{-2}\right) &= \exp\left(-(\overline{\mathbf{C}}_i\overline{\boldsymbol{\nu}}_i)_j + \frac{1}{2}(\overline{\mathbf{Z}}_i\overline{\boldsymbol{\Sigma}}_i\overline{\mathbf{Z}}_i^T)_{jj}\right), \\ \mathbb{E}_\delta \log[\overline{\mathbf{u}}_i] &= \frac{\overline{K}_i}{2}\log(\overline{\sigma}_i^2) - \frac{\|\overline{\boldsymbol{\mu}}_i\|^2 + \mathrm{tr}(\overline{\boldsymbol{\Sigma}}_i)}{2\overline{\sigma}_i^2} \\ \text{and } \mathcal{H}_{\delta_{\overline{\mathbf{u}}_i}} &= \frac{1}{2}\log|\overline{\boldsymbol{\Sigma}}_i|\end{aligned}$$

where $\overline{\mathbf{C}}_i \equiv [\overline{\mathbf{X}}_i, \overline{\mathbf{Z}}_i]$ and $\overline{\boldsymbol{\nu}}_i \equiv (\overline{\boldsymbol{\beta}}_i, \overline{\boldsymbol{\mu}}_I)$ for $1 \leq i \leq v$.

6.5.1 *Optimisation*

Maximization of the function $\ell_L(\boldsymbol{\theta}; \boldsymbol{\xi})$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ is difficult due to the large number of parameters and the complex interactions between them. Newton-Raphson and quasi-Newton methods on their own are also unsatisfactory because of the storage requirements for storing the Hessian or approximate Hessian (for quasi-Newton methods) are high due to the large number of parameters.

Let us first consider the first derivatives of ℓ_L with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. The first derivatives of ℓ_L with respect to $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$ are given by

$$\begin{aligned} \mathsf{D}_{\boldsymbol{\nu}}\ell_L &= \mathbf{C}^T \mathrm{diag}(\mathbf{w})(\mathbf{y} - \mathbf{C}\boldsymbol{\nu}) - \mathbf{B}\boldsymbol{\nu} && (6.15) \\ \mathsf{D}_{\Sigma_{ij}} &= \mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{-1} - \mathbf{Z}^T \mathrm{diag}(\mathbf{w})\mathbf{Z} - \mathbf{D}\right)\mathbf{E}_{ij}\right)/2 && (6.16) \end{aligned}$$

where $\mathbf{B} = \mathrm{blockdiag}\{\mathbf{0}_p, \mathbf{D}\}$, $\mathbf{D} = \mathrm{blockdiag}(\mathrm{diag}(\mathbf{d}_1), \ldots, \mathrm{diag}(\mathbf{d}_v))$, $[\mathbf{w}]_i = A_i \mathbb{E}_q(\sigma_{y,i}^{-2})/B_i$, $[\mathbf{d}_i]_j = \mathbb{E}_q(\sigma_{ij}^{-2})$ and $\mathbf{E}_{ij}$ is a matrix of zeros except the (i,j)th entry which is 1 and has the same dimensions as $\boldsymbol{\Sigma}$.

The first derivatives of ℓ_L with respect to $A_{y,i}$, $B_{y,i}$ and ν_y are given by

$$\begin{aligned} \mathsf{D}_{A_{y,i}}\ell_L &= -\frac{\nu_y + \widetilde{y}_i \mathbb{E}_q(\sigma_{y,i}^{-2})}{2} \cdot \frac{1}{B_{y,i}} + \left(\frac{1+\nu_y}{2} - A_{y,i}\right)\psi'(A_{y,i}) + 1, \\ \mathsf{D}_{B_{y,i}}\ell_L &= \frac{\nu_y + \widetilde{y}_i \mathbb{E}_q(\sigma_{y,i}^{-2})}{2} \cdot \frac{A_{y,i}}{B_{y,i}^2} - \frac{1+\nu_y}{2} \cdot \frac{1}{B_{y,i}} \qquad (6.17) \\ \text{and } \mathsf{D}_{\nu_y}\ell_L &= \frac{n}{2}\left(\log\left(\frac{\nu_y}{2}\right) + 1 - \psi\left(\frac{\nu_y}{2}\right)\right) + \sum_{i=1}^{n} \frac{\psi(A_{y,i}) - \log(B_{y,i})}{2} - \frac{A_{y,i}}{2B_{y,i}}. \end{aligned}$$

The first derivatives of ℓ_L with respect to $\widetilde{\boldsymbol{\nu}}$, $\widetilde{\boldsymbol{\Sigma}}$ and $\widetilde{\sigma}_i^2$ are given by

$$\begin{aligned} \mathsf{D}_{\widetilde{\boldsymbol{\nu}}}\ell_L &= \widetilde{\mathbf{C}}^T(\widetilde{\mathbf{y}} \odot \widetilde{\mathbf{w}} - \mathbf{1})/2 - \widetilde{\mathbf{B}}_{\widetilde{\boldsymbol{\sigma}}^2}\widetilde{\boldsymbol{\nu}}, \\ \mathsf{D}_{\widetilde{\Sigma}_{ij}}\ell_L &= \mathrm{tr}\left(\left(\widetilde{\boldsymbol{\Sigma}}^{-1} - \widetilde{\mathbf{Z}}^T \mathrm{diag}(\widetilde{\mathbf{y}} \odot \widetilde{\mathbf{w}}/2)\widetilde{\mathbf{Z}} - \widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2}\right)\widetilde{\mathbf{E}}_{ij}\right)/2 \\ \text{and } \mathsf{D}_{\widetilde{\sigma}_i^2} &= \frac{\widetilde{\boldsymbol{\mu}}^T \widetilde{\mathbf{D}}_i \widetilde{\boldsymbol{\mu}} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{D}}_i)}{2(\widetilde{\sigma}_i^2)^2} - \frac{\widetilde{K}_i}{2\widetilde{\sigma}_i^2} \end{aligned}$$

where $\widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2} = \sum_{i=1}^{\widetilde{v}} \widetilde{\sigma}_i^{-2}\mathbf{I}_{\widetilde{K}_i}$, $\widetilde{\mathbf{B}} = \mathrm{blockdiag}(\mathbf{0}_{\widetilde{q}}, \widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2})$, $[\widetilde{\mathbf{w}}]_i = A_{y,i}\mathbb{E}_q(\sigma_{y,i}^{-2})/B_{y,i}$ and $\widetilde{\mathbf{E}}_{ij}$ is a matrix of zeros except the (i,j)th entry which is 1 and has the same dimensions as $\widetilde{\boldsymbol{\Sigma}}$.

The first derivatives of ℓ_L with respect to $\overline{\boldsymbol{\nu}}_i$, $\overline{\boldsymbol{\Sigma}}_i$ and $\overline{\sigma}_i^2$ are given by

$$\begin{aligned} \mathsf{D}_{\overline{\boldsymbol{\nu}}_i} &= \overline{\mathbf{C}}_i^T(\overline{\mathbf{y}}_i \odot \overline{\mathbf{w}}_i - \mathbf{1})/2 - \overline{\boldsymbol{\sigma}}_i^{-2}\overline{\boldsymbol{\nu}}_i, \\ \mathsf{D}_{\overline{\Sigma}_{i,jk}} &= \mathrm{tr}\left(\left(\overline{\boldsymbol{\Sigma}}_i^{-1} - \overline{\mathbf{Z}}_i^T \mathrm{diag}(\overline{\mathbf{y}}_i \odot \overline{\mathbf{w}}_i/2)\overline{\mathbf{Z}}_i - \overline{\boldsymbol{\sigma}}_i^{-2}\mathbf{I}_{\overline{K}_i}\right)\overline{\mathbf{E}}_{ijk}\right)/2 \\ \text{and } \mathsf{D}_{\overline{\sigma}_i^2} &= \frac{\|\overline{\boldsymbol{\mu}}_i\|^2 + \mathrm{tr}(\overline{\boldsymbol{\Sigma}}_i)}{2(\overline{\sigma}_i^2)^2} - \frac{\overline{K}_i}{2\overline{\sigma}_i^2} \end{aligned}$$

where $[\overline{\mathbf{w}}_i]_j = \mathbb{E}_q(\sigma_{ij}^{-2})$ and $\overline{\mathbf{E}}_{ijk}$ is a matrix of zeros except the jkth entry which is 1 and has the same dimensions as $\overline{\boldsymbol{\Sigma}}_i$. Thus we could apply Newton-Raphson updates for ν_y

which are given by

$$\nu_y := \nu_y - \left(\frac{\partial^2 \ell_L}{\partial \nu_y^2}\right)^{-1} \frac{\partial \ell_L}{\partial \nu_y}.$$

Unfortunately ν_y is subject to the implicit constraint $\nu_y > 0$ and Newton-Raphson updates may make $\nu_y \leq 0$. Instead we propose to first make the transformation $\nu = e^r$ and then use Newton-Raphson updates on r. The first derivatives of ℓ_L with respect to r are

$$\begin{aligned} \frac{\partial \ell_L}{\partial r} &= \frac{\partial \nu_y}{\partial r}\frac{\partial \ell_L}{\partial \nu_y} &&= \nu_y \frac{\partial \ell_L}{\partial \nu_y} \\ \frac{\partial^2 \ell_L}{\partial r^2} &= \frac{\partial^2 \nu_y}{\partial r^2}\frac{\partial \ell_L}{\partial \nu_y} + \left(\frac{\partial \nu_y}{\partial r}\right)^2 \frac{\partial^2 \ell_L}{\partial \nu_y^2} &&= \nu_y \frac{\partial \ell_L}{\partial \nu_y} + \nu_y^2 \frac{\partial^2 \ell_L}{\partial \nu_y^2} \end{aligned}$$

where

$$\frac{\partial^2 \ell_L}{\partial \nu_y^2} = \frac{n}{2}\left(\frac{1}{\nu_y} - \frac{1}{2}\psi'\left(\frac{\nu_y}{2}\right)\right).$$

The fixed point updates for the base model is

$$\begin{aligned} \boldsymbol{\Sigma} &:= \left(\mathbf{Z}^T \mathrm{diag}(\mathbf{w})\mathbf{Z} + \mathbf{D}\right)^{-1} \\ \text{and } \boldsymbol{\nu} &:= \left(\mathbf{C}^T \mathrm{diag}(\mathbf{w})\mathbf{C} + \mathbf{B}\right)^{-1} \mathbf{C}^T \mathrm{diag}(\mathbf{w})\mathbf{y}. \end{aligned}$$

The fixed point updates for the Student's t response parameters are

$$\begin{aligned} A_{y,i} &:= \frac{1+\nu_y}{2} \\ B_{y,i} &:= \frac{\nu_y + \widetilde{y}_i \mathbb{E}_q(\sigma_{y,i}^{-2})}{2} \\ \text{and } \nu_y &:= \nu_y \exp\left(-\left(\frac{\partial \ell_L}{\partial \nu_y} + \nu_y \frac{\partial^2 \ell_L}{\partial \nu_y^2}\right)^{-1} \frac{\partial \ell_L}{\partial \nu_y}\right). \end{aligned}$$

The fixed point updates for the variance function is

$$\begin{aligned} \widetilde{\boldsymbol{\Sigma}} &:= \left(\widetilde{\mathbf{Z}}^T \mathrm{diag}(\widetilde{\mathbf{y}} \odot \widetilde{\mathbf{w}}/2)\widetilde{\mathbf{Z}} + \widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2}\right)^{-1}. \\ \widetilde{\boldsymbol{\nu}} &:= \left(\widetilde{\mathbf{C}}^T \mathrm{diag}(\widetilde{\mathbf{y}} \odot \widetilde{\mathbf{w}}/2)\widetilde{\mathbf{C}} + \widetilde{\mathbf{B}}_{\widetilde{\boldsymbol{\sigma}}^2}\right)^{-1} \left(\widetilde{\mathbf{C}}^T \left(\widetilde{\mathbf{y}} \odot \widetilde{\mathbf{w}} - 1\right)/2 - \widetilde{\mathbf{B}}_{\widetilde{\boldsymbol{\sigma}}^2}\widetilde{\boldsymbol{\nu}}\right) \\ \text{and } \widetilde{\sigma}_i^2 &:= \frac{\widetilde{\boldsymbol{\mu}}^T \widetilde{\mathbf{D}}_i \widetilde{\boldsymbol{\mu}} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{D}}_i)}{\widetilde{K}_i}. \end{aligned}$$

Finally, the fixed point update for the adaptive variance components are

$$\begin{aligned} \overline{\boldsymbol{\Sigma}}_i &:= \left(\overline{\mathbf{Z}}_i^T \mathrm{diag}(\overline{\mathbf{y}}_i \odot \overline{\mathbf{w}}_i/2)\overline{\mathbf{Z}}_i + \overline{\sigma}_i^{-2}\mathbf{I}_{\overline{K}_i}\right)^{-1}, \\ \overline{\boldsymbol{\nu}}_i &:= \left(\overline{\mathbf{C}}_i^T \mathrm{diag}(\overline{\mathbf{y}}_i \odot \overline{\mathbf{w}}_i/2)\overline{\mathbf{C}}_i + \overline{\mathbf{B}}_i\right)^{-1} \left(\overline{\mathbf{C}}_i^T \left(\overline{\mathbf{y}}_i \odot \overline{\mathbf{w}}_i - 1\right)/2 - \overline{\mathbf{B}}_i\widetilde{\boldsymbol{\nu}}_i\right) \\ \text{and } \overline{\sigma}_i^2 &:= \frac{\|\overline{\boldsymbol{\mu}}_i\|^2 + \mathrm{tr}(\overline{\boldsymbol{\Sigma}}_i)}{\overline{K}_i} \end{aligned}$$

where $\overline{\mathbf{B}}_i = \mathrm{blockdiag}(\mathbf{0}, \overline{\sigma}_i^{-2}\mathbf{I}_{\overline{K}_i})$.

A little care is needed when applying these updates. We use the following starting points

$$\begin{aligned}
\boldsymbol{\beta} &:= \boldsymbol{\beta}^*, \\
\mathbf{u} &:= \mathbf{u}^*, \\
\boldsymbol{\Sigma} &:= (\sigma_y^{-2*}\mathbf{Z}^T\mathbf{Z} + \mathbf{D}_{\boldsymbol{\sigma}^{2*}})^{-1}, \\
\widetilde{\beta}_1 &:= \log(\sigma_y^{2*}), \\
\overline{\beta}_{i1} &:= \log(\sigma_{u,i}^{2*}) && \text{for } 1 \le i \le v, \\
\nu_y &:= 2, \\
\widetilde{\sigma}_i^2 &:= 1,000 && \text{for } 1 \le i \le \widetilde{v} \\
\text{and } \overline{\sigma}_i^2 &:= 1,000 && \text{for } 1 \le i \le v
\end{aligned} \tag{6.18}$$

where $(\boldsymbol{\beta}^*, \mathbf{u}^*, \sigma_y^{2*}, \boldsymbol{\sigma}^{2*})$ are the parameter values obtained from the solution of a LMM.

An update strategy that works is to first update the Student's t response parameters, variance function parameters (expect the $\widetilde{\sigma}_i^2$s) and the adaptive variance components parameters (expect the $\overline{\sigma}_i^2$s) until these parameters converge. We then apply the updates for Student's t response, variance function and the adaptive variance components parameters which are interleaved by updates for the base model parameters. This is the basis for Algorithm 10.

6.5.2 *Alternatives and Extensions*

It is now easy to remove "robustness" options by making some simple changes.

- For Gaussian response and random effects one might be tempted to set ν_y to a large constant, say 1000. We have found, however, that this strategy leads to incorrect results since, for example, if most $\widetilde{y}_i$s are larger than 1000 then this is not a sufficiently large constant. Instead for Gaussian response models we set $A_{y,i} = B_{y,i} = 1$, $1 \le i \le n$.
- For constant variance function we let $\delta_{\widetilde{\mathbf{u}}}(\widetilde{\mathbf{u}}) = 1$ and

$$\mathbb{E}_\delta\left(\log(\sigma_{y,i}^2)\right) = \log(\sigma_y^2) \quad \text{and} \quad \mathbb{E}_\delta\left(\sigma_{y,i}^{-2}\right) = \sigma_y^{-2}$$

  for some σ_y^2. Also let $\mathbb{E}_\delta(\log[\widetilde{\mathbf{u}}]) = 0$ and $\mathcal{H}_{\delta_{\widetilde{\mathbf{u}}}} = 0$. We also replace (6.20), (6.21) and (6.25) with

$$\sigma_y^2 := n^{-1}\sum_{i=1}^{n} A_{y,i}\widetilde{y}_i/B_{y,i} \tag{6.27}$$

- For constant variance components we let $\delta_{\overline{\mathbf{u}}}(\overline{\mathbf{u}}) = 1$ and

$$\mathbb{E}_\delta \log(\sigma_{ij}^2) = \log(\sigma_i^2) \quad \text{and} \quad \mathbb{E}_\delta\left(\sigma_{ij}^{-2}\right) = \sigma_i^{-2}$$

  for some σ_i^2. Also let $\mathbb{E}_\delta \log[\overline{\mathbf{u}}] = 0$ and $\mathcal{H}_{\delta_{\overline{\mathbf{u}}}} = 0$. We also replace (6.22), (6.23) and (6.26) with

$$\sigma_i^2 := K_i^{-1}\sum_{i=1}^{n} \overline{y}_i$$

Algorithm 10 Robust Spatially Adaptive Penalised Splines with Heteroscedastic Errors

1. Set initial values using (6.18)

2. Cycle

Apply updates

$$\begin{aligned} A_{y,i} &:= \frac{1+\nu_y}{2} \\ B_{y,i} &:= \frac{\nu_y + \widetilde{y}_i \mathbb{E}_q(\sigma_{y,i}^{-2})}{2} \\ \nu_y &:= \nu_y \exp\left(-\left(\frac{\partial \ell_L}{\partial \nu_y} + \nu_y \frac{\partial^2 \ell_L}{\partial \nu_y^2}\right)^{-1} \frac{\partial \ell_L}{\partial \nu_y}\right) \end{aligned} \tag{6.19}$$

and

$$\widetilde{\boldsymbol{\Sigma}} := \left(\widetilde{\mathbf{Z}}^T \mathrm{diag}(\widetilde{\mathbf{y}} \odot \widetilde{\mathbf{w}}/2)\widetilde{\mathbf{Z}} + \widetilde{\mathbf{D}}_{\widetilde{\boldsymbol{\sigma}}^2}\right)^{-1} \tag{6.20}$$

$$\widetilde{\boldsymbol{\nu}} := \left(\widetilde{\mathbf{C}}^T \mathrm{diag}(\widetilde{\mathbf{y}} \odot \widetilde{\mathbf{w}}/2)\widetilde{\mathbf{C}} + \widetilde{\mathbf{B}}_{\widetilde{\boldsymbol{\sigma}}^2}\right)^{-1} \left(\widetilde{\mathbf{C}}^T \left(\widetilde{\mathbf{y}} \odot \widetilde{\mathbf{w}} - 1\right)/2 - \widetilde{\mathbf{B}}_{\widetilde{\boldsymbol{\sigma}}^2}\widetilde{\boldsymbol{\nu}}\right) \tag{6.21}$$

$$\overline{\boldsymbol{\Sigma}}_i := \left(\overline{\mathbf{Z}}_i^T \mathrm{diag}(\overline{\mathbf{y}}_i \odot \overline{\mathbf{w}}_i/2)\overline{\mathbf{Z}}_i + \overline{\sigma}_i^{-2}\mathbf{I}_{\overline{K}_i}\right)^{-1} \tag{6.22}$$

$$\overline{\boldsymbol{\nu}}_i := \left(\overline{\mathbf{C}}_i^T \mathrm{diag}(\overline{\mathbf{y}}_i \odot \overline{\mathbf{w}}_i/2)\overline{\mathbf{C}}_i + \overline{\mathbf{B}}_i\right)^{-1} \left(\overline{\mathbf{C}}_i^T \left(\overline{\mathbf{y}}_i \odot \overline{\mathbf{w}}_i - 1\right)/2 - \overline{\mathbf{B}}_i\widetilde{\boldsymbol{\nu}}_i\right) \tag{6.23}$$

Until convergence.

3. Cycle

Apply updates (6.19) and then

$$\begin{aligned} \boldsymbol{\Sigma} &:= \left(\mathbf{Z}^T \mathrm{diag}(\mathbf{w})\mathbf{Z} + \mathbf{D}\right)^{-1} \\ \boldsymbol{\nu} &:= \left(\mathbf{C}^T \mathrm{diag}(\mathbf{w})\mathbf{C} + \mathbf{B}\right)^{-1} \mathbf{C}^T \mathrm{diag}(\mathbf{w})\mathbf{y} \end{aligned} \tag{6.24}$$

Apply updates (6.20-6.25), (6.24) and then

$$\widetilde{\sigma}_i^2 := \frac{\widetilde{\boldsymbol{\mu}}^T \widetilde{\mathbf{D}}_i \widetilde{\boldsymbol{\mu}} + \mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{D}}_i)}{\widetilde{K}_i} \tag{6.25}$$

Apply updates (6.22-6.26), (6.24) and then

$$\overline{\sigma}_i^2 := \frac{\|\overline{\boldsymbol{\mu}}_i\|^2 + \mathrm{tr}(\overline{\boldsymbol{\Sigma}}_i)}{\overline{K}_i} \tag{6.26}$$

Until convergence.

- Finally we could model $\mathbf{y}$ as a non-normal response and then use

$$\begin{aligned}\boldsymbol{\nu} &:= \boldsymbol{\nu} + \left(\mathbf{C}^T\mathbf{S}\mathbf{C} + \mathbf{B}\right)^{-1}\left(\mathbf{C}^T\boldsymbol{\varepsilon} - \mathbf{B}\right) \\ \boldsymbol{\Sigma} &:= \left(\mathbf{Z}^T\mathbf{W}\mathbf{Z} + \mathbf{D}\right)^{-1}\end{aligned}$$

where the values for $\mathbf{W}$, $\mathbf{S}$ and $\boldsymbol{\varepsilon}$ are available from Table 5.2.2 and as noted in Chapter 4 the update equations for the nuisance parameter ϕ are available from Table 5.2.3. We also envisage that we could, in theory, have non-constant ϕ and mimic variance function estimation for the Gaussian case but the need to do this is likely to be quite rare so we do not pursue this here.

6.5.3 *Numerical Experience*

To test the effectiveness of the above Algorithm 10 for fitting variance functions we will use the same functions and settings as (6.12). We will also use the following variance functions

$$\begin{aligned}g_5(x) &= \log(\sigma_y^2) \\ g_6(x) &= \frac{\sigma_y^2}{0.25}\log\left(\frac{r}{32} + \frac{3r}{32}x^2\right) \\ g_7(x) &= \frac{\sigma_y^2}{0.25}\left(-3.9 + 1.7\exp(\sin(5\pi x))\right)\end{aligned}$$

The xs will be equally distributed between 0 and 1. We will use thin plate splines (see Chapter 1) for these experiments with $m = 3$, $K_1 = 80$ knots for the mean function, $\widetilde{K}_1 = 20$ knots for the variance function and $\overline{K}_1 = 20$ knots for the adaptive variance components for the construction of the matrices $\mathbf{X}$, $\mathbf{Z}$, $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{Z}}$ and $(\overline{\mathbf{X}}_i, \overline{\mathbf{Z}}_i)$, $1 \leq i \leq v$. These knots are spaced using the quantities of the unique xs as per equation (6.8). Note that we standardised the xs to have zero mean and unit variance which typically improves numerical stability.

Finally we will use the following noise settings

1. Gaussian noise ($\nu_y \to \infty$),
2. Student's t noise with 2 degree of freedom ($\nu_y = 2$) and
3. Student's t noise with 4 degrees of freedom ($\nu_y = 4$).

Note that for some noise settings are significantly heteroscedastic and the Student's t noise settings contain a substantial fraction of outliers.

We will compare the variational approximation of the Robust Spatially Adaptive Penalised Splines with Heteroscedastic Errors (RSAPSHE) model with the most similar alternative method `AdaptFit`. Although is in terms of the actual model the model proposed in Crainiceanu *et al.* (2007) is closer in terms of statistical goals of RSAPSHE the fitting times are in the order of hours rather than minutes or seconds. This limits extensive comparisons between the method described here and the method described in Crainiceanu *et al.* (2007). Note that we use the same knot locations for `AdaptFit` as RSAPSHE but that `AdaptFit` uses a cubic power spline basis for univariate splines.

The mean MSE for each setting using the LMM, RSAPSHE and `AdaptFit` methods for 30 trials is summarised in Table 6.5.4. From Table 6.5.4 we see that, although there

does not seem to be a strong pattern about which method does best, RSAPSHE has the smallest MSE in most cases and often givens significantly better fits than `AdaptFit`. Finally Figure 6.5 illustrates the case with f_6, g_6 and degrees of freedom $\nu_y = 2$. Note that even though there are a substantial number of outliers for this dataset and heteroscedastic noise the RSAPSHE method does a remarkable job of approximating the true mean function.

f_i	g_j	ν_y	LMM	RSAPSHE	`AdaptFit`
5	5	2	0.02546	**0.00874**	0.01379
5	5	4	0.00960	0.00492	**0.00396**
5	5	∞	0.00586	0.00271	**0.00218**
5	6	2	0.01176	**0.00128**	0.00535
5	6	4	0.00368	**0.00103**	0.00124
5	6	∞	0.00203	0.00077	**0.00076**
5	7	2	0.05895	0.04810	**0.04405**
5	7	4	0.03662	0.02278	**0.01558**
5	7	∞	0.01904	**0.00919**	0.00974
6	5	2	0.00407	**0.00119**	0.00323
6	5	4	0.00124	**0.00104**	0.00107
6	5	∞	0.00057	0.00057	**0.00051**
6	6	2	0.00183	**0.00033**	0.00149
6	6	4	0.00040	**0.00027**	0.00031
6	6	∞	0.00023	0.00017	**0.00016**
6	7	2	0.02954	**0.00243**	0.01600
6	7	4	0.00800	**0.00253**	0.00711
6	7	∞	0.00266	**0.00146**	0.00258
7	5	2	0.01554	0.01065	**0.00783**
7	5	4	0.00298	**0.00195**	0.00247
7	5	∞	0.00150	**0.00125**	0.00127
7	6	2	0.00603	**0.00151**	0.00569
7	6	4	0.00167	**0.00111**	0.00185
7	6	∞	0.00098	**0.00085**	0.00105
7	7	2	0.07382	0.06556	**0.03004**
7	7	4	0.02554	0.01516	**0.01380**
7	7	∞	0.00707	**0.00557**	0.00585

Table 6.5.4: *Mean square errors (MSE) for linear mixed model (LMM), variational approximation of the robust spatially adaptive penalised splines with heteroscedastic errors (RSAPSHE) model and* `AdaptFit`. *Method with smallest MSE are highlighted in bold.*

6.6 Conclusion

The assumption of homoscedastic Gaussian noise is often clearly false in many real world applications. Dealing with this problem in a fast and effective way has been, thus far, an unattained goal in semiparametric regression. Variational methods are a simple class of approximations which, as we have shown in this chapter, are able to seamlessly combine a number of types of robustness. Variational approximations allows fitting the RSAPSHE model in a matter of minutes whereas the model developed by Crainiceanu *et al.* (2007), the closest model in terms of its statistical goals, fits in hours.

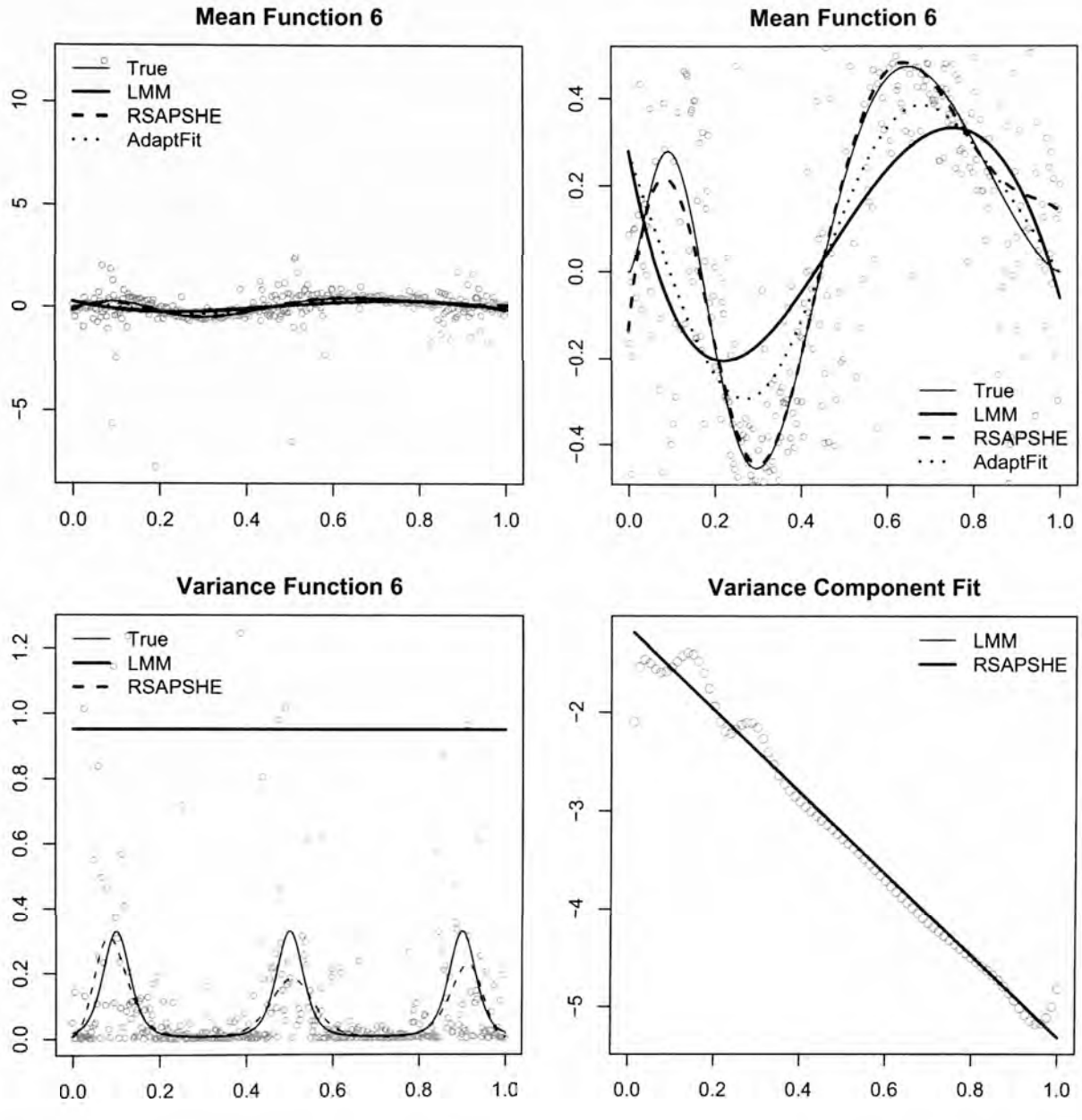


Figure 6.5: *Exemplar plots for f_6, g_6 with Student's t noise with $\nu_y = 2$ for linear mixed model (LMM), variational approximation of the robust spatially adaptive penalised splines with heteroscedastic errors (RSAPSHE) model and* `AdaptFit`. *The top left panel shows original data with fits, the top right panel shows data in the range of the fits, the bottom left panel shows the estimated variance function for LMM and RSAPSHE and the bottom right panel shows the estimated variance component function for RSAPSHE.*

Typical maximization routines using Newton-Raphson or quasi-Newton iterates are inappropriate to maximise the variational approximations of the RSAPSHE model. Alternative methods to those described here could potentially decrease fitting times.

APPENDIX A

General Probability

A.1 General Probability

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be a random vectors, i.e. vectors whose components are random variables with probability density functions $[\mathbf{x}]$ and $[\mathbf{y}]$, joint density function $[\mathbf{x}, \mathbf{y}]$ and conditional densities $[\mathbf{y}|\mathbf{x}]$ and $[\mathbf{x}|\mathbf{y}]$.

Bayes theorem is the following result, which expresses $[\mathbf{x}|\mathbf{y}]$ in terms of $[\mathbf{y}|\mathbf{x}]$:

$$[\mathbf{x}|\mathbf{y}] = \frac{[\mathbf{y}|\mathbf{x}][\mathbf{x}]}{\int [\mathbf{y}|\mathbf{x}][\mathbf{x}] d\mathbf{x}}.$$

If $\mathbf{A}$ is a constant matrix, and $\mathbf{b}$ is a constant vector whose dimensions are such that the vector $\mathbf{Ax} + \mathbf{b}$ is defined, then

$$\mathbb{E}(\mathbf{Ax} + \mathbf{b}) = \mathbf{A}\mathbb{E}(\mathbf{x}) + \mathbf{b}$$

and

$$\mathrm{Cov}(\mathbf{Ax} + \mathbf{b}) = \mathbf{A}\mathrm{Cov}(\mathbf{x})\mathbf{A}^T.$$

Finally, the mean of a quadratic form $\mathbf{x}^T\mathbf{Ax}$, is given by

$$\mathbb{E}(\mathbf{x}^T\mathbf{Ax}) = \mathbb{E}(\mathbf{x})\mathbf{A}\mathbb{E}(\mathbf{x}) + \mathrm{tr}(\mathbf{A}\mathrm{Cov}(\mathbf{x})).$$

A.2 Multivariate Gaussian Distribution

The multivariate Gaussian distribution is covered by almost all textbooks on multivariate statistics and probability theory.

Let $\mathbf{x}$ be an n-dimensional multivariate Gaussian random variable with mean $\boldsymbol{\mu}$ and (positive definite symmetric) covariance matrix $\boldsymbol{\Sigma}$, then we denote this by

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and its probability density function is given by

$$[\mathbf{x}] = \phi_{\boldsymbol{\Sigma}}(\mathbf{x} - \boldsymbol{\mu}) = \frac{1}{|2\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}.$$

The mean and covariance of $\mathbf{x}$ are $\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$ respectively.

Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ be a partitions of $\mathbf{x}$ such that

$$(\mathbf{x}_1, \mathbf{x}_2) \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right).$$

Then the marginal distributions of $\mathbf{x}_1$ and $\mathbf{x}_2$ are

$$\begin{aligned}\mathbf{x}_1 &\sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ \mathbf{x}_2 &\sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})\end{aligned}$$

and the conditional distributions are

$$\begin{aligned}\mathbf{x}_1|\mathbf{x}_2 &\sim N(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-2}\boldsymbol{\Sigma}_{21}) \\ \mathbf{x}_2|\mathbf{x}_1 &\sim N(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-2}\boldsymbol{\Sigma}_{12}).\end{aligned}$$

Finally, let $\mathbf{A}$ be a constant matrix, and $\mathbf{b}$ be a constant vector whose dimensions are such that the vector $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$ is defined. Then

$$\mathbf{y} \sim N(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T).$$

A.2.1 Multivariate Gaussian Expectations

There are many expectations results for the multivariate Gaussian distribution. Some of the most common of these are

$$\begin{aligned}\mathbb{E}\{\mathbf{x}^T\mathbf{A}\mathbf{x}\} &= \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{A}) && \text{(Quadratic Expectation)} \\ \mathbb{E}\{\exp(\mathbf{x}^T\mathbf{t})\} &= \exp\left(\boldsymbol{\mu}^T\mathbf{t} + \mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}/2\right) && \text{(Moment Generating Function)} \\ -\mathbb{E}\{\log[\mathbf{x}]\} &= \frac{1}{2}\log|2e\pi\boldsymbol{\Sigma}| && \text{(Entopy)}\end{aligned}$$

assuming all of the vectors and matrices are appropriately sized. We remind the reader that $[\mathbf{x}]$ denotes the probability density function for the vector $\mathbf{x}$.

A.2.2 Other Results

This result, appearing in Wand & Jones (1993,1995), can be useful in simplifying multivariate Gaussian expectations:

$$\phi_{\boldsymbol{\Sigma}}(\mathbf{x} - \boldsymbol{\mu})\phi_{\boldsymbol{\Sigma}'}(\mathbf{x} - \boldsymbol{\mu}') = \phi_{\boldsymbol{\Sigma}+\boldsymbol{\Sigma}'}(\boldsymbol{\mu} - \boldsymbol{\mu}')\phi_{\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\boldsymbol{\Sigma}')^{-1}\boldsymbol{\Sigma}'}(\mathbf{x} - \boldsymbol{\mu}^*)$$

where $\boldsymbol{\mu}^* = \boldsymbol{\Sigma}'(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}')^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{\Sigma}')^{-1}\boldsymbol{\mu}'$ assuming all of the vectors and matrices are appropriately sized. Hence

$$\int \phi_{\boldsymbol{\Sigma}}(\mathbf{x} - \boldsymbol{\mu})\phi_{\boldsymbol{\Sigma}'}(\mathbf{x} - \boldsymbol{\mu}')d\mathbf{x} = \phi_{\boldsymbol{\Sigma}+\boldsymbol{\Sigma}'}(\boldsymbol{\mu} - \boldsymbol{\mu}').$$

A.3 Uniform Distribution

Let x be a uniform random variable with upper and lower bounds a and b respectively. Then we denote this by

$$x \sim \text{Unif.}(a, b)$$

and its probability density function is given by

$$[x] = \frac{1}{b-a}, \text{for } x \in [a, b].$$

The mean and covariance of x are $\mathbb{E}(x) = \frac{a+b}{2}$ and $\text{Cov}(x) = \frac{(b-a)^2}{12}$ respectively.

A.4 Gamma Distribution

Let x be a Gamma random variable with shape $\alpha > 0$ and rate (or inverse-scale) $\beta > 0$. Then we denote this by

$$x \sim \text{Gamma}(\alpha, \beta)$$

and its probability density function is given by

$$[x] = g(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x),$$

for $x > 0$. The mean and covariance of x are $\mathbb{E}(x) = \frac{\alpha}{\beta}$ and $\text{Cov}(x) = \frac{\alpha}{\beta^2}$ respectively.

A.4.1 Gamma Expectations

$$\begin{aligned}
\mathbb{E}(\exp(xt)) &= (1 - t/\beta)^{-\alpha}, \text{ for } t < \beta && \text{(Moment Generating Function)} \\
-\mathbb{E}(\log[x]) &= \alpha - \log(\beta) + \log\Gamma(\alpha) - (\alpha - 1)\psi(\alpha) && \text{(Entopy)} \\
\mathbb{E}(x^{-1}) &= \frac{\beta}{\alpha - 1} \\
\mathbb{E}(\log(x)) &= \psi(\alpha) - \log(\beta)
\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function, $\psi(x) = d\log\Gamma(x)/dx$ is the digamma function (see Abramowitz & Stegun, 1964, Chapter 6). The last integral can be verified using integration by parts. The $\Gamma(\cdot)$ function has the properties

$$\begin{aligned}
\Gamma(x + 1) &= x\Gamma(x) \\
\Gamma(1/2) &= \sqrt{\pi}.
\end{aligned}$$

A.5 Inverse-Gamma Distribution

Let x be a inverse-gamma random variable with shape $\alpha > 0$ and scale $\beta > 0$, then we denote this by

$$x \sim IG(\alpha, \beta)$$

and its probability density function is given by

$$[x] = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp(-\beta/x), \; x > 0.$$

As implied by the name the inverse-gamma random variable arises by considering the reciprocal of the gamma random variable, i.e. if $x \sim \text{Gamma}(\alpha, \beta)$ then $1/x \sim \text{IG}(\alpha, \beta)$. The mean and covariance of x are $\mathbb{E}(x) = \frac{\beta}{\alpha-1}$ and $\text{Cov}(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ respectively.

A.5.1 Inverse-Gamma Expectations

$$\begin{aligned}
-\mathbb{E}(\log[x]) &= \alpha + \log(\beta) + \log\Gamma(\alpha) - (\alpha + 1)\psi(\alpha) && \text{(Entopy)} \\
\mathbb{E}(x^{-1}) &= \frac{\alpha}{\beta} \\
\mathbb{E}(\log(x)) &= \log(\beta) - \psi(\alpha)
\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function, $\psi(\cdot)$ is the digamma function (see Abramowitz & Stegun, 1964, Chapter 6). The last integral can be verified using integration by parts.

A.6 Beta Distribution

Let x be a beta random variable with shape parameters $\alpha, \beta > 0$, then we denote this by

$$x \sim \text{Beta}(\alpha, \beta)$$

and its probability density function is given by

$$[x] = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \; 0 \leq x \leq 1$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. The mean and covariance of x are $\mathbb{E}(x) = \frac{\alpha}{\alpha+\beta}$ and $\text{Cov}(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ respectively. The special case $x \sim \text{Beta}(1,1)$ is equivalent to the uniform distribution.

A.6.1 Beta Expectations

$$\begin{aligned} -\mathbb{E}(\log[x]) &= \log B(\alpha, \beta) - (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta) \\ &\quad +(\alpha+\beta-2)\psi(\alpha+\beta) \qquad \text{(Entopy)} \\ \mathbb{E}(\log(x)) &= \psi(\alpha) - \psi(\alpha+\beta) \\ \mathbb{E}(\log(1-x)) &= \psi(\beta) - \psi(\alpha+\beta) \end{aligned}$$

where $\psi(\cdot)$ is the digamma function (see Abramowitz & Stegun, 1964, Chapter 6). The last two integrals can be verified using integration by parts.

A.7 Student's t-Distribution

Let x be a univariate Student's t random variable with degrees of freedom parameter ν:

$$x \sim t(\nu).$$

The probability density function for the univariate Student's t-distribution is

$$[x] = \mathcal{S}(x; \nu) = \frac{\Gamma\left(\frac{1+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{\frac{1}{2}}} \left[1 + \frac{x^2}{\nu}\right]^{-\frac{1+\nu}{2}}$$

As $\nu \to \infty$ the univariate Student's t-distribution approaches the univariate standard Gaussian distribution, i.e. as $\nu \to \infty$, $\mathcal{S}(x; \nu) \to \phi_1(x)$.

A possible scale-location extension for the univariate Student's t-distribution is

$$[x] = \mathcal{S}(x; \mu, \sigma^2, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}} \left[1 + \frac{(x-\mu)^2}{\sigma^2\nu}\right]^{-\frac{1+\nu}{2}}$$

and has mean $\mathbb{E}(x) = \mu$ (for $\nu > 1$) and covariance $\text{Cov}(x) = \frac{\nu}{\nu-2}\sigma^2$ (for $\nu > 2$). For the special case $\nu = 1$ the random variable x becomes a univariate Cauchy random variable.

A possible, but not the only (see Kotz, Balakrishnan & Johnson, 2000, for example), multivariate scale-location extension for the Student's t-distribution is

$$\mathcal{S}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{n+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)|\pi\nu\boldsymbol{\Sigma}|^{\frac{1}{2}}} \left[1 + \frac{(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{n+\nu}{2}}$$

and has mean $\mathbb{E}(\mathbf{y}) = \boldsymbol{\mu}$ (for $\nu > 1$) and covariance $\text{Cov}(\mathbf{y}) = \frac{\nu}{\nu-2}\boldsymbol{\Sigma}$ (for $\nu > 2$).

The expression for the entropy $\mathcal{H}$ for this extension is not obvious but was first derived by Guerrero-Cusumano (1996) and is given by

$$\mathcal{H} = \log\left[\frac{|\nu\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{n+\nu}{2}\right)}\right] + \frac{n+\nu}{2}\left[\psi\left(\frac{n+\nu}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right].$$

for $\nu > 2$.

APPENDIX B
Matrix Algebra

Matrix results play a dominant role in this thesis. This appendix contains reference formula for matrix computations used throughout this thesis. Other references include Magnus & Neudecker (1988) and Harville (1997). A standard reference for matrix analysis is Golub & van Loan (1996).

B.1 Some Matrix Algebra Rules

$$\begin{aligned}
\operatorname{tr}(\mathbf{AB}) &= \operatorname{tr}(\mathbf{BA}) && \mathbf{A} \in \mathbb{R}^{n\times m}, \mathbf{B} \in \mathbb{R}^{m\times n} \\
|\mathbf{AB}| &= |\mathbf{BA}| && \mathbf{A} \in \mathbb{R}^{n\times m}, \mathbf{B} \in \mathbb{R}^{m\times n} \\
|c\mathbf{A}| &= c^n|\mathbf{A}| && c \in \mathbb{R}, \mathbf{A} \in \mathbb{R}^{n\times n} \\
|\mathbf{A}^c| &= |\mathbf{A}|^c && c \in \mathbb{R}, \mathbf{A} \in \mathbb{R}^{n\times n} \\
\log|\mathbf{I}+\mathbf{AB}| &= \log|\mathbf{I}+\mathbf{BA}| && \mathbf{A} \in \mathbb{R}^{n\times m}, \mathbf{B} \in \mathbb{R}^{m\times n}
\end{aligned}$$

B.2 Matrix Calculus

B.2.1 Derivatives of Linear Operators

$$\begin{aligned}
\frac{\partial(\mathbf{A}+\mathbf{B})}{\partial x} &= \left(\frac{\partial \mathbf{A}}{\partial x}\right) + \left(\frac{\partial \mathbf{B}}{\partial x}\right) \\
\frac{\partial \operatorname{diag}(\mathbf{a})}{\partial x} &= \operatorname{diag}\left(\frac{\partial \mathbf{a}}{\partial x}\right) \\
\frac{d\operatorname{tr}(\mathbf{A})}{\partial x} &= \operatorname{tr}\left(\frac{\partial \mathbf{A}}{\partial x}\right)
\end{aligned}$$

B.2.2 Product and Quotient Rules

$$\begin{aligned}
\frac{\partial(\mathbf{A}\odot\mathbf{B})}{\partial x} &= \left(\frac{\partial \mathbf{A}}{\partial x}\right)\odot\mathbf{B} + \mathbf{A}\odot\left(\frac{\partial \mathbf{B}}{\partial x}\right) \\
\frac{\partial \mathbf{AB}}{\partial x} &= \left(\frac{\partial \mathbf{A}}{\partial x}\right)\mathbf{B} + \mathbf{A}\left(\frac{\partial \mathbf{B}}{\partial x}\right) \\
\frac{\partial \mathbf{a}/\mathbf{b}}{\partial x} &= \left(\left(\frac{\partial \mathbf{a}}{\partial x}\right)\odot\mathbf{b} - \mathbf{b}\odot\left(\frac{\partial \mathbf{a}}{\partial x}\right)\right)\Big/(\mathbf{b}\odot\mathbf{b})
\end{aligned}$$

B.2.3 Rules for Determinants and Inverses

$$\begin{aligned}
\frac{\partial|\mathbf{A}|}{\partial x} &= |\mathbf{A}|\operatorname{tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right) \\
\frac{\partial \mathbf{A}^{-1}}{\partial x} &= -\mathbf{A}^{-1}\left(\frac{\partial \mathbf{A}}{\partial x}\right)\mathbf{A}^{-1} \\
\frac{\partial \log|\mathbf{A}|}{\partial x} &= \operatorname{tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right)
\end{aligned}$$

B.3 Special Matrix Formulae

B.3.1 Inverse Identities

Let $\mathbf{A}$ and $\mathbf{B}$ be non-singular square $m \times m$ matrices. The inverse of the product of the two matrices can be written in terms of the individual inverses

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.$$

For the sum of two matrices the following identities are valid:

$$\begin{aligned} \mathbf{A}^{-1} + \mathbf{B}^{-1} &= \mathbf{A}^{-1}(\mathbf{A} + \mathbf{B})\mathbf{B}^{-1} \\ (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} &= \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B} \\ &= \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A}. \end{aligned}$$

B.3.2 Sherman-Morrison-Woodbury Inversion Formula

$$\left(\mathbf{\Lambda} + \mathbf{UDV}^T\right)^{-1} = \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\mathbf{U}\left(\mathbf{D}^{-1} + \mathbf{V}^T\mathbf{\Lambda}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}^T\mathbf{\Lambda}^{-1} \tag{B.1}$$

assuming the inverses matrices $\mathbf{\Lambda}^{-1}$ and $\left(\mathbf{D}^{-1} + \mathbf{V}^T\mathbf{\Lambda}^{-1}\mathbf{U}\right)^{-1}$ above exist. This formula is more efficient than straight inversion when either $\mathbf{\Lambda}^{-1}$ is known or easy to calculate.

B.3.3 Partitioned Matrix Inversion Formulae

$$\begin{aligned} \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} &= \begin{bmatrix} (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{C} & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{BD}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\mathbf{A}} & \tilde{\mathbf{B}} \\ \tilde{\mathbf{C}} & \tilde{\mathbf{D}} \end{bmatrix} \end{aligned} \tag{B.2}$$

where

$$\begin{aligned} \tilde{\mathbf{A}} &= (\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}, \\ \tilde{\mathbf{B}} &= -\tilde{\mathbf{A}}\mathbf{BD}^{-1}, \\ \tilde{\mathbf{C}} &= -\mathbf{D}^{-1}\mathbf{C}\tilde{\mathbf{A}} \\ \text{and } \tilde{\mathbf{D}} &= \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\tilde{\mathbf{A}}\mathbf{BD}^{-1}, \end{aligned} \tag{B.3}$$

assuming the inverse matrices $\mathbf{D}^{-1}$ and $(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}$ above exist. This formula is more efficient than straight inversion when both $(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^{-1}$ and $\mathbf{D}^{-1}$ are known or easy to calculate.

B.3.4 Partitioned Matrix Determinant Formula

$$\left|\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}\right| = |\mathbf{A}||\mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}| \tag{B.4}$$

This formula can be used to simplify determinants when $\mathbf{A}^{-1}$ is known or is easy to calculate.

APPENDIX C
Multivariate Optimisation

Multivariate optimisation plays a dominant role in Statistics via the concept of maximum likelihood. This appendix is short primer for some of the optimisation concepts used in this thesis. We will first describe some of background material used in this thesis concerning optimisation. The stated results, or variants of these results, can be found in the optimisation texts by Dennis & Schnabel (1983), Nocedal & Wright (1999) and Ruszczyński (2006) amongst others.

We also explore an inexact-Newton method in Section C.3.3. There we derive results concerning a not uncommon modification of the Newton-Raphson method which, to the best of our knowledge, are *new*. The modification involves only calculating Hessian matrix only every r iterations. We show that the rate of convergence over $r = 2$ iterations, under appropriate conditions, is cubic and that these iterations have the same asymptotic computational cost as the Newton-Raphson method. Since, in many situations, the computational cost of calculating the Hessian is high this can result in significant computational improvements over the Newton-Raphson method. We call this the *repeated-Hessian* Newton's method.

C.1 Definitions

- In unconstrained optimisation we seek to minimise (or maximise) a function $f : \mathbb{R}^n \to \mathbb{R}$, the *objective function*, with respect to variables $\mathbf{x} \in \mathbb{R}^n$, called *decision variables*, with no restrictions on the values these decision variables take, i.e.

$$\min_{\mathbf{x}} f(\mathbf{x}). \tag{C.1}$$

- A point $\mathbf{x}_*$ is a *global minimiser* if $f(\mathbf{x}_*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$.
- A point $\mathbf{x}_*$ is a *local minimiser* if there is a neighbourhood $\mathcal{N}$ of $\mathbf{x}_*$ such that $f(\mathbf{x}_*) \leq f(\mathbf{x})$ for $\mathbf{x} \in \mathcal{N}$.
- A point $\mathbf{x}_*$ is a *strict local minimiser* if there is a neighbourhood $\mathcal{N}$ of $\mathbf{x}_*$ such that $f(\mathbf{x}_*) < f(\mathbf{x})$ for $\mathbf{x} \in \mathcal{N}$ with $\mathbf{x} \neq \mathbf{x}_*$.
- A function f is *convex* if and only if for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and for all $0 \leq \alpha \leq 1$ we have

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

- Let $\{\mathbf{x}_k\}$ be a sequence in $\mathbb{R}^n$ that converges to $\mathbf{x}_*$. We say that the convergence is *Q-linear* if there is a constant $r \in (0,1)$ such that

$$\frac{\|\mathbf{x}_{k+1} - \mathbf{x}_*\|}{\|\mathbf{x}_k - \mathbf{x}_*\|} \leq r, \qquad \text{for all } k \text{ sufficiently large.}$$

We say that the Q-order of convergence is p (with $p > 1$) if there is a positive constant M such that

$$\frac{\|\mathbf{x}_{k+1} - \mathbf{x}_*\|}{\|\mathbf{x}_k - \mathbf{x}_*\|^p} \leq M, \qquad \text{for all } k \text{ sufficiently large.}$$

and the Q-order of convergence is super-p (with $p \geq 1$) if

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_*\|}{\|\mathbf{x}_k - \mathbf{x}_*\|^p} = 0.$$

In particular if the Q-order of convergence is super-1 we call this simply *superliner* convergence.

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. A *stationary point* of f, $\mathbf{x}_*$ satisfies $\mathsf{D}_{\mathbf{x}} f(\mathbf{x}_*) = \mathbf{0}$.

C.2 Optimality Conditions

At different points in this thesis we refer to first order or second order optimality conditions. This terminology refers to the following theorems. For brevity we will now adopt the following notation. If $\mathbf{x}_k \in \mathbb{R}^n$ then

$$f_k = f(\mathbf{x}_k), \qquad \mathbf{g}_k = \mathsf{D}_{\mathbf{x}} f(\mathbf{x}_k) \qquad \text{and} \qquad \mathbf{H}_k = \mathsf{H}_{\mathbf{x}} f(\mathbf{x}_k).$$

and use similarly notation for $\mathbf{x}_* \in \mathbb{R}^n$.

Theorem C.1 [First-Order Necessary Conditions, e.g. Nocedal & Wright, 1999, Theorem 2.2]: *If $\mathbf{x}_*$ is a local minimiser and f is continuously differentiable in an open neighbourhood of $\mathbf{x}_*$, then $\mathbf{g}_* = \mathbf{0}$.*

Theorem C.2 [Second-Order Necessary Conditions, e.g. Nocedal & Wright, 1999, Theorem 2.3]: *If $\mathbf{x}_*$ is a local minimiser and f and $\mathsf{H}_{\mathbf{x}} f(\mathbf{x})$ is continuous in an open neighbourhood of $\mathbf{x}_*$, then $\mathbf{g}_* = \mathbf{0}$ and $\mathbf{H}_*$ is positive semidefinite.*

Theorem C.3 [Second-Order Sufficient Conditions, e.g. Nocedal & Wright, 1999, Theorem 2.4]: *Suppose that $\mathsf{H}_{\mathbf{x}} f(\mathbf{x})$ is continuous in an open neighbourhood of $\mathbf{x}_*$ and that $\mathbf{g}_*$ and $\mathbf{H}_*$ is positive definite. Then $\mathbf{x}_*$ is a strict local minimiser of f.*

Theorem C.4 [e.g. Nocedal & Wright, 1999, Theorem 2.5]: *When f is convex, any local minimiser $\mathbf{x}_*$ is a global minimiser of f. If, in addition, f is differentiable, then any stationary point $\mathbf{x}_*$ is a global minimiser of f.*

C.3 Optimisation Methods

The above theorems concern the characterisation of minimisers of (C.1). There are two main classes of methods which may be used to find local minimisers of (C.1). These are called *line search* methods and *trust region* methods.

Line search methods solve a sequence of one dimensional minimisation problems of the form

$$\alpha_k = \underset{\alpha>0}{\operatorname{argmin}} \{\phi(\alpha) = f(\mathbf{x}_k + \alpha \mathbf{p}_k)\} \tag{C.2}$$

for some $\mathbf{p}_k \in \mathbb{R}^n$ is a *descent direction*, i.e.

$$\mathbf{p}_k^T \mathbf{g}_k < 0.$$

The vector $\mathbf{p}_k$ may chosen using the Newton-Raphson method, quasi-Newton method, or a variety of other methods. In practice the step length α is restricted using some conditions, for example Wolfe conditions or Goldstein conditions, to ensure convergence to a local minimise of f. The optimisation of (C.2) is typically performed using polynomial interpolation methods. Once a suitable step length α_k is found we apply the update $\mathbf{x}_k + \alpha_k \mathbf{p}_k$. See Dennis & Schnabel (1983, Chapter 6) or Nocedal & Wright (1999, Chapter 3) for details.

The basis idea behind trust region methods is to approximate f with a simpler function, say $\widehat{f}$, which reasonably reflects the shape of the function f in a neighbourhood $\mathcal{N}$, called the trust region, around the current point $\mathbf{x}_k$. A trial step $\mathbf{x}_{k+1}$ is computed by approximately minimizing $\widehat{f}$ over the region $\mathcal{N}$. Let

$$\mathbf{s}_k = \underset{\mathbf{s}\in\mathcal{N}}{\operatorname{argmin}}\, \widehat{f}(\mathbf{s}). \tag{C.3}$$

If $f(\mathbf{x}_k + \mathbf{s}_k) < f(\mathbf{x}_k)$ we assign $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$, otherwise the trust region $\mathcal{N}$ is shrunk and (C.3) is recomputed.

Almost all trust region methods use a second order Taylor series expansion of f around $\mathbf{x}_k$ for $\widehat{f}$ and the neighbourhood $\mathcal{N}$ is spherical or ellipsoidal in shape. The trust region subproblem (C.3) is then typically stated as

$$\mathbf{s}_k = \underset{\mathbf{s}}{\operatorname{argmin}} \left\{ \frac{1}{2}\mathbf{s}_k^T \mathbf{H}_k \mathbf{s}_k + \mathbf{s}_k^T \mathbf{g}_k \text{ such that } \|\mathbf{D}\mathbf{s}_k\| \leq \triangle \right\}, \tag{C.4}$$

where $\mathbf{D}$ is a diagonal scaling matrix and $\triangle$ is a positive scalar. Many good algorithms exist for solving (C.4). See for example Moré & Sorensen (1983), Dennis & Schnabel (1983), Celis, Dennis & Tapia (1994), Byrd, Schnabel & Schultz (1994), Nocedal & Wright (1999, Chapters 4 and 6) and Ruszczyński (2006). Note that (C.4) may also be modified by using an approximate $\mathcal{H}_k$ to produce different search directions, for example quasi-Newton directions (see Nocedal & Wright, 1999, Chapters 4 and 6 for details).

C.3.1 *Newton-Raphson Method*

Consider the sequence $\{\mathbf{x}_k\}$ defined by

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k \tag{C.5}$$

where $\mathbf{p}_k$ is the Newton-Raphson search direction is given by

$$\mathbf{p}_k = -\mathbf{H}_k^{-1}\mathbf{g}_k. \tag{C.6}$$

The sequence $\{\mathbf{x}_k\}$ generated by (C.5) and (C.6) are the Newton-Raphson iterates. Note that without using a line search or trust region method Newton-Raphson iterates are not guaranteed to converge (see Dennis & Schnabel, 1983, Chapter 6 for examples).

We will now prove that if $\mathbf{x}_k$ is sufficiently close to a strict local minimiser $\mathbf{x}_*$ then the rate of convergence of the Newton-Raphson iterates is quadratic. The proof we present is identical to Theorem 3.7. of Nocedal & Wright, (1999) except that we break this theorem into several lemmas which we will later use to prove convergence properties of the repeated-Hessian Newton's method.

The quadratic convergence of Newton-Raphson iterates may be proved with the help of some additional assumptions and the use of Taylor's Theorem (stated below).

Assumptions C.5: *Let $\mathbf{x}_*$ be a strict local minimiser of f, i.e. $\mathsf{H}_\mathbf{x} f(\mathbf{x})$ is continuous in an open neighbourhood of $\mathbf{x}_*$ and that $\mathbf{g}_* = \mathbf{0}$ and $\mathbf{H}_*$ is positive definite. Assume*

1. *that f is twice differentiable;*
2. *the gradient $\mathsf{D}_\mathbf{x} f(\mathbf{x})$ and the Hessian $\mathsf{H}_\mathbf{x} f(\mathbf{x})$ are Lipschitz continuous with constants $L_1 > 0$ and $L_2 > 0$ respectively in a neighbourhood of $\mathbf{x}_*$, i.e.*

$$\begin{aligned} \|\mathsf{D}_\mathbf{x} f(\mathbf{x}) - \mathsf{D}_\mathbf{x} f(\mathbf{y})\| &\leq L_1\|\mathbf{x} - \mathbf{y}\| \\ \|\mathsf{H}_\mathbf{x} f(\mathbf{x}) - \mathsf{H}_\mathbf{x} f(\mathbf{y})\| &\leq L_2\|\mathbf{x} - \mathbf{y}\| \end{aligned} \tag{C.7}$$

   for all $\mathbf{x}$ and $\mathbf{y}$ in a neighbourhood of $\mathbf{x}_*$.
3. The sequence $\{\mathbf{x}_k\}$ defined by (C.5) and (C.6) converges to $\mathbf{x}_*$.

Theorem C.6 [Taylor's Theorem, e.g. Nocedal & Wright, 1999, Theorem 2.1]: *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and that $\mathbf{p} \in \mathbb{R}^n$. Then we have that*

$$f(\mathbf{x} + \mathbf{p}) = f(\mathbf{x}) + (\mathsf{D}_\mathbf{x} f(\mathbf{x} + t\mathbf{p}))^T \mathbf{p}, \tag{C.8}$$

for some $t \in (0, 1)$. Moreover, if f is twice continuously differentiable, we have that

$$\mathsf{D}_\mathbf{x} f(\mathbf{x} + \mathbf{p}) = \mathsf{D}_\mathbf{x} f(\mathbf{x}) + \int_0^1 \mathsf{H}_\mathbf{x} f(\mathbf{x} + t\mathbf{p})\mathbf{p}\, dt, \tag{C.9}$$

and that

$$f(\mathbf{x} + \mathbf{p}) = f(\mathbf{x}) + (\mathsf{D}_\mathbf{x} f(\mathbf{x}))^T \mathbf{p} + \frac{1}{2}\mathbf{p}^T[\mathsf{H}_\mathbf{x} f(\mathbf{x} + t\mathbf{p})]\mathbf{p}, \tag{C.10}$$

for some $t \in (0, 1)$.

Lemma C.6 [Nocedal & Wright (1999), part of Theorem 3.7]: *Suppose that Assumptions C.5 hold. Consider the sequence of iterates* $\{\mathbf{x}_k\}$ *generated by (C.5) and (C.6). Then*

$$\|\mathbf{g}_{k+1}\| \leq \frac{L_2}{2}\|\mathbf{H}_k^{-1}\|^2\|\mathbf{g}_k\|^2 \tag{C.11}$$

Proof: Using the relations $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$ and $\mathbf{g}_{k+1} + \mathbf{H}_k\mathbf{p}_k = \mathbf{0}$ we have

$$\begin{aligned}
\|\mathbf{g}_{k+1}\| &= \|\mathbf{g}_{k+1} - \mathbf{g}_k - \mathbf{H}_k\mathbf{p}_k\| \\
&= \left\| \int_0^1 [\mathsf{H}_{\mathbf{x}} f(\mathbf{x}_k + t\mathbf{p}_k)\mathbf{p}_k]dt - \mathbf{H}_k\mathbf{p}_k \right\| \\
&\leq \int_0^1 \|\mathsf{H}_{\mathbf{x}} f(\mathbf{x}_k + t\mathbf{p}_k) - \mathbf{H}_k\|\|\mathbf{p}_k\|dt \\
&\leq L_2\|\mathbf{p}_k\|^2/2
\end{aligned}$$

The second line follows from (C.9), the third follows from properties of vector norms and the forth line follows from the fact that the Hessian is Lipschitz continuous with constant L_2. Finally the lemma follows by applying the properties of vector norms to (C.6).

■

Lemma C.7 [Nocedal & Wright (1999), part of Theorem 3.7]: *Suppose that Assumptions C.5 hold. Consider the sequence of iterates* $\{\mathbf{x}_k\}$ *generated by (C.5) and (C.6). Then*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\| \leq \frac{L_2}{2}\|\mathbf{H}_k^{-1}\|^2\|\mathbf{x}_k - \mathbf{x}_*\|^2 \tag{C.12}$$

Proof: Again, using the relations $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$ and $\mathbf{g}_k + \mathbf{H}_k\mathbf{p}_k = \mathbf{0}$ we have

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_*\| &= \|\mathbf{x}_k - \mathbf{x}_* - \mathbf{H}_k^{-1}\mathbf{g}_k\| \\
&= \|\mathbf{H}_k^{-1}(\mathbf{H}_k(\mathbf{x}_k - \mathbf{x}_*) - (\mathbf{g}_k - \mathbf{g}_*))\| \\
&\leq \|\mathbf{H}_k^{-1}\| \left\| \int_0^1 [\mathbf{H}_k - \mathbf{H}(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k))](\mathbf{x}_k - \mathbf{x}_*)dt \right\| \\
&\leq \|\mathbf{H}_k^{-1}\| \int_0^1 \|[\mathbf{H}_k - \mathbf{H}(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k))]\| \, \|\mathbf{x}_k - \mathbf{x}_*\|dt
\end{aligned}$$

Again, the third line follows from (C.9) in Taylor's Theorem, the forth line follows from properties of vector norms and the lemma follows from the fact that the Hessian is Lipschitz continuous with constant L_2.

■

Assumption C.8: Assume the sequence of iterates $\{\mathbf{x}_k\}$ generated by (C.5) and (C.6) converges to a strict local minima $\mathbf{x}_*$.

Theorem C.9: *Suppose that Assumptions C.5 and C.8 hold. If*

$$\|\mathbf{g}_k\| < \frac{1}{2L_2\|\mathbf{H}_*^{-1}\|^2} \quad \text{and} \quad \|\mathbf{x}_k - \mathbf{x}_*\| < \frac{1}{2L_2\|\mathbf{H}_*^{-1}\|^2} \tag{C.13}$$

for some $k > k_0$ *then*

1. *the rate of convergence of* $\{\mathbf{g}_k\}$ *is quadratic* $k > k_0$*;*
2. *the rate of convergence of* $\{\mathbf{x}_k - \mathbf{x}_*\}$ *is quadratic for* $k > k_0$*.*

Proof: Using Assumption C.8, since $\{\mathbf{x}_k\}$ converges to $\mathbf{x}_*$, $\mathbf{H}_*$ is nonsingular, and $\mathbf{H}_k \rightarrow \mathbf{H}_*$, then for all sufficiently large k we have

$$\|\mathbf{H}_k^{-1}\| \leq 2\|\mathbf{H}_*^{-1}\|. \tag{C.14}$$

Applying this inequality and the inequalities in Lemma C.6 and Lemma C.7 we have

$$\begin{aligned} \|\mathbf{g}_{k+1}\| &\leq 2L_2\|\mathbf{H}_*^{-1}\|^2\|\mathbf{g}_k\|^2 \\ \|\mathbf{x}_{k+1} - \mathbf{x}_*\| &\leq 2L_2\|\mathbf{H}_*^{-1}\|^2\|(\mathbf{x}_k - \mathbf{x}_*)\|^2. \end{aligned} \tag{C.15}$$

Using the above inequalities recursively

$$\begin{aligned} \|\mathbf{g}_{k+k'}\| &\leq (2L_2\|\mathbf{H}_*^{-1}\|^2)^{-1}(2L_2\|\mathbf{H}_*^{-1}\|^2\|\mathbf{g}_k\|)^{2^{k'}} \\ \|\mathbf{x}_{k+k'} - \mathbf{x}_*\| &\leq (2L_2\|\mathbf{H}_*^{-1}\|^2)^{-1}(2L_2\|\mathbf{H}_*^{-1}\|^2\|(\mathbf{x}_k - \mathbf{x}_*)\|)^{2^{k'}}. \end{aligned} \tag{C.16}$$

The right hand sides of (C.16) approach 0 if the conditions (C.13) stated in the theorem are satisfied. Furthermore, for a sufficiently large $k > k_0$

$$\begin{aligned} \frac{\|\mathbf{g}_{k+1}\|}{\|\mathbf{g}_k\|^2} &\leq 2L_2\|\mathbf{H}_*^{-1}\|^2, \\ \frac{\|\mathbf{x}_{k+1} - \mathbf{x}_*\|}{\|(\mathbf{x}_k - \mathbf{x}_*)\|^2} &\leq 2L_2\|\mathbf{H}_*^{-1}\|^2 \end{aligned} \tag{C.17}$$

for all $k > k_0$ so that the rate of convergence of $\{\mathbf{g}_k\}$ and $\{\mathbf{x}_k - \mathbf{x}_*\}$ is quadratic.

■

Note that under different conditions for convex problems the Newton-Raphson method may converge linearly when the current point $\mathbf{x}_k$ is far from a local minimiser $\mathbf{x}_*$ (Boyd & Vandenberghe, 2004, Section 9.5).

While Newton-Raphson iterates are often favoured in practice, due to their quadratic convergence properties, there are a number of drawbacks to Newton-Raphson iterates including:

1. Newton-Raphson iterates are not globally convergent.
2. The Hessian matrix $\mathbf{H}_k$ is often expensive to calculate or store.
3. At each iteration the solution to a system of linear equations involving a matrix which may be singular or ill-conditioned is required.

While 1. may be handled by using a line search or trust region approach and 3. may be handled using a variety of modifications (see Nocedal & Wright, 1999, Chapter 6 for examples), point 2. can mean that alternative methods can perform better in practice.

C.3.2 Quasi-Newton Methods

Quasi-Newton methods are a class of inexact Newton methods which, instead of calculating the Hessian matrix $\mathbf{H}_k$ at each iteration, use an approximate Hessian $\widehat{\mathbf{H}}_k$ in its place. There are several ways of doing this which only require the derivatives $\mathbf{g}_k$ at each iteration. The most popular of these is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method.

Let $\widehat{\mathbf{H}}_k$ be the current approximation of the Hessian at the current value of $\mathbf{x}_k$. Then sequence of steps taken in the BFGS method are as follows

1. Obtain $\mathbf{p}_k$ by solving $\mathbf{B}_k\mathbf{p}_k = -\mathbf{g}_k$.
2. Perform a line search to find the optimal α_k in the direction found in 1., then perform the update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{p}_k$.
3. Let $\mathbf{y}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$.
4. $\widehat{\mathbf{H}}_{k+1} = \widehat{\mathbf{H}}_k + \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\mathbf{p}_k} - \frac{\widehat{\mathbf{H}}_k\mathbf{p}_k(\widehat{\mathbf{H}}_k\mathbf{p}_k)^T}{\mathbf{p}_k^T\widehat{\mathbf{H}}_k\mathbf{p}_k}$.

Note that instead of performing the last step we can update the inverse approximate Hessian $\widehat{\mathbf{H}}_k^{-1}$ using the Shermann-Morrison-Woodbury formula

$$\widehat{\mathbf{H}}_{k+1}^{-1} = \widehat{\mathbf{H}}_{k+1}^{-1} + (\mathbf{p}_k\mathbf{p}_k^T)\frac{\mathbf{p}_k\mathbf{y}_k^T + \mathbf{y}_k^T\widehat{\mathbf{H}}_k^{-1}\mathbf{y}_k}{(\mathbf{p}_k^T\mathbf{y}_k)^2} - \frac{\widehat{\mathbf{H}}_k^{-1}\mathbf{y}_k\mathbf{p}_k^T + \mathbf{p}_k\mathbf{y}_k^T\widehat{\mathbf{H}}_k^{-1}}{\mathbf{p}_k^T\mathbf{y}_k}.$$

Often the initial matrix $\widehat{\mathbf{H}}_0^{-1}$ is set to $\mathbf{I}$ which results in the first step being a steepest descent step. As the sequence of BFGS steps progress the approximate Hessian $\widehat{\mathbf{H}}_k$ (or approximate inverse approximate Hessian $\widehat{\mathbf{H}}_k^{-1}$) becomes increasingly close to the true Hessian $\mathbf{H}_k$.

Under certain conditions the sequence of steps taken in the BFGS method can be shown to converge superlinearly. The development of the BGFS and similar quasi-Newton methods have virtually replaced Newton-Raphson-like methods in practice (Nocedal & Wright, 1999).

Finally, we note that the quasi-Newton method BFGS is implemented in the `R` function `optim` and is used to fit several of the models in this thesis.

C.3.3 Repeat Hessian Newton's Method

In many of the models considered in this thesis the cost of calculating the gradient $\mathbf{g}_k$ vector of the likelihood is $O(nm)$ and the cost of calculating the Hessian matrix $\mathbf{H}$ is $O(nm^2)$ where n is the number of observations and m is O(the number of basis functions). When m becomes large the computational burden of calculating the Hessian dominates compared to the cost of inverting the Hessian, i.e $O(m^3)$. This problem also occurs in other situations in statistics, for example generalised linear models with a large number of predictors.

An alternative inexact Newton method to quasi-Newton methods for reducing the computational cost of Newton's method is not to calculate the Hessian at every iteration. Suppose that we want to use the Newton-Raphson method but only want recalculate the Hessian every r iterations. Consider the sequence $\{\mathbf{x}_{i,j}\}$ defined by

$$\{\mathbf{x}_{0,1}, \ldots, \mathbf{x}_{0,r}, \mathbf{x}_{1,1}, \ldots, \mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,r}, \ldots\}$$

where

$$\begin{aligned} \mathbf{x}_{i,j+1} &= \mathbf{x}_{i,j} - \mathbf{H}_i^{-1}\mathbf{g}_{i,j} \quad \text{for } j = 1, \ldots, r-1 \\ \mathbf{x}_{i+1,1} &= \mathbf{x}_{i,r} - \mathbf{H}_i^{-1}\mathbf{g}_{i,r} \end{aligned} \tag{C.18}$$

with

$$\mathbf{f}_{i,j} = f(\mathbf{x}_{i,j}), \qquad \mathbf{g}_{i,j} = \mathsf{D}_{\mathbf{x}}f(\mathbf{x}_{i,j}) \qquad \text{and } \mathbf{H}_i = \mathsf{H}_{\mathbf{x}}f(\mathbf{x}_{i,1}).$$

Also for convenience of notation we will denote $\mathbf{p}_{i,j} = \mathbf{x}_{i,j+1} - \mathbf{x}_{i,j}$ if $j = 1, \ldots, r-1$ and $\mathbf{p}_{i,r} = \mathbf{x}_{i+1,1} - \mathbf{x}_{i,r}$. We also note that if $r = 1$ then the Newton-Raphson iterates are retained.

Suppose the computational cost of calculating and inverting the Hessian is $O(nm^2 + m^3)$ and the cost of calculating the gradient and multiplying it by the inverse Hessian is $O(nm + m^2)$. If $r < m$ then the computational cost over the r iterations is no more than twice the cost of one Newton-Raphson iterate. For this reason we may also be interested in the convergence of the sequence $\{\mathbf{x}_{i,1}\}$ for different values of r.

We will now consider the convergence properties of $\{\mathbf{x}_{i,j}\}$. But before we do so we note that Lemma C.6 and Lemma C.7 hold if we simply replace $\mathbf{x}_k$ with $\mathbf{x}_{i,1}$, $\mathbf{x}_{k+1}$ with $\mathbf{x}_{i,2}$ or $\mathbf{x}_{i+1,1}$ for the cases $r > 1$ and $r = 1$ respectively.

Lemma C.10: *Suppose that Assumptions C.5 hold. Consider the sequence of iterates* $\{\mathbf{x}_{i,j}\}$ *defined in (C.18). Then, for* $j = 1, \ldots, r-1$

$$\|\mathbf{g}_{i,j+1}\| \leq L_2 \|\mathbf{H}_i^{-1}\|^2 \|\mathbf{g}_{i,j}\| \left(\sum_{k=1}^{j-1} \|\mathbf{g}_{i,k}\| \right) + \frac{L_2}{2} \|\mathbf{H}_i^{-1}\|^2 \|\mathbf{g}_{i,j}\|^2 \tag{C.19}$$

and if $j = r$ *we replace the right hand side of (C.19) by* $\|\mathbf{g}_{i+1,1}\|$.

Proof: Using Assumptions C.5, and the relation, $\mathbf{g}_{i,j} + \mathbf{H}_i \mathbf{p}_{i,j} = \mathbf{0}$, for the cases $j = 1, \ldots, r-1$

$$\begin{aligned}
\|\mathbf{g}_{i,j+1}\| &= \|\mathbf{g}_{i,j+1} - \mathbf{g}_{i,j} - \mathbf{H}_i \mathbf{p}_{i,j}\| \\
&= \left\| \int_0^1 [\mathsf{H}_{\mathbf{x}} f(\mathbf{x}_{i,j} + t\mathbf{p}_{i,j}) \mathbf{p}_{i,j}] dt - \mathbf{H}_i \mathbf{p}_{i,j} \right\| \\
&\leq \int_0^1 \|\mathsf{H}_{\mathbf{x}} f(\mathbf{x}_{i,j} + t\mathbf{p}_{i,j}) - \mathbf{H}_i\| \|\mathbf{p}_{i,j}\| dt \\
&\leq \int_0^1 L_2 \|\mathbf{x}_{i,j} - \mathbf{x}_{i,1} + t\mathbf{p}_{i,j}\| \|\mathbf{p}_{i,j}\| dt \\
&= \int_0^1 L_2 \left\| \sum_{k=1}^{j-1} \mathbf{p}_{i,k} + t\mathbf{p}_{i,j} \right\| \|\mathbf{p}_{i,j}\| dt
\end{aligned}$$

The second line follows from (C.9) in Taylor's Theorem; the third line follows from properties of vector norms; the forth line follows from the fact that $\mathsf{H}_{\mathbf{x}} f(\mathbf{x})$ is Lipschitz continuous with Lipschitz constant L_2 and the last line follows from

$$\mathbf{x}_{i,j} = \mathbf{x}_{i,1} + \sum_{k=1}^{j-1} \mathbf{p}_{i,k}$$

which we get from recursive application of (C.18). Finally, the lemma follows from the triangle inequality of vector norms and $\|\mathbf{p}_{i,j}\| \leq \|\mathbf{H}_i^{-1}\| \|\mathbf{g}_{i,j}\|$. The case $j = r$ follows from almost identical arguments.

■

Lemma C.11 *Suppose that Assumptions C.5 hold. Consider the sequence of iterates* $\{\mathbf{x}_{i,j}\}$ *defined in (C.18). Then, for* $j = 1, \ldots, r-1$

$$\|\mathbf{x}_{i,j+1} - \mathbf{x}_*\| \leq \frac{L_2}{2}\|\mathbf{H}_i^{-1}\|\|\mathbf{x}_{i,j} - \mathbf{x}_*\|^2 + L_1 L_2 \|\mathbf{H}_i^{-1}\|^2 \|\mathbf{x}_{i,j} - \mathbf{x}_*\| \sum_{k=1}^{j-1} \|\mathbf{x}_{i,k} - \mathbf{x}_*\| \tag{C.20}$$

and if $j = r$ *we replace the right hand side of (C.19) by* $\|\mathbf{x}_{i+1,1} - \mathbf{x}_*\|$.

Proof: Using Assumptions C.5 and the relations in (C.18) we have

$$\begin{aligned}
\|\mathbf{x}_{i,j+1} - \mathbf{x}_*\| &= \|\mathbf{x}_{i,j} - \mathbf{x}_* - \mathbf{H}_i^{-1}\mathbf{g}_{i,j}\| \\
&= \|\mathbf{H}_i^{-1}(\mathbf{H}_i(\mathbf{x}_{i,j} - \mathbf{x}_*) - (\mathbf{g}_{i,j} - \mathbf{g}_*))\| \\
&\leq \|\mathbf{H}_i^{-1}\| \left\| \mathbf{H}_i(\mathbf{x}_{i,j} - \mathbf{x}_*) - \int_0^1 \mathsf{H}_{\mathbf{x}} f(\mathbf{x}_{i,j} + t(\mathbf{x}_* - \mathbf{x}_{i,j}))(\mathbf{x}_{i,j} - \mathbf{x}_*)\, dt \right\| \\
&\leq \|\mathbf{H}_i^{-1}\|\|\mathbf{x}_{i,j} - \mathbf{x}_*\| \left\| \int_0^1 \mathsf{H}_{\mathbf{x}} f(\mathbf{x}_{i,j} + t(\mathbf{x}_* - \mathbf{x}_{i,j})) - \mathbf{H}_i\, dt \right\| \\
&\leq L_2\|\mathbf{H}_i^{-1}\|\|\mathbf{x}_{i,j} - \mathbf{x}_*\| \int_0^1 \|\mathbf{x}_{i,j} - \mathbf{x}_{i,1} + t(\mathbf{x}_* - \mathbf{x}_{i,j})\| dt \\
&\leq L_2\|\mathbf{H}_i^{-1}\|\|\mathbf{x}_{i,j} - \mathbf{x}_*\| \left(\int_0^1 t\|\mathbf{x}_{i,j} - \mathbf{x}_*\| dt + \|\mathbf{x}_{i,j} - \mathbf{x}_{i,1}\| \right) \\
&= L_2\|\mathbf{H}_i^{-1}\|\|\mathbf{x}_{i,j} - \mathbf{x}_*\| \left(\tfrac{1}{2}\|\mathbf{x}_{i,j} - \mathbf{x}_*\| + \|\textstyle\sum_{k=1}^{j-1} \mathbf{p}_{i,k}\| \right) \\
&\leq L_2\|\mathbf{H}_i^{-1}\|\|\mathbf{x}_{i,j} - \mathbf{x}_*\| \left(\tfrac{1}{2}\|\mathbf{x}_{i,j} - \mathbf{x}_*\| + \|\mathbf{H}_i^{-1}\| \textstyle\sum_{k=1}^{j-1} \|\mathbf{g}_{i,k} - \mathbf{g}_*\| \right) \\
&\leq L_2\|\mathbf{H}_i^{-1}\|\|\mathbf{x}_{i,j} - \mathbf{x}_*\| \left(\tfrac{1}{2}\|\mathbf{x}_{i,j} - \mathbf{x}_*\| + L_1\|\mathbf{H}_i^{-1}\| \textstyle\sum_{k=1}^{j-1} \|\mathbf{x}_{i,k} - \mathbf{x}_*\| \right)
\end{aligned}$$

The third line follows from (C.9) in Taylor's Theorem and properties of vector norms; the forth line follows from properties of vector norms; the fifth line follows from the fact that $\mathsf{H}_{\mathbf{x}} f(\mathbf{x})$ is Lipschitz continuous with Lipschitz constant L_2; the sixth line follows from the relations defined by (C.18); the seventh line follows from,

$$\mathbf{x}_{i,j} = \mathbf{x}_{i,1} + \sum_{k=1}^{j-1} \mathbf{p}_{i,k}$$

so that

$$\|\mathbf{x}_{i,j} - \mathbf{x}_{i,1}\| \leq \|\mathbf{H}_i^{-1}\| \sum_{k=1}^{j-1} \|\mathbf{g}_{i,k} - \mathbf{g}_*\|$$

since $\mathbf{g}_* = \mathbf{0}$. Finally the last line follows form the fact that $\mathsf{D}_{\mathbf{x}} f(\mathbf{x})$ is Lipschitz continuous with Lipschitz constant L_1. The case $j = r$ follows from almost identical arguments. ■

Again, following Theorem 3.7 of Nocedal & Wright (1999), if we assume that the iterates (C.18) converge to $\mathbf{x}_*$ then the inequality (C.14) holds. Using this inequality and Lemmas C.6, C.7, C.10 and C.11 we have the fixed point relations

$$u_{i,2} \leq c_1 u_{i,1}^2, \quad u_{i,j+1} \leq c_1 u_{i,j}^2 + c_2 u_{i,j} \sum_{k=1}^{j-1} u_{i,k}, \quad \text{and} \quad u_{i+1,1} \leq c_1 u_{i,r}^2 + c_2 u_{i,r} \sum_{k=1}^{r-1} u_{i,k}$$

where c_1, c_2 are fixed positive constants and $u_{i,j}$ may be replaced by $\|\mathbf{g}_{i,j}\|$ or $\|\mathbf{x}_{i,j} - \mathbf{x}_*\|$. We can see that if $\mathbf{x}_{i,j}$ is sufficiently close to $\mathbf{x}_*$ then the first step, the Newton step will

halve the number of decimal places between iterates. However it is unclear for general r under what conditions the rates of convergence the sequence $\{u_{i+1,1}\}$ will have. Here we will only consider the simplest case $r = 2$.

Theorem C.12: *Suppose that Assumptions C.5 and C.8 holds. For the case where $r = 2$ if*

$$\|\mathbf{g}_{i,1}\| < \frac{-1+\sqrt{5}}{2L_2\|\mathbf{H}_*^{-1}\|^2} \tag{C.21}$$

for sufficiently a large $i > i_0$ then the rate of convergence of $\{\mathbf{g}_{i,1}\}$ is cubic.

Proof: Following Theorem 3.7 of Nocedal & Wright (1999) we have(C.14). Using this inequality and the inequalities in Lemmas C.6 and C.10 we have

$$\begin{aligned} \|\mathbf{g}_{i,2}\| &\leq c\|\mathbf{g}_{i,1}\|^2 \\ \|\mathbf{g}_{i+1,1}\| &\leq c\|\mathbf{g}_{i,2}\|^2 + 2c\|\mathbf{g}_{i,2}\|\|\mathbf{g}_{i,1}\| \end{aligned} \tag{C.22}$$

where $c = 2L_2\|\mathbf{H}_*^{-1}\|^2 > 0$. Combining these two inequalities

$$\|\mathbf{g}_{i+1,1}\| \leq c^3\|\mathbf{g}_{i,1}\|^4 + 2c^2\|\mathbf{g}_{i,1}\|^3. \tag{C.23}$$

Treating (C.23) as a fixed point iteration

$$u_{i+1} = G(u_i) = c^3u_i^4 + 2c^2u_i^3 \tag{C.24}$$

where $u_i = \|\mathbf{g}_{i,1}\|$. If $u_i > 0$ then $u_{i+k} > 0$ for all $k > 0$. Fixed points satisfy $u_i = G(u_i)$. The fixed points of (C.24) are

$$0, -\frac{1}{c}, \frac{-1+\sqrt{5}}{2c} \text{ and } -\frac{1+\sqrt{5}}{2c}.$$

Since $-\frac{1+\sqrt{5}}{2c}$ and $-\frac{1}{c}$ are not possible (since $u_i > 0$) we only need to consider the stability of the other two positive fixed points. The condition for stability of a fixed point u_* is $|G'(u_*)| < 1$.

$$\begin{aligned} G'(0) &= 0 \\ G'\left(\frac{-1+\sqrt{5}}{2c}\right) &\approx 3.2361 > 1. \end{aligned}$$

Thus, 0 is a stable fixed point and $\frac{-1+\sqrt{5}}{2c}$ is an unstable fixed point. So if

$$\|\mathbf{g}_{i,1}\| < \frac{-1+\sqrt{5}}{2L_2\|\mathbf{H}_*^{-1}\|^2}$$

then $\{\|\mathbf{g}_{i,1}\|\}$ converges to 0. Finally,

$$\lim_{i\to\infty} \frac{\|\mathbf{g}_{i+1,1}\|}{\|\mathbf{g}_{i,1}\|^3} \leq \lim_{i\to\infty} c^3\|\mathbf{g}_{i,1}\| + 2c^2 = 2c^2$$

so that the rate of convergence of $\{\|\mathbf{g}_{i,1}\|\}$ is cubic.

■

Various constants in the above theorem rely on a providential algebraic form for the fixed points. An analogous theorem for the convergence of the $\{\mathbf{x}_{i,1} - \mathbf{x}_*\}$ would be possible if the constants L_1, L_2 and $\|\mathbf{H}_*^{-1}\|$ where known.

We note that since for Newton-Raphson iterates, for $\mathbf{x}_{i,1}$ sufficiently close to $\mathbf{x}_*$,

$$\frac{\|\mathbf{x}_{i+2,1} - \mathbf{x}_*\|}{\|\mathbf{x}_{i+1,1} - \mathbf{x}_*\|^2} \frac{\|\mathbf{x}_{i+1,1} - \mathbf{x}_*\|^2}{\|\mathbf{x}_{i,1} - \mathbf{x}_*\|^4} = \frac{\|\mathbf{x}_{i+2,1} - \mathbf{x}_*\|}{\|\mathbf{x}_{i,1} - \mathbf{x}_*\|^4} \leq M^3$$

the rate of convergence of two consecutive Newton-Raphson iterates is quartic. This means that there is a loss of efficiency of the repeat Hessian Newton's method if we solely consider the rate of convergence. On the other hand, for expensive to compute Hessians, repeated Hessian steps are much faster to compute since only the gradient need be calculated for these steps. In Chapter 3, where this strategy was adopted for GLMMs, the cost of fitting such models was reduced by a factor of 2 or more.

Bibliography

Abramowitz, M. and Stegun, I. (1964). *Handbook of Mathematical Functions*. New York: Dover.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Andrews, D.F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B 36*(1), 99–102.

Attias, H (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, San Francisco, CA, pp. 21–30. Morgan Kaufmann.

Attias, H (2000). A variational Bayesian framework for graphical models. In Solla, S., Leen, T., and Muller, K.-R. (Eds.), *Advances in Neural Information Processing Systems*, Volume 12, pp. 209–215. Cambridge, MA: MIT press.

Büchner, A.G. and Mulvenna, M.D. (1998). Discovering internet marketing intelligence through online analytical web usage mining. *SIGMOD Rec. 27*(4), 54–61.

Bachrach, L.K., Hastie, T., Wang, M.-C., Narasimhan, B., and Marcus, R. (1999). Bone mineral acquisition in healthy Asian, Hispanic, Black and Caucasian youth. A longitudinal study. *Journal of Clinical Endocrinology and Metabolism 84*, 4702–12.

Baladandayuthapani, V., Mallick, B.K., and Carroll, R.J. (2005). Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computational and Graphical Statistics 14*(2), 378–394.

Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics*. London: Chapman & Hall.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*. London: Chapman & Hall.

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. Ph. D. thesis, Gatsby Computational Neuroscience Unit.

Beal, M. and Ghahramani, Z. (2002). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics*, Volume 7. Oxford: Oxford University Press.

Beliakov, G. (2004). Least squares splines with free knots: global optimization approach. *Applied Mathematics and Computation 149*(3), 783–798.

Berndt, E.R. (1991). *The Practice of Econometrics: Classical and Contemporary*. Reading, Massachusetts: Addison-Wesley.

Berry, M.J.A. and Linoff, G.S. (2004). *Data mining techniques: For marketing, sales, and customer relationship management* (2nd ed.). Indiana: Wiley Publishing, Inc.

Bishop, C.M. (1999). Variational PCA. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN)*.

Bishop, C.M. and Tipping, M.E. (2000). Variational relevance vector machines. In *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, USA, pp. 46–53. Morgan Kaufmann Publishers Inc.

Bishop, C.M. and Winn, J. (2003). Structural variational distributions in VIBES. In Bishop, C.M. and Frey, B. (Eds.), *Proceedings of Artificial Intelligence*, Florida, USA.

Booth, J.G. and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association 93*, 262–272.

Box, G.E.P. (1979). Robustness in the strategy of scientific model building. In Launer, R.L. and Wilkinson, G.N. (Eds.), *Robustness in Statistics*, pp. 201–236. New York: Academic Press.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. New York: Cambridge University Press.

Breiman, L. (2001). Statistical modeling: the two cultures (with discussion). *Statistical Science 16*, 199–231.

Breiman, L., Friedman, J.H, Olshen R.A, and Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth International Group.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal American Statistical Association 88*(1), 9–25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika 82*(1), 81–91.

Browne, W.J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis 1*(3), 473–514.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics 17*, 453–510.

Byrd, R.H., S. and Schultz, G.A. (1987). A trust region algorithm for nonlinearly constrained optimisation. *SIAM Journal of Numerical Analysis 24*, 1152–1170.

Cantoni, E. and Hastie, T. (2002). Degrees of freedom tests for smoothing splines. *Biometrika 89*, 251–265.

Cao, R., Cuevas, A., and González Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis 17*(2), 153–176.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman & Hall.

Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models* (2nd ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Celeux, G., Forbes, F., Robert, C., and Titterington, D.M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis 1*(4), 651–674.

Celis, M., Dennis, J.E., and Tapia, R.A. (1994). A trust region strategy for nonlinear equality constrained optimisation. In Boggs, P., Byrd, R., and Schnabel, R. (Eds.), *Numerical optimisation*, pp. 71–82. Philadelphia: SIAM.

Chambers, J.M. and Hastie, T.J. (1992). *Statistical Models in S*. New York: Chapman & Hall.

Chau, K.-W. and Muttil, N. (2007). Data mining and multivariate statistical analysis for ecological system in coastal waters. *Journal of Hydroinformatics 9*(4), 305–317.

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

Clayton, D. (1996). Generalized linear mixed models. In Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 275–301. London: Chapman & Hall.

Consonni, G. and Marin, J.-M. (2007). Mean-field variational approximate Bayesian inference for latent variable models. *Computational Statistics Data Analysis 52*(2), 790–798.

Corduneanu, A. and Bishop, C.M. (2001). Variational Bayesian model selection for mixture distributions. In Jaakkola, T. and Richardson, T. (Eds.), *Artificial Intelligence and Statistics*, pp. 27–34. Los Altos, CA: Morgan Kaufmann.

Cowles, M.K. and Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of American Statistics Association 91*, 883–904.

Cowles, M.K. and Rosenthal, J.S. (1998). A simulation approach to convergence rates for markov chain Monte Carlo. *Statistics and Computing 8*, 115–124.

Cox, D. and Koh, E. (1989). A smoothing spline based test of model adequacy in polynomial regression. *Annals of the Institute of Statistical Mathematics 41*, 383–400.

Crainiceanu, C., Ruppert, D., Carroll, R. J., Adarsh, J., and Goodner, B. (2007). Spatially adaptive Bayesian P-splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics 16*, 265–288.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*(4), 377–403.

Csiszár, I. and Shields, P. (2004). Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory 1*(4), 417–528.

Davidian, M. and Carroll, R.J. (1997). Variance function estimation. *Journal of the American Statistical Association 82*, 1072–1091.

de Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer-Verlag.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B 68*, 411–436.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 339*, 1–38.

Denison, D.G.T., Holmes, C.C., Mallick, D.K., and Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley.

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*(2), 333–350.

Dennis, J.E. and Schnabel, R.B. (1983). *Numerical methods for unconstrained optimization and nonlinear equations*. Englewood Cliffs, NJ: Prentice Hall.

Diggle, P., Heagerty, P., Liang, K.-L., and Zeger, S. (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford University Press.

DiMatteo, I., Genovese, C.R., and Kass, R.E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika 88*(4), 1055–1071.

Donoho, D.L. and Johnstone, J.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*(3), 425–455.

Donoho, D.L. and Johnstone, J.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association 90*(432), 1200–1224.

Donoho, D.L., Johnstone, J.M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society. Series B 57*(2), 301–369.

Duffett, P. and Vernik, R. (1997). Software system visualisation: Netmap investigations. Technical Report DSTO-TR-0558, Australian Government Department of Defence: Defence Science and Technology Organisation.

Eubank, R.L. (1994). A simple smoothing spline. *The American Statistician 48*, 103–106.

Eubank, R.L. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. London: Chapman & Hall.

Fan, Y., Leslie, D.S., and Wand, M.P. (2007). Generalised linear mixed model analysis via sequential Monte Carlo sampling. Technical Report 07:10, University of Bristol, Bristol, U.K.

Fasshauer, G.E. (2007). *Meshfree Approximation Methods with Matlab*, Volume 6 of *Interdisciplinary Mathematical Sciences*. Singapore: World Scientific.

Faul, A.C. and Tipping, M.E. (2001). A variational approach to robust regression. In *Proceedings of the International Conference on Artificial Neural Networks*, Berlin / Heidelberg, pp. 95–102. Springer.

Fisher, R.A. ([1935]1956). Mathematics of a lady tasting tea. In Newman, J.R. (Ed.), *The World of Mathematics*, pp. 1512–1521. New York: Simon & Schuster.

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I.H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics 20*(15), 2479–2481.

French, J.L. and Wand, M.P. (2004). Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics 5*(2), 177–191.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics 19*(1), 1–67.

Friston, K.J., Glaser, D.E., Henson, R.N., Kiebel, S.J., Phillips, C., and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: Applications. *Neuroimage 16*, 484–512.

Garthwaite, P.H., Jolliffe, I.T., and Jones, B. (2002). *Statistical Inference* (2nd ed.). Oxford: Oxford University Press.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. (Comment on an article by Browne & Draper, 2006). *Bayesian Analysis 1*(3), 515–534.

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

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science 7*, 473–482.

Ghahramani, Z. and Beal, M. (2000). Propagation algorithms for variational Bayesian learning. In Touretzky, D.S., Mozer, M.C., and Hasselmo, M.E. (Eds.), *Advances in Neural Information Processing*, Volume 13, Cambridge, MA, USA, pp. 507–513. MIT Press.

Ghahramani, Z. and Beal, M. (2001). Graphical Models and Variational Methods. In Opper, M. and Saad, D. (Eds.), *Advanced Mean Field Methods - Theory and Practice*. MIT Press.

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

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics 41*, 337–348.

Giri, N. (1964a). On the LRT of a normal multivariate testing problem. *Annals of Statistics 35*, 181–189.

Giri, N. (1964b). Correction to "On the LRT of a normal multivariate testing problem". *Annals of Statistics 35*, 1388.

Golub, G.H. and Van Loan, C.F. (1996). *Matrix Computations* (3rd ed.). Baltimore & London: Johns Hopkins University Press.

Gray, R.J. (1994). Spline-based tests in survival analysis. *Biometrics 50*, 640–652.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer.

Guo, G., Li, S.Z., and Chan, K. (2000). Face recognition by support vector machines. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, Los Alamitos, CA, USA, pp. 196–201. IEEE Computer Society.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research 3*, 1157–1182.

Hall, P., Humphreys, K., and Titterington, D. M. (2002). On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values. *Journal of the Royal Statistical Society Series B 64*, 549–564.

Hall, P. and Opsomer, J.D. (2005). Theory for penalised spline regression. *Biometrika 92*, 105–118.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Hand, D.J (2006). Classifier technology and the illusion of progress (with discussion). *Statistical Science 21*, 1–34.

Hannig, J. and Marron, J.S. (2006). Advanced distribution theory for SiZer. *Journal of the American Statistical Association 101*, 484–499.

Harville, D.A. (1997). *Matrix Algebra From a Statistican's Perspective*. New York: Springer.

Hastie, T. (2008). *gam: Generalized Additive Models*. R package version 1.0.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.

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

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Heisele, B., Ho, P., Wu, J., and Poggio, T. (2003). Face recognition: Comparing component-based and global approaches. *Computer Vision and Image Understanding 91*(1/2), 6–21.

Hickernell, F.J., Lemieux, C., and Owen, A.B. (2005). Control variates for quasi-Monte Carlo. *Statistical Science 20*, 1–31.

Hinton, G.E. and van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *COLT '93: Proceedings of the sixth annual conference on Computational learning theory*, New York, USA, pp. 5–13. Association for Computing Machinery.

Hoover, K.D. and Perez, S.J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *The Econometrics Journal 2*(2), 167–191.

Horton, N. J. and Laird, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research 8*, 37–50.

Horton, N.J. and Kleinman, K.P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *American Statistical Association 61*(1), 79–90.

Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.

Humphreys, K. and Titterington, D. (2000). Approximate Bayesian inference for simple mixtures. In *COMPSTAT'2000, Proceedings in Computational Statistics*, Berlin. Springer.

Humphreys, K. and Titterington, D. (2001). Some examples of recursive variational approximations for Bayesian inference. In Opper, M. and Saad, D. (Eds.), *Advances Mean Field Methods: Theory and Practice*. Cambridge, MA: MIT Press.

Hurvich, C.M., Simonoff, J.S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B 60*(2), 271–293.

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

Ibrahim, J.G., Chen, M.-H., and Lipsitz, S.R. (1999a). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics 55*, 591–596.

Ibrahim, J.G., Chen, M.-H., and Lipsitz, S.R. (1999b). Missing responses in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society, Series B 61*, 173–190.

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

Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika 48*, 419–426.

INSIGHTFUL CORPORATION (2007). S-plus. Seattle, WA.

Jaakkola, T. S. and Jordan, M.I. (1997). Recursive algorithms for approximating probabilities in graphical models. In Mozer, M.C., Jordan, M.I., and Petsche, T. (Eds.), *Advances in neural information processing systems 9*, Cambridge, MA. MIT Press.

Jaakkola, T.S. (2001). Tutorial on variational approximation methods. In Opper, M. and Saad, D. (Eds.), *Advanced Mean Field Methods: Theory and Practice*. Cambridge, MA: MIT Press.

Jaakkola, T.S. and Jordan, M.I. (1998). Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models*, pp. 163–173. Kluwer Academic Publishers.

Jaakkola, T.S. and Jordan, M.I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing 10*, 25–37.

Jones, M. C., Marron, J.S., and Sheather, S.J. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics 11*(3), 337–381.

Jordan, M.I. (2004). Graphical models. *Statistical Science (Special Issue on Bayesian Statistics) 19*, 140–155.

Jordan, M.I., Ghahramani, Z., Jaakkola T.S., and Saul, L.K. (1999). An introduction to variational methods for graphical models. *Machine Learning 37*(2), 183–233.

Jupp, D.L.B. (1978). Approximation to data by splines with free knots. *SIAM Journal on Numerical Analysis 15*(2), 328–343.

Karimabadi, H., Sipes, T.B., White, H., Marinucci, M., Dmitriev, A., Chao, J.K., Driscoll, J., and Balac, N. (2007). Data mining in space physics: MineTool algorithm. *Journal of Geophysical Research 112*(A11215).

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kauermann, G., Ormerod, J.T., and Wand, M.P. (2008). Parsimonious classification via generalised linear mixed models. Technical report. (submitted).

Kooperberg, C., Bose, S., and Stone, C.J. (1997). Polychotomous regression. *Journal of the American Statistical Association 92*, 117–127.

Kotz, S., Balakrishnan, N., and Johnson, N.L. (2000). *Continuous Multivariate Distributions*, Volume 1: Models and Applications. John Wiley.

Kovalerchuk, B. and Vityaev E. (2000). *Data mining in finance: advances in relational and hybrid methods*. Norwell, MA, USA: Kluwer Academic Publishers.

Krivobokova, T. (2008). *AdaptFit: Adaptive Semiparametic Regression.*

Krivobokova, T., Crainiceanu C.M., and Kauermann, G. (2007). Fast adaptive penalised splines. *Journal of Computational & Graphical Statistics 17*(1), 1–20.

Kruk, S., Muramatsu, M., Rendl, F., Vanderbei, R.J., and Wolkowicz, H. (2001). The Gauss-Newton direction in semidefinite programming. *Optimization Methods & Software 15*(1), 1–28.

Kuk, A.Y.C. and Cheng, Y.W. (1997). The Monte Carlo Newton-Raphson algorithm. *Journal of Statistical Computing and Simulation 59*, 233–250.

Kuo, F., Dunsmuir, W.T.M., Sloan, I.H., Wand, M.P., and Womersley, R.S. (2008). Quasi-Monte Carlo for highly structured generalised response models. *Methodology and Computing in Applied Probability 10*(2), 239–275.

Kuss, M. (2006). *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*. Ph. D. thesis, Technische Universität Darmstadt.

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

Lee, Y.-J., Mangasarian, O.L., and Wolberg, W.H. (1999). Breast cancer survival and chemotherapy: A support vector machine analysis. Technical Report 99-10, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Volume 55, 2000, 1-10.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Journal of the Royal Statistical Society, Applied Statistics, Series C 50*(3), 325–335.

Lin, X (1997). Variance component testing in generalised linear models with random effects. *Biometrika 84*, 309–326.

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*(435), 1007–1016.

Lindstrom, M.J. (1999). Penalized estimation of free-knot splines. *Journal of Computational & Graphical Statistics 8*, 333–352.

Little, R.J.A. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association 87*, 1227–1237.

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association 88*, 125–134.

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

Liu, Q. and Pierce, D.A. (1994). A note on Gauss-Hermite quadrature. *Biometrika 81*(3), 624–629.

Loader, C. (1999). *Local Regression and Likelihood*. New York: Springer-Verlag.

Lonergan, B. (1958). *Insight: A Study of Human Understanding* (2nd ed.). London: Longmans, Green and Co.

Lovell, B.C. and Chen, S. (2005). Robust face recognition for data mining. In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining*, pp. 965–972. Idea Group.

Luengo, F., Cofiño, A. S., and Gutierrez, J. M. (2004). GRID oriented implementation of self-organizing maps for data mining in meteorology. *Lecture Notes in Artificial Intelligence 2970*, 163–171.

Lui, C. and Rubin, D.B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECEM. *Statistica Sinica 5*, 19–39.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association 92*, 107–116.

Lyche, T. and Schumaker, L.L. (1973). Computation of smoothing and interpolating natural splines via local bases. *SIAM Journal on Numerical Analysis 10*(6), 1027–1038.

MacKay, D.J.C. (1995). Developments in probabilistic modelling with neural networks – ensemble learning. In Kappen, B. and Gielen, S. (Eds.), *Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks*, Netherlands. Nijmegen.

MacKay, D.J.C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press.

Madeira, S.C., Oliveira, A.L., and Conceição, C.S. (2003). A data mining approach to credit risk evaluation and behaviour scoring. In Moura-Pires, F. and Abreu, S. (Eds.), *Progress in Artificial Intelligence, 11th Protuguese Conference on Artificial Intelligence, EPIA 2003, Beja, Portugal, December 4-7, 2003, Proceedings*, Volume 2902 of *Lecture Notes in Computer Science*, pp. 184–188. Springer.

Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus*. Chichester: John Wiley & Sons.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics 15*(4), 661–675.

Mangasarian, O.L., Street, W.N., and Wolberg, W.H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research 43*(4), 570–577.

Mao, V.W. and Zhao, L.H. (2003). Free-knot polynomial splines with confidence intervals. *Journal of the Royal Statistical Society. Series B 65*(4), 901–919.

Marron, J.S. and Nolan, D. (1988). Canonical kernels for density estimation. *Statistics & Probability Letters 7*(3), 195–199.

Marx, B.D. and Eilers, P. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science 11*(2), 89–121.

McCullagh, P. (1994). Maximum likelihood estimation of variance components for binary data. *Journal of the American Statistical Association 89*, 330–335.

McCullagh, P. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92*, 162–170.

McCullagh, P. and Nelder, J. (1989). *Generalised Linear Models* (2nd ed.). London: Chapman & Hall.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. New York: John, Wiley & Sons.

McGrory, C.A. and Titterington, D.M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis 51*(11), 5352–5367. Advances in Mixture Models.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*(6), 1087–1092.

Miyata, S. and Shen, X. (2003). Adaptive free-knot splines. *Journal of Computational & Graphical Statistics 12*, 197–213.

Moré, J.J. and Sorensen, D.C. (1983). Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing 3*, 553–572.

Naylor, J.C. and Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics 31*(3), 214–225.

Neal, R.M. and Hinton, G.E. (1998). A new view of the EM algorithm that justifies incremental, sparse and other variants. In Jordan, M.I. (Ed.), *Learning in Graphical Models*, pp. 355–368. Kluwer Academic Publishers.

Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A 231*, 289–337.

Ngo, L. and Wand, M.P. (2004). Smoothing with mixed model software. *Journal of Statistical Software 9*(1), 1–54.

Nocedal, J. and Wright, S.J. (1999). *Numerical Optimization*. New York: Springer.

Nussbaum, M (1985). Spline smoothing in regression models and asymptotic efficiency in L2. *The Annals of Statistics 13*, 984–997.

Nychka, D. and Cummins, D.J. (1996). Comment on paper by Eilers and Marx. *Statistical Science 11*, 104–105.

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

Nychka, D. and Saltzman, N. (1998). Design of air quality monitoring networks. In Nychka, D., Piegorsch, W.W., and Cox, L.H. (Eds.), *Case Studies in Environmental Statistics*, pp. 51–76. New York: Springer-Verlag.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science 1*, 505–527.

Park, B.U. and Turlach, B.A. (1992). Practical performance of several data driven bandwidth selectors (with discussion). *Computational Statistics 7*(3), 251–270. Correction in Vol. 9, p. 79.

Patterson, H.D. and Thompson, H. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrica 58*, 545–554.

Penny, W.D. and Roberts, S.J. (2000). Variational Bayes for 1-dimensional mixture models. Technical Report PARG-2000-01, Oxford University.

Perez-Iratxeta, C., Bork, P., and Andrade, M.A. (2002). Association of genes to genetically inherited diseases using data mining. *Nature Genetics 31*, 316–319.

Phua, C., Lee, V., Gayler, R., and Smith, K. (2006). *Intelligence and Security Informatics*, Volume 3917/2006 of *Lecture Notes in Computer Science*, Chapter Temporal Representation in Spike Detection of Sparse Personal Identity Streams, pp. 115–126. Berlin/Heidelberg: Springer.

Pintore, A., Speckman, P., and Holmes, C.C. (2006). Spatially adaptive smoothing splines. *Biometrika 93*(1), 113–125.

Pittman, J. (2002). Adaptive splines and genetic algorithms. *Journal of Computational & Graphical Statistics 11*, 615–638.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: John Wiley & Sons.

Raudenbush, S.W., Yang, M.-L., and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of Computational & Graphical Statistics 9*(1), 141–157.

Rijmen, F. and Vomlel, J. (2007). Assessing the performance of variational methods for mixed logistic regression models. *Journal of Statistical Computation and Simulation*. (accepted).

Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods* (2 ed.). New York: Springer-Verlag.

Rockafellar, R. (1972). *Convex Analysis*. Princeton, New Jersey: Princeton University Press.

Rosenthal, J.S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of American Statistics Association 90*, 558–566.

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

Rubin, D. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Rubinstein, B.Y. (1981). *Simulation and the Monte Carlo Method*. New York: Wiley & Sons.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational & Graphical Statistics 11*(4), 735–757.

Ruppert, D. and Carroll, R.J. (2000). Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics 42*(2), 205–223.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. New York: Cambridge University Press.

Ruszczyński, A. (2006). *Nonlinear Optimization*. Princeton: Princeton University Press.

Salford Systems (2000, Version 2). MARS for Windows. http://www.salford-systems.com/. San Diego, California.

SAS INSTITUTE, INC. (2007). Sas. Cary, NC.

Saul, L.K., Jaakkola, T. S., and Jordan, M.I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research 4*, 61–76.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. Cambridge, MA: MIT Press.

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

Schoenberg, I. (1964). Spline functions and the problem of gradation. *Proceedings of the National Academy of Sciences of the United States of America 52*, 947–950.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York, Chichester: John Wiley & Sons.

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

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association 82*, 605–610.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B 53*, 683–690.

Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*. New York: Wiley.

Silvapulle, M.J. and Sen, P.K. (2005). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. New York: John Wiley & Sons.

Smyth, G. (2008). *statmod: Statistical Modeling*. R package version 1.6.1.

Solo, V. (2000). A simple derivation of the smoothing spline. *The American Statistician 54*, 40–43.

Solomon, P.J. and Cox, D.R. (1992). Nonlinear component of variance models. *Biometrika 79*(1), 1–11.

Speed, T. (1991). Comment on "That BLUP is a good thing: The estimation of random effects." paper by Robinson. *Statistical Science 6*, 42–44.

Spiegelhalter, D., Thomas, A., and Best, N. (2000). WinBUGS Version 1.3 User Manual. www.hrc-bsu.cam.ac.uk/bugs.

Staudenmayer, D., Lake, E., and Wand, M.P. (2008). Robustness for general design mixed models using the t-distribution. *Statistical Modelling*. (submitted).

Staudte, R.G. and Sheather, S.J. (1990). *Robust Estimation and Testing*. New York: Wiley.

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.

Sutradhar, B.C. and Rao, R.P. (2001). On marginal quasi-likelihood inference in generalized linear mixed models. *Journal of Multivariate Analysis 76*, 1–34.

THE MATHWORKS, INC (2007). Matlab. Natick, MA.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostat 3*(2), 245–265.

Tierney, L., Kass, R.E., and Kadane, J.B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association 84*, 710–16.

Tipping, M.E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research 1*, 211–244.

Tipping, M.E. and Lawrence, N.D. (2003). A variational approach to robust Bayesian interpolation. In Molina, C. , Adali, T. , Larsen, J., Hulle, M.V., Douglas, S.C., and Rouat, J. (Eds.), *Proceedings of the IEEE 2003 Neural Networks for Signal Processing Workshop*, Piscataway, New Jersey, pp. 229–238. IEEE Press.

Tipping, M.E. and Lawrence, N.D. (2005). Variational inference for Student-t models: Robust Bayesian interpolation and generalised component analysis. *Neurocomputing 69*(1-3), 123–141.

Titterington, D.M. (2004). Bayesian methods for neural networks and related models. *Statistical Science 19*, 128–139.

Tuerlinckx, F., Rijmen, F., Verbeke, G., and de Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology 59*, 225–255.

Tukey, J.W. (1962). The future of data analysis. *The Annals of Mathematical Statistics 33*(1), 1–67.

Tukey, J.W. (1965). The technical tools of statistics. *The American Statistician 19*(2), 23–28.

Ueda, N. and Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks 15*(10), 1223–1241.

Vandenberghe, L. and Boyd, S. (1996). Semidefinite programming. *SIAM Review. A Publication of the Society for Industrial and Applied Mathematics 38*(1), 49–95.

Vapnik, V.N. (1998). *Statistical Learning Theory*. New York: Wiley.

Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory* (2nd ed.). New York: Springer-Verlag.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer. ISBN 0-387-95457-0.

Venables, W. N. and Ripley, B. D. (2002b). *MASS: Modern Applied Statistics with S*.

Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models in Practice*. New York: Springer-Verlag.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Wager, C., Vaida, F., and Kauermann, G (2007). Model selection for P-spline smoothing using Akaike information criteria. *Australian & New Zealand Journal of Statistics 49*, 173–190.

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

Wakefield, J.C., Best, N.G., and Waller, L.A. (2000). Bayesian approaches to disease mapping. In Elliott, P., Wakefield, J.C., Best, N.G., and Briggs, D.J. (Eds.), *In Spatial Epidemiology: Methods and Applications*, pp. 104–127. Oxford: Oxford University Press.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society 54*, 426–482.

Wand, M.P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics 15*(4), 443–462.

Wand, M.P. (2002). Vector differential calculus in statistics. *The American Statistician 56*, 55–62.

Wand, M.P. (2003). Smoothing and mixed models. *Computational Statistics 18*, 223–249.

Wand, M.P. (2007). Fisher information for generalised linear mixed models. *Journal of Multivariate Analysis 98*, 1412–1416.

Wand, M.P. and Jones, M.C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association 88*, 520–528.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. London: Chapman & Hall.

Wand, M.P. and Ormerod, J.T. (2008). On semiparametric regression with O'Sullivan penalised splines. *Australian & New Zealand Journal of Statistics 50*(2), 179–198.

Wang, B. and Titterington, D.M. (2003a). Lack of consistency of mean field and variational Bayes approximations for state space models. Technical report, University of Glasgow.

Wang, B. and Titterington, D.M. (2003b). Local convergence of variational Bayes estimators for mixing coefficients. Technical report, University of Glasgow.

Wang, B. and Titterington, D.M. (2004). Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In Chickering, M. and Halpern, J. (Eds.), *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*. AUAI Press.

Wang, B. and Titterington, D.M. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In Cowell, R.G. and Ghahramani, Z. (Eds.), *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 373–380. Society for AISTATS.

Wang, B. and Titterington, D.M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis 1*(3), 625–650.

Waterhouse, S., MacKay, D., and Robinson, T. (1996). Bayesian methods for mixtures of experts. In Touretzky, D.S., Mozer, M.C., and Hasselmo, M.E. (Eds.), *Advances in Neural Information Processing Systems*, Volume 8, pp. 351–357. The MIT Press.

WATERLOO MAPLE INC (2007). Maple. Waterloo (Ontario), Canada.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association 85*, 699–704.

Welham, S.J., Cullis, B.R., Kenward, M.G., and Thompson, R. (2007). A comparison of mixed model splines for curve fitting. *Australian & New Zealand Journal of Statistics 49*, 1–23.

Whittaker, E.T. and Robinson, G. (1967). The Newton-Cotes formulae of integration. section 76. In *The Calculus of Observations: A Treatise on Numerical Mathematics, 4th Edition*, pp. 152–156. Dover.

Wilcox, R.R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*. San Diego: Academic Press.

Williams, D., Liao, X., Xue, Y., and Carin, L. (2005). Incomplete-data classification using logistic regression. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 977–984.

Winn, J. and Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research 6*, 661–694.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika 80*(4), 791–795.

Wolfram Research, Inc. (2007). Mathematica. Champaign, IL.

Wood, S.N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society, Series B 65*, 95–114.

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

Wood, S.N. (2006b). *mgcv: GAMs with GCV smoothness estimation and GAMMs by REML/PQL*. R package version 1.3.

Yau, P., Kohn, R., and Wood, S. (2003). Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *Journal of Computational & Graphical Statistics 12*, 1–32.

Young, G.A. and Smith, R.L. (2005). *Essentials of Statistical Inference*. Cambridge, U.K.: Cambridge University Press.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* 4(1), 57–74.

Zhao, Y., Staudenmayer, J., Coull, B.A., and Wand, M.P. (2006). General Design Bayesian Generalized Linear Mixed Models. *Statistical Science 21*, 35–51.

Zhou, S. and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association 96*, 247–259.

Zhu, Z. and Fung, W.K. (2004). Variance component testing in semiparametric mixed models. *Journal of Multivariate Analysis 91*(1), 107–118.

In Ub.