UNIVERSITY OF GLASGOW

Spatiotemporal Modelling of Groundwater Contaminants

by

Daniel Alberto Molinari

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

School of Mathematics and Statistics

December 2014

Declaration of Authorship

I, DANIEL ALBERTO MOLINARI, declare that this thesis titled, 'SPATIOTEMPORAL MODELLING OF GROUNDWATER CONTAMINANTS' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"Considerate la vostra semenza:
fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza"
Dante Alighieri
La Divina Commedia
Inferno - Canto XXVI - 118-120

"Consider the seed from which you sprang;
You were not created to live the life of brutes,
But virtue to pursue and knowledge high"
Dante Alighieri
The Divine Comedy
Hell - Canto XXVI - 118-120

UNIVERSITY OF GLASGOW

Abstract

School of Mathematics and Statistics

Doctor of Philosophy

by Daniel Alberto Molinari

Spatiotemporal data have become very common, particularly through environmental settings where a spatial array of sampling sites generates data over time. This thesis deals with a specific spatio-temporal setting of groundwater contamination and aims to construct suitable statistical models. One of the motivating features of the application is that the model has to be implemented in an unsupervised manner and there is a high premium on the results being available very quickly, with a response time of a few seconds only.

Many routes to spatiotemporal models are possible, but in order to achieve the aims outlined above we have proposed a model based on P-splines. A Bayesian approach to fitting is used to provide the stability required in an unsupervised setting. The speed requirement makes computationally intensive methods such as MCMC unsuitable for the determination of the optimal penalisation parameter and so conjugate priors and highly efficient methods of linear algebra have been brought to bear.

Use of the model identified a problematic issue due to the irregular spatio-temporal design of some data sets, giving rise to cases of "ballooning", where unexpectedly high predictions, not supported by the observations, can appear. This matter was also tackled within the Bayesian framework mentioned above. The proposed procedures were assessed both by means of a simulation study and on real data.

Finally, as an extension of the proposed methodology, we address the issue of non-detects, namely observations which are known only to lie below some limit of detection. The task is accomplished using a Laplace-type approximation to the

posterior distribution of the parameters and the suitability of this approximation is analysed through examples.

The problems addressed in the thesis are motivated by the need to ensure environmental quality in and around installations operated by the multinational company Shell. The assistance of Shell in advising on the context of the issues, and in providing data sets for case studies, is much appreciated.

Acknowledgements

I would like to thank my supervisors Prof. Adrian W. Bowman and Dr. Ludger Evers for their invaluable support and guidance throughout my research. I am extremely grateful for their encouragement and patience, without which the production of this thesis would not have been possible.

I would also like to acknowledge Dr. Wayne Jones and Shell for providing the funding, the data and their assistance for this project.

I would also like to thank the School of Mathematics and Statistics for contributing to the funding of my PhD.

Finally, I would like to thank to everyone in the School of Mathematics and Statistics, to all the friends I made and to all those people who made possible the present work.

Contents

List of Figures

List of Tables

Abbreviations

AIC	**A**kaike's **I**nformation **C**riterion
AICc	**A**kaike's **I**nformation **C**riterion (**c**orrected)
BIC	**B**ayesian **I**nformation **C**riterion
CV	Ordinary **C**ross-**V**alidation
EM	**E**xpectation **M**aximisation
FLOPS	**Fl**oating Point **Op**erations
GCV	**G**eneral **C**ross **V**alidation
GWSDAT	**G**round**W**ater **S**patio-Temporal **D**ata **A**nalysis **T**ool
IG	**I**nverse **G**amma distribution
MAP	**M**aximum **a** **P**osteriori
MCMC	**M**arkov **C**hain **M**onte **C**arlo
ML	**M**aximum **L**ikelihood
MLE	**M**aximum **L**ikelihood **E**stimates
MSE	**M**ean **S**quared **E**rror
REML	**RE**stricted **M**aximum **L**ikelihood
SVM	**S**upport **V**ector **M**achine

To my family and Rivkah

Chapter 1

Introduction

Spatiotemporal data have become ubiquitous. In some settings this has been driven by the development of affordable technology for data collection where spatially located networks of sensors collect data over time. In environmental monitoring multiple sensors are routinely used to gather data over time, in air, water or land settings. Brain imaging using EEG (electro-encephalography) or MEG (magneto-encephalography) is another example where around 200 sensors each record brain signals at very high time resolution, generating large volumes of data. In many scientific contexts, measurements are increasingly made automatically, leading to high resolution data with a strong degree of regularity, while on other occasions visits to sites of interest by trained personnel may be required, leading to sparser and more irregular data patterns.

Models for the analysis and interpretation of spatiotemporal data have developed rapidly to match the demands of the data now available and the underlying questions. Sometimes prediction is the aim while on other occasions interest can be directed at assessing the mean levels of the measurement and evidence for change over time. Banerjee et al. (2004), Finkenstädt et al. (2007) and Cressie and Wikle (2011) provide excellent entry points to the large literature on spatiotemporal modelling, with the last book very helpfully giving coverage of modern hierarchical and dynamic methods in both breadth and depth. These models are usually implemented in a Bayesian setting. In the wider literature, a unifying theme is the expectation that the spatial and temporal patterns exhibited will not follow simple parametric forms, so that models which can express flexible, but generally

smooth, shapes are required. One approach is to apply flexible forms of regression, described for example by Wood (2006), in the spatiotemporal setting. Bowman et al. (2009) take this approach to the modelling of sulphur dioxide over Europe throughout the 1990's. P-splines, described by Eilers and Marx (1996), and more general regression splines, offer a very interesting approach through the use of relatively low-dimensional sets of basis functions and Lee and Durbán (2011) apply this to the spatiotemporal modelling of ozone over Europe. The formulation of P-splines offers an interpretation in terms of mixed effects and Ruppert et al. (2003) showed the wide range of settings to which these models can be applied when the random effect interpretation is appropriate. A fully Bayesian P-splines model was introduced by Lang and Brezger (2004), with inference carried out by MCMC. Fahrmeir et al. (2004) adopted a model of this type in the specific setting of spatiotemporal data, with an empirical Bayes approach which returns again to a mixed-model representation. Brezger and Lang (2006) provided a wider range of models and efficient updating schemes while Brezger and Lang (2008) discussed simultaneous probability statements for Bayesian p-spline models, again in the context of MCMC implemetation. More recently, Wood (2011) explored the REML approach in detail and developed a fast implementation in a generalised linear modelling framework.

The context of the application discussed in this work is the monitoring of contamination in groundwater. It is clearly important to assess water quality and its associated risks to human health and the wider environment, and in particular to detect sudden increases in contaminant concentration due to possible releases. The contaminants in the groundwater are measured using water samples collected from wells and sent for subsequent lab analysis. The practicalities and cost of this inevitably lead to irregularity in time and also in space, even when operating within a fixed set of sampling locations determined by the well positions. The data collection and assessment activity is generally undertaken by staff who have science or engineering background, but may not have had advanced training in statistical methods. However it is impractical that results should always be referred back to others for statistical analysis and so there is a practical need for statistical tools that can be implemented easily and robustly as a routine part of the work of those environmental professionals. The analysis therefore needs to be fully automatic and to be fast to carry out in an unsupervised setting, but

also to produce results which are reliable, informative, and aid robust project decision-making.

The aim of the present research is to address these issues. In order to allow the construction of flexible models over space and time, P-splines are used because of their ability to provide compact representations and to express smoothness control in simple forms, as described in chapter 2. A fully Bayesian spatiotemporal model is introduced in chapter 3 using conjugate priors to avoid the need for MCMC implementation. In particular, the issue of selecting the degree of smoothness in the model is also addressed in order to produce a fully automatic procedure. A focus will be on issues of "ballooning", where predictions can be high in areas where there is no data, and this is identified and addressed by appropriate choices of the number of basis functions and the type of smoothness penalty used. The need for speed is addressed through matrix decompositions which enable the parameter which controls smoothness to be separated out from the computationally intensive parts of the calculation.

In order to assess the practical effectiveness of the approach proposed, a simulation study is performed in chapter 4. A comparison with other model selection criteria shows that very good results can be achieved following our fully Bayesian model.

Chapter 5 discusses the same topics working on real cases provided by Shell. The strategy proposed for dealing with the issue of ballooning is also shown to be effective in the cases under analysis.

An extension of our Bayesian model is studied in Chapter 6 in order to deal with *non-detect* data, i.e. data for which it is only known to be below a certain detection limit. The basic idea consists of trying to approximate the posterior distribution of the parameters by using a Laplace-type approximation approach.

An overall discussion of the present work, in which possible future developments are suggested, is the matter of the last chapter 7.

To finish, it should be stated that much of this research has led to a manuscript (see Bowman et al., 2013) submitted for publication. In particular, the work described in chapter 4 is based on data published in the aforementioned paper.

Chapter 2

Theoretical Framework

2.1 Non-parametric regression methods

Given a set of observations $(\boldsymbol{x}_i,\, y_i) = \big((x_{i1},\, \ldots,\, x_{im}),\, y_i\big)\ i = 1, \ldots, n,$ the objective of regression methods is to model the response variable y_i as a function of the predictors $\boldsymbol{x}_i,$ allowing point and interval predictions for future values of the covariates $\boldsymbol{x}$ as well as for the unknown fixed parameters involved in the model. In general these models have the form

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i \tag{2.1}$$

where $f(\boldsymbol{x}_i)$ represents the deterministic relationship between predictors and response. The uncertainty due to random variation is accounted for by ε_i with the assumptions $\mathbb{E}(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2$ and mutual independence.

Parametric approaches proceed by proposing a known form for the function $f(\boldsymbol{x})$ which traditionally is linear in the unknown parameters. In this case their estimation is carried out using methods such as least squares. Likelihood inference requires an additional assumption on the distribution of the random vector of errors. Generally it is assumed that $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0},\, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ is a positive definite covariance matrix (typically $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$).

But if these fairly strong assumptions fail, predictions, estimations and their associated degrees of certainty can be unrealistic and misleading. A more general approach relaxes the conditions on $f(\boldsymbol{x})$ by simply requiring it to be a smooth function. Typical smoothing regression techniques such as *local linear regression* rely on the estimation of a non-parametric smooth density or *kernel* function.

We will describe briefly such technique and mention some drawbacks associated with non-parametric smoothing. A short account on a commonly used model in the spatio-temporal context called *kriging* will follow. Finally, we will address a particular smoothing regression approach known as *splines*, on which most of the present work relies on.

2.2 Kernel density estimation

Statistical inference aims at inferring something about the underlying process which generated a given set of data $x_1, \ldots, x_n$.

The traditional approach is to assume that these observations follow a known parametric density, whose parameters need to be estimated on the basis of the available data.

Non-parametric smoothing methods are a more general although generally less powerful approach which simply impose a smooth pattern to the underlying density function. The starting point is to construct a suitable density function by "smoothing out" the histogram built from the given data.

By considering the definition of a (continuous) probability density function for a random variable X

$$\varphi(x) \quad = \quad \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

we can construct the approximate density function based on the observed data $x_1, \ldots, x_n$ as

$$\begin{aligned} \hat{\varphi}(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\frac{I\left\{\frac{|x-x_i|}{h}<1\right\}}{2} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\Big(\frac{|x-x_i|}{h}\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n} w(x-x_i,\, h) \end{aligned} \tag{2.2}$$

The function $K(x)$ is known as the *kernel function*. In this case, it corresponds to $K(x) = \frac{1}{2}I\{|x|<1\}$ but as a generalization any (generally symmetric) probability density function can be used. Table 2.1 lists some typical kernel functions.

Kernel function	$K(x)$
Epanechnikov	$\frac{3}{4\sqrt{5}}\left(1-\frac{x^2}{5}\right) I\{\lvert x\rvert<\sqrt{5}\}$
Biweight	$\frac{15}{16}\left(1-x^2\right)^2 I\{\lvert x\rvert<1\}$
Triweight	$\frac{35}{32}\left(1-x^2\right)^3 I\{\lvert x\rvert<1\}$
Triangular	$(1-\lvert x\rvert)\, I\{\lvert x\rvert<1\}$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$
Rectangular	$\frac{1}{2}I\{\lvert x\rvert<1\}$

TABLE 2.1: Some typical kernel functions

The graphical representation of $\hat{\varphi}(x)$ provides a more realistic idea of the distribution of the data than that of an assumed parametric density. We can tell for

instance whether it presents multimodality, skewness or heavy tails. This density representation becomes more accurate as h decreases and n increases jointly.

Nevertheless, it should be noticed that the smoothness of $\hat{\varphi}(x)$ is determined by the smoothing parameter h which reflects the unavoidable tradeoff between bias and variance. A small value of h will yield an undersmoothed density function with a low bias but high variance. Conversely, a large value of h will produce an oversmoothed curve with low variance but large bias. In the first case, the resulting density will be too "bumpy" whereas in the second it will be too smooth.

Whereas the the kernel function adopted does not have a great impact on the efficiency, the choice of the optimal value of h is a matter of crucial interest and it is tackled using elements of asymptotical theory (see e.g. Silverman, 1986, for details).

Simonoff (1996) points out some weak points of kernel density estimation and therefore source of potential problems when used in non-parametric smoothing.

Firstly, it is prone to boundary bias when the domain of the data is not unbounded. This is typically the case when the data are non-negative. In this case the kernel formulation may produce values biased downward near the origin.

Secondly, ordinary kernel estimation does not allow for different levels of smoothing at different points of the domain, as the smoothing parameter h is unique. Nevertheless, equation (2.2) can be generalised by allowing larger values of the smoothing parameter for regions with a low density of points and smaller values of h otherwise.

Finally, the bias of the kernel estimator often tends to flatten peaks and valleys of the density.

2.3 Local linear regression

The function $w(x - x_i,\, h) = \frac{1}{h}K\big(\frac{x-x_i}{h}\big)$ is generally referred to as the *weight* function and is central in the context of flexible regression. The underlying idea is that closer points should be more alike and hence ought to be given higher weights. The idea is to fit locally a proposed function $f(x)$.

Local linear regression proceeds by solving the least squares problem

$$\min_{\alpha,\,\beta} \sum_{i=1}^{n} \big\{y_i - \alpha - \beta(x - x_i)\big\}^2 w(x - x_i,\, h)$$

and taking $\hat{f}(x) = \hat{\alpha}(x)$. The local linear approach can be easily extended to two dimensions. Typically, this is the case when the response variable y_i is defined over geographical coordinates $(x_{1i},\, x_{2i})$ $i = 1, \ldots, n$. Here the weighted least squares objective function to be fitted is

$$\min_{\alpha,\beta,\gamma} \sum_{i=1}^{n} \big\{y_i - \alpha - \beta(x_1 - x_{1i}) - \gamma(x_2 - x_{2i})\big\}^2 w(x_1 - x_{1i},\, h_1) w(x_2 - x_{2i},\, h_2)$$

and again $\hat{\alpha}(x_1,\, x_2)$ is taken to be the value of the fitted surface at $(x_1,\, x_2)$.

2.4 Kriging

Given n observations over space and time $Y(\boldsymbol{s}_i,\, t_{ij})$ $\boldsymbol{s}_i \in \mathbb{R}^2$, $t_{ij} \in \mathbb{R}$ $i = 1, \ldots, I$, $j = 1, \ldots, T_i$ with $n = \sum_{i=1}^{I} T_i$, the kriging method proposes the model

$$Y(\boldsymbol{s}_i, t_{ij}) = Z(\boldsymbol{s}_i, t_{ij}) + \varepsilon(\boldsymbol{s}_i, t_{ij}) \qquad \text{with} \qquad \boldsymbol{K} = Var(\boldsymbol{Y}) = \boldsymbol{\Sigma_Z} + \sigma^2 \boldsymbol{I}_n \tag{2.3}$$

where $\boldsymbol{\Sigma_Z}(\boldsymbol{s}_1, \boldsymbol{s}_2, t_1, t_2) = Cov\left(\boldsymbol{Z}(\boldsymbol{s}_1, t_1), \boldsymbol{Z}(\boldsymbol{s}_2, t_2)\right)$ corresponds to a (semi) positive definite spatio-temporal covariance matrix and $\boldsymbol{\varepsilon}$ represents the vector measurement errors such that $\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}_n$ and $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$ with $\boldsymbol{Z}$ and $\boldsymbol{\varepsilon}$ independent.

Distances in space and distances in time should be treated differently in the spatio-temporal covariance matrix $\boldsymbol{\Sigma_Z}$. Cressie and Wikle (2011) suggest examples of possible structures for such matrices.

The objective is to predict $\eta = Y(\boldsymbol{s}_0, t_0)$ for a given spatio-temporal location $(\boldsymbol{s}_0, t_0) \in \mathbb{R}^3$. *Simple kriging* proceeds by assuming that

$$\mu(\boldsymbol{s}, t) = \mathbb{E}\left(\boldsymbol{Z}(\boldsymbol{s}, t)\right) \tag{2.4}$$

is known $\forall \ s, t$. Let us call

$$\boldsymbol{k}_0 = Cov\left(\boldsymbol{Y}, Y(\boldsymbol{s}_0, t_0)\right) \tag{2.5}$$

$$c_0 = Var\left(Y(\boldsymbol{s}_0, t_0)\right) \tag{2.6}$$

If $\boldsymbol{Z}(\boldsymbol{s}, t)$ and $\boldsymbol{\varepsilon}(\boldsymbol{s}, t)$ are assumed to be Gaussian processes then it holds that

$$\begin{pmatrix} \boldsymbol{Y} \\ Y(\boldsymbol{s}_0, t_0) \end{pmatrix} \sim \mathcal{N}_{n+1}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \mu(\boldsymbol{s}_0, t_0) \end{pmatrix}, \begin{pmatrix} \boldsymbol{K} & \boldsymbol{k}_0 \\ \boldsymbol{k}_0' & c_0 \end{pmatrix} \right) \tag{2.7}$$

where $\boldsymbol{\mu} = \mathbb{E}\left(\boldsymbol{Z} + \boldsymbol{\varepsilon}\right) = \mathbb{E}\left(\boldsymbol{Y}\right)$ is the vector of known expectations of the observed values. Under these premises

$$Y(\boldsymbol{s}_0, t_0)|\boldsymbol{Y} \quad \sim \quad \mathcal{N}\left(\mu(\boldsymbol{s}_0, t_0) + \boldsymbol{k}_0'\boldsymbol{K}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}),\, c_0 - \boldsymbol{k}_0'\boldsymbol{K}^{-1}\boldsymbol{k}_0\right) \tag{2.8}$$

A sensible choice for $\hat{\eta}(\boldsymbol{Y})$, the estimator of $\eta = Y(\boldsymbol{s}_0, t_0)$, is the statistic minimising $MSE(\hat{\eta}) = \mathbb{E}\big((\hat{\eta} - \eta)^2\big)$. It can be shown (see e.g. Diggle and Ribeiro, 2007) that such estimator is $\hat{\eta} = \hat{\eta}(\boldsymbol{Y}) = \mathbb{E}\left(\eta|\boldsymbol{Y}\right)$. Therefore, taking into account (2.8)

$$\hat{\eta} \quad = \quad \mu(\boldsymbol{s}_0, t_0) + \boldsymbol{k}_0'\boldsymbol{K}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{2.9}$$

$$\text{and} \qquad Var\left(\hat{\eta}\right) \quad = \quad c_0 - \boldsymbol{k}_0'\boldsymbol{K}^{-1}\boldsymbol{k}_0 \tag{2.10}$$

Ordinary kriging is a more general approach which assumes that $\mu(\boldsymbol{s}, t) = \mathbb{E}\left(\boldsymbol{Z}(\boldsymbol{s}, t)\right) = \mu$ is constant but unknown. Cressie and Wikle (2011) show that the generalised-least squares estimator of μ is $\hat{\mu} = \dfrac{\mathbf{1}'\boldsymbol{K}^{-1}\boldsymbol{Y}}{\mathbf{1}'\boldsymbol{K}^{-1}\mathbf{1}}$ where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones. In this case, $\hat{\eta}$ is obtained by replacing $\mu(\boldsymbol{s}, t)$ by its estimator $\hat{\mu}$ in (2.9)

$$\hat{\eta} \quad = \quad \hat{\mu} + \boldsymbol{k}_0'\boldsymbol{K}^{-1}(\boldsymbol{Y} - \hat{\mu}\mathbf{1}) \tag{2.11}$$

We shall only mention that the predictor $\hat{\eta}$ can be derived under even more general conditions, assuming that μ is a linear combination of covariates. This approach is known as *universal kriging* (see e.g. Cressie, 1993, for details).

A noticeable drawback of kriging is that the estimations require the inversion of the $n \times n$ matrix $\boldsymbol{K}$ which may be very time consuming. If $T_i = T \;\; \forall \;\; i$, a more efficient computation results under the assumption of separability of space and time when modelling $\boldsymbol{K}$. If we consider that the covariance matrix can be written as

$$\boldsymbol{K} = \boldsymbol{K}^{(s)} \otimes \boldsymbol{K}^{(t)}$$

where $\otimes$ indicates the Kronecker product, $\boldsymbol{K}^{(s)}$ is an $I \times I$ covariance matrix of purely spatial covariances and $\boldsymbol{K}^{(t)}$ is a $T \times T$ covariance matrix of purely temporal covariances then

$$\boldsymbol{K}^{-1} = \left(\boldsymbol{K}^{(s)}\right)^{-1} \otimes \left(\boldsymbol{K}^{(t)}\right)^{-1}$$

then the computation of $\boldsymbol{K}^{-1}$ involves inverting matrices of dimension $I \times I$ and $T \times T$ which are much smaller than $IT \times IT$.

2.5 P-splines

2.5.1 Introduction

Splines come from the field of numerical analysis. They were initially used for constructing smooth interpolating functions. Splines consist of piecewise polynomials connected by points called *knots*. There are two different types of smoothing techniques using splines (see e.g. Durbán, 2009)

- **Regression splines**: In these models it is necessary to select the number and location of the knots in such a way that the smoothness of the fitted function can be adjusted. Additionally, restrictions must be imposed leading to a smooth connection between the adjacent polynomial pieces. The model is then fitted using least squares.
- **Smoothing splines**: They arise as the solution to a non-parametric regression problem in which it is desired to find a function (with two continuous derivatives) minimising the *penalised sum of squares:*

$$PSS = \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2 + \lambda \int_x \left(f''\left(x\right)\right)^2 dx$$

  where the last term represents a penalisation on the second derivative of the curve and λ is the *smoothing parameter* controlling the smoothness of such

curve. If $\lambda = 0$ the smoothing spline reduces to an interpolation function whereas, if $\lambda \to \infty$, the second derivative is constrained to tend to zero, yielding an estimator which approaches a linear fit.

It should be mentioned that according to the *Representer Theorem*, the solution to a smoothing spline is a regression spline with knots at every observation (see e.g. Hastie et al., 2009).

However, both techniques present some drawbacks. The smoothness of the fitted function in the regression splines depends on the choice of the knots and involves using somewhat awkward algorithms which are difficult to extend to the multidimensional case. As for the smoothing splines, the issues are of a computational nature, as they use as many knots as the number of observations.

Penalised splines or **P-splines**, introduced by Eilers and Marx (1992, 1996, 2010), are a trade-off between both techniques combining the best of these approaches. This approach to smoothing has become widely spread because of its simple representation of the function of interest.

P-splines use fewer parameters than smoothing splines, but the choice of the knots is not as significant as for regression splines. P-splines use a number of knots which is far smaller than the dimensionality of the data and they are computationally efficient in particular when the sample size is very large. In addition, the use of the penalisation relaxes the importance of the choice and location of the knots.

Essentially, the methodology of P-splines is based on

(a) Adopting a convenient basis of functions for the regression,

(b) Modification of the likelihood function by introducing a penalisation based on the differences of adjacent coefficients.

2.5.2 The P-spline approach

For a given set of pairs $(x_i,\, y_i) \quad i = 1,\, \ldots,\, n$ of response-covariates points, we can describe their dependence by means of a model $y_i = f(x_i) + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}\,(0,\, \sigma^2)$ and $f(x)$ describes a non-parametric regression function whose shape is largely unconstrained. It is important to underline that this approach should only be used to fit data which do not present abrupt changes in the signal.

A convenient representation for the non-parametric smooth regression function $f(x)$ is a linear combination of a conveniently chosen set of basis functions $b_j(x)$ as $f(x) = \sum_{j=1}^{m} \alpha_j b_j(x)$. By modifying the values of the coefficients α_j a huge range of smooth functions can be created.

In more detail, let us suppose we are given n points (x_i, y_i) and we consider the regression model

$$\boldsymbol{Y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon} = \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \qquad \text{where} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\boldsymbol{0},\, \sigma^2 \boldsymbol{I}_n\right) \tag{2.12}$$

$\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{x})$ is a $n \times m$ matrix of the form

$$\boldsymbol{B} = \begin{pmatrix} b_1(x_1) & \cdots & b_j(x_1) & \cdots & b_m(x_1) \\ b_1(x_2) & \cdots & b_j(x_2) & \cdots & b_m(x_2) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ b_1(x_i) & \cdots & b_j(x_i) & \cdots & b_m(x_i) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ b_1(x_n) & \cdots & b_j(x_n) & \cdots & b_m(x_n) \end{pmatrix}$$

where the $b_j(x) \;\; j = 1, \ldots m$ represent the m functions making up the basis and $\boldsymbol{\alpha}$ is an m-dimensional vector of parameters to be estimated.

Although the functions $b_j(x)$ may be computed in several different ways, the typical choice falls upon **B-splines**, as they are generally more stable than other bases and they can be efficiently constructed from polynomial pieces.

- A **B-spline of order p** consists of $p+1$ pieces of a polynomial of order p,
- These pieces are connected by p internal knots,
- The derivatives up to the order $p-1$ are continuous on these knots,
- The B-spline is positive in the domain spanned by $p+2$ knots and 0 otherwise,
- Except on the extremes, it overlaps with $2p$ pieces of the polynomials on its neighborhood and
- For each value of x, there are $p+1$ splines which are not null at x.

Usually B-splines of order $p = 3$ are considered. All these 3-order polynomials have the same shape but they are horizontally shifted. This shifting depends on the distance between knots. Figure 2.1 depicts this situation for a B-spline of order 3. Each function in the basis is made up of 4 pieces of polynomials of order 3 connected by 3 internal knots. As mentioned, the choice and location of the knots is not determined in advance as happens with smoothing splines. It suffices to take a sensible large number (> 20, for instance) of equidistant knots.

2.5.3 Penalisation

If, for a given basis $\boldsymbol{B}$, we use least squares to fit the model, the objective function to be minimised is

$$S(\boldsymbol{\alpha}) = \|\boldsymbol{y} - \boldsymbol{B\alpha}\|^2 = (\boldsymbol{y} - \boldsymbol{B\alpha})'(\boldsymbol{y} - \boldsymbol{B\alpha})$$

yielding $\hat{\boldsymbol{\alpha}} = (\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{y}$ and hence the fitted curve $\hat{f}(\boldsymbol{x}) = \boldsymbol{B}\hat{\boldsymbol{\alpha}}$ will depend on the basis size. Eilers and Marx (1996) proposed to use a dense set of basis functions. The counterpart of this proposal is that in addition to the signal also the noise tends to be fitted, yielding always a wigglier function as the number of basis functions increases. The extreme situation corresponds to an equal number of knots and points, in which case the fitting curve interpolates the data.

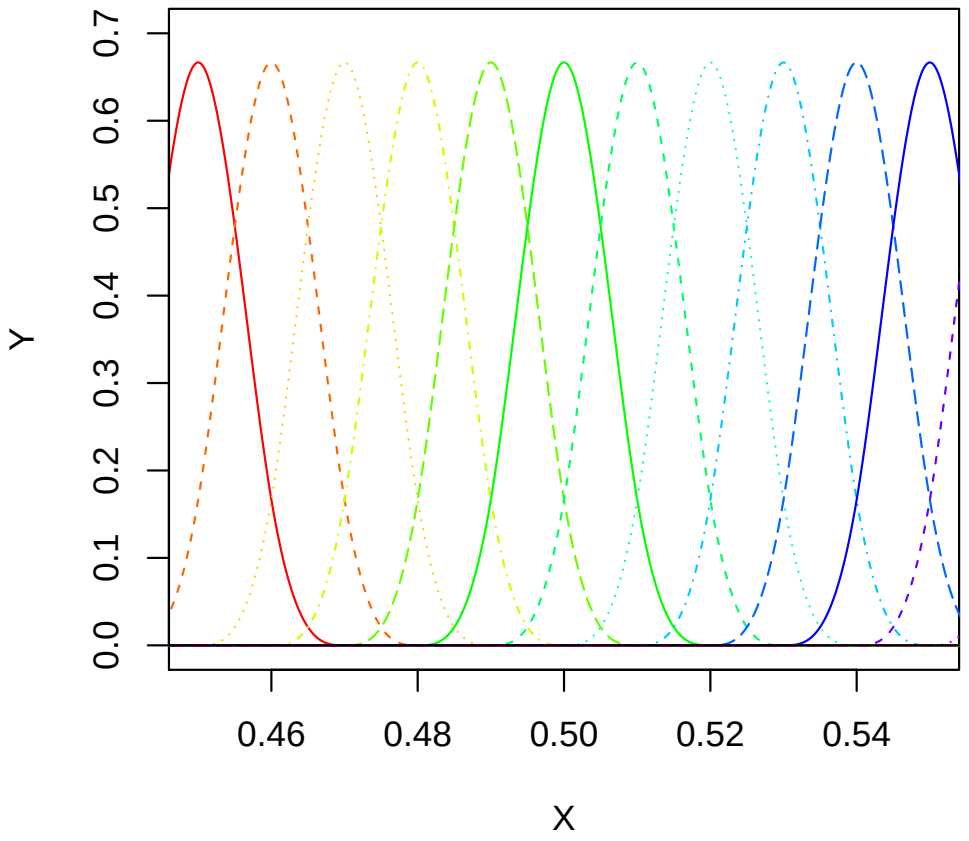


FIGURE 2.1: B-spline basis of order 3

In order to avoid this circumstance O'Sullivan (1986) introduced a penalisation on the second derivative of the curve, and hence the objective function to be minimised turned out to be

$$\begin{aligned} S(\boldsymbol{\alpha}, \lambda) &= \|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}\|^2 + \lambda \int_{\boldsymbol{x}} \left(\boldsymbol{B}''(\boldsymbol{x})\boldsymbol{\alpha}\right)^2 d\boldsymbol{x} \\ &= (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}) + \lambda \int_{\boldsymbol{x}} \left(\boldsymbol{B}''(\boldsymbol{x})\boldsymbol{\alpha}\right)^2 d\boldsymbol{x} \end{aligned}$$

where λ is a non-negative parameter that penalises the overall smoothness of the fitted function. Typically, a penalisation based on the second derivative is used but any order of derivatives may be employed.

Eilers and Marx (1996) advocate a penalisation based on the difference of order d between the adjacent coefficients in the bases of the B-splines. This kind of penalisation is more flexible as it is independent of the order of the polynomial used to construct the B-spline and represents a good discrete approximation to the integral of the square of the *d-th* derivative. Besides, it acts directly upon the coefficients rather than on the curve itself.

This approximation is implemented by means of the $(m-d)\times m$ difference matrix of order d, $\boldsymbol{D_d}$. Using this approach, the objective function to be minimised is

$$\begin{aligned} S(\boldsymbol{\alpha},\lambda) &= \|\boldsymbol{y}-\boldsymbol{B\alpha}\|^2+\lambda\|\boldsymbol{D_d\alpha}\|^2 \\ &= (\boldsymbol{y}-\boldsymbol{B\alpha})'(\boldsymbol{y}-\boldsymbol{B\alpha})+\lambda\boldsymbol{\alpha' D_d' D_d\alpha} \end{aligned} \tag{2.13}$$

producing $\hat{\boldsymbol{\alpha}}=\left(\boldsymbol{B'B}+\ \lambda\boldsymbol{D_d'D_d}\right)^{-1}\boldsymbol{B'y}$. Usually $d=2$ is used but other orders may be employed depending on the variability of the curve and the amount of noise in the data. For example, for a penalisation of order $d=2$, it is

$$D_2=\begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & \cdots \\ 0 & 0 & 1 & -2 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

and thus

$$\boldsymbol{\alpha' D_2' D_2\alpha} = (\alpha_1-2\alpha_2+\alpha_3)^2+\ldots+(\alpha_{m-2}-2\alpha_{m-1}+\alpha_m)^2 \tag{2.14}$$

Figure 2.2 shows the effect of penalisation: to force the coefficients to yield a smooth pattern. The fitting process of a function using B-splines is pictured with and without penalisation, together with the functions making up the basis (the columns of the B matrix). The left plot results from not penalising ($\lambda=0$) the

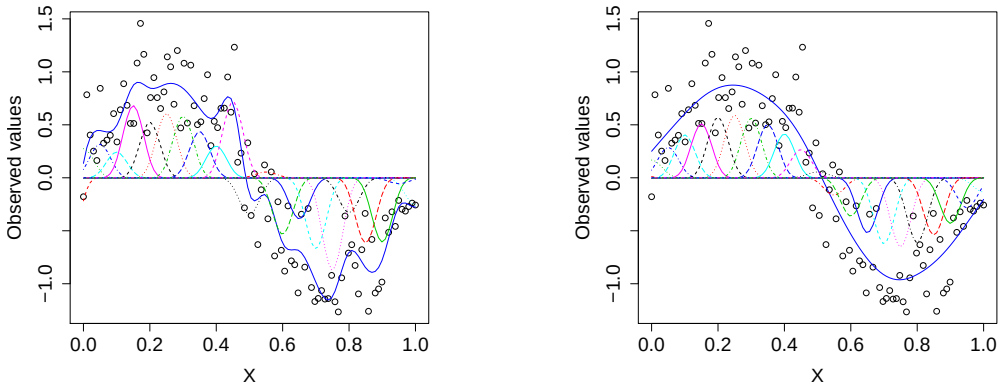


FIGURE 2.2: Curve based on 20 knots in the basis, without penalisation (left), with penalisation (right)

term in the objective function that accounts for the smoothness; it can be noticed that it yields a rather wiggly regression function. On the right plot instead, a suitable choice for λ constrains the optimisation method to find values for the coefficients $\hat{\alpha}$ which result in a smoother regression curve.

For a given value of λ, the fitted values are given by

$$\hat{\boldsymbol{y}} = \boldsymbol{B}\hat{\boldsymbol{\alpha}} = \boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}_d'\boldsymbol{D}_d\right)^{-1}\boldsymbol{B}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{2.15}$$

Although the "hat matrix" H is not a projection matrix (it is symmetric, but not idempotent) it plays a similar role as its counterpart in the linear model. Following with this analogy the trace of the hat matrix

$$p = tr(\boldsymbol{H}) \tag{2.16}$$

is known as the *degrees of freedom (df)* or *effective dimension (ED)* of the model and it can be thought of as the number of free parameters that are being estimated

giving an idea of the complexity of that model. Because it provides a more intuitive scale on which the smoothness can be expressed, the optimal regression function is sometimes described in terms of the degrees of freedom rather than by means of λ.

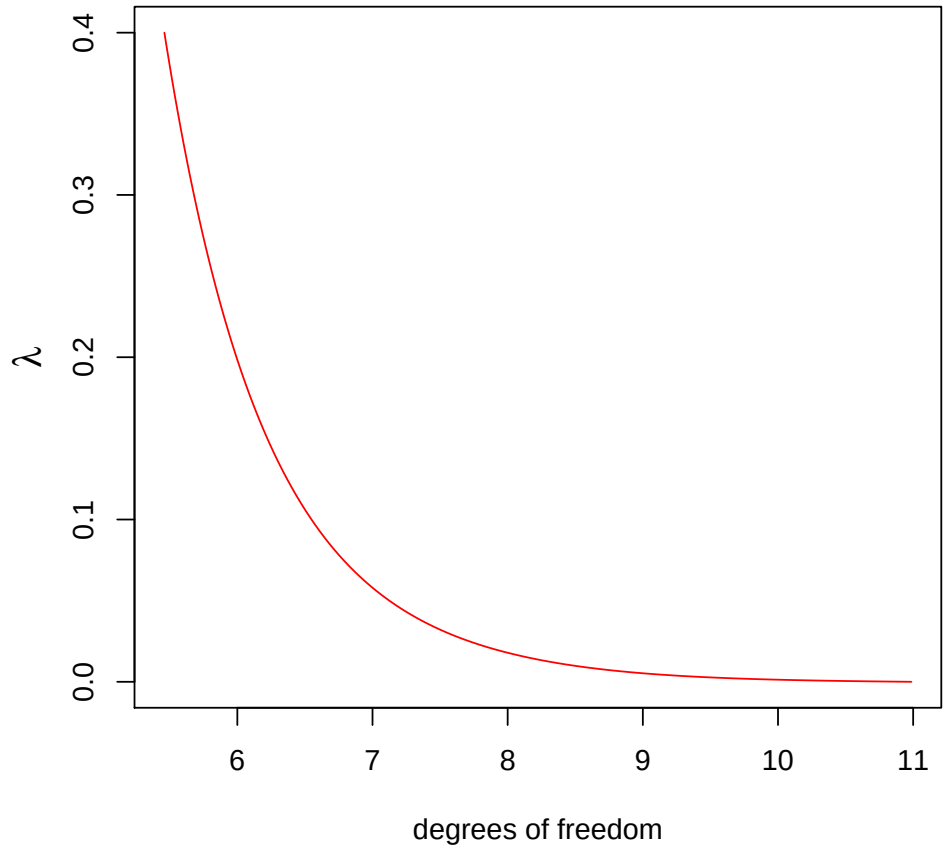


FIGURE 2.3: Effective dimension vs. λ for one-dimensional simulation

It should be noted that there is a one-to-one decreasing relationship between the degrees of freedom and the penalisation parameter, indicating that increasing the complexity of the model lessens the need for smoothing. This is shown in Figure 2.3 which corresponds to the illustrative example presented later in section 3.7, Figures 3.5, 3.6, 3.7, 3.8, 3.9, 3.10 and 3.11.

When $\lambda = 0$, the expression for the estimator of the parameters $\hat{\boldsymbol{\alpha}}$ boils down to the classical solution in linear models theory with the degrees of freedom equal

to the number of parameters m. As $\lambda \to \infty$, the fitted function tends to a linear function and the degrees of freedom tend to 2.

The variance of the fitted values can be computed as

$$Var(\hat{\boldsymbol{y}}) = Var(\boldsymbol{H}\boldsymbol{y}) = \boldsymbol{H}\,\sigma^2\boldsymbol{I}_n\,\boldsymbol{H}' = \boldsymbol{H}^2\sigma^2$$

By using $\hat{\sigma}^2 = \frac{\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2}{n-tr(\boldsymbol{H})}$ as an estimator of σ^2, an approximated 95% confidence interval for the mean of each observation y_i can be constructed as $\hat{y}_i \pm 1.96\,\sqrt{(\boldsymbol{H}^2)_{ii}}\,\hat{\sigma}$.

It is worth mentioning that the larger the value of the penalisation parameter, the larger is the bias, which vanishes when $\lambda = 0$.

The generalisation to the three-dimensional case of the above approach for a spatio-temporal model is carried out by considering now the pairs response-covariates as $(\boldsymbol{x}_i,\, y_i) = \big((x_{i1},\, x_{i2},\, t_i),\, y_i\big) \quad i = 1, \ldots, n$. The regresion function is expressed as $f(x_1,\, x_2,\, t) = \sum_j \sum_k \sum_l \alpha_{jkl} b_j(x_1) b_k(x_2) b_l(t)$ using a basis set which is simply the product of all triples of the marginal basis functions over x_1, x_2 and t. This yields a design matrix $\boldsymbol{B}$ of dimension $n \times m^3$ and a difference matrix $\boldsymbol{D_d}$ of dimension $(m-d)^3 \times m^3$.

In addition to Eilers and Marx (1996), details of these methods are also described by Ruppert et al. (2003) and Wood (2006).

As an example, Figure 2.4 depicts part of the design matrix $\boldsymbol{B}$ corresponding to a B-spline basis of order 3 for a bidimensional (spatial) case.

The P-splines approach is very flexible because of its low rank representation which can encapsulate a flexible curve conveniently. Here, the matrices to be dealt with are of dimension $m \times m$ where m represents the number of parameters to be used. This number of parameters depends on the number of basis functions which can be controlled dynamically.

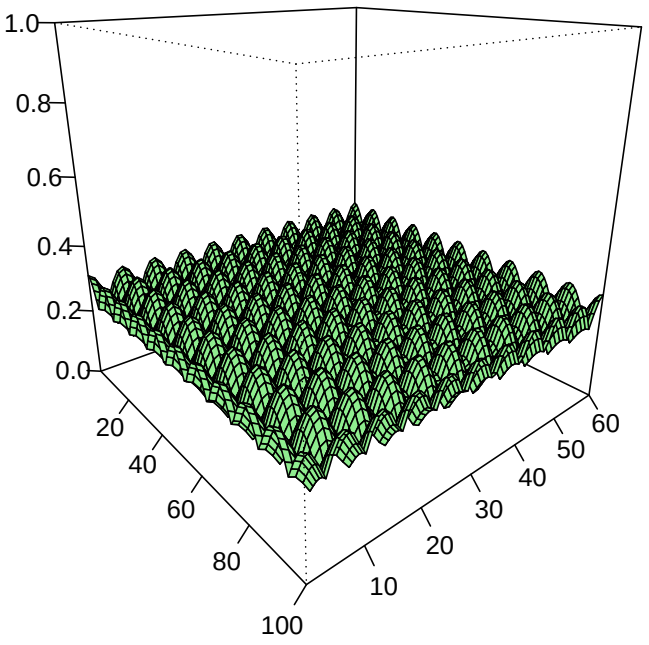


FIGURE 2.4: Bidimensional B-spline basis of order 3

In comparison with the kriging method described in section 2.4, P-splines do not require a balanced spatio-temporal design for an efficient implementation, though such design would further improve the computational efficiency of P-splines. It should be mentioned nevertheless, that if a large number of basis functions is used, there might not be a real benefit with the P-splines approach.

As we will see in section 5.1, the spatio-temporal model to be proposed needs to run fast on large data sets. In order to meet this constraint we have chosen to use P-splines for modeling the data.

2.6 P-splines regarded as a Gaussian process

If we are mostly interested in prediction rather than in the estimation of the vector of coefficients $\boldsymbol{\alpha}$, we can approach the P-splines model thinking of it as a Gaussian process.

If we consider that σ^2 is known, we can give a Bayesian interpretation to the objective formula 2.13 to be optimised with respect to $\boldsymbol{\alpha}$ under the P-splines framework. Let us assume

$$\begin{aligned} \boldsymbol{\alpha} &\sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{V}) && (2.17) \\ \boldsymbol{Y}|\boldsymbol{\alpha} &\sim \mathcal{N}_n(\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I}_n) && (2.18) \end{aligned}$$

with $\boldsymbol{V}^{-1} = \frac{\lambda}{\sigma^2}\boldsymbol{P}$ assuming for notational simplicity that $\boldsymbol{P}$ has full rank. Hence the posterior distribution of the parameters is

$$f(\boldsymbol{\alpha}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\alpha})\, f(\boldsymbol{\alpha}) \propto \exp\left\{-\frac{\|\boldsymbol{y}-\boldsymbol{B}\boldsymbol{\alpha}\|^2}{2\sigma^2}\right\} \exp\left\{-\frac{\lambda \boldsymbol{\alpha}'\boldsymbol{P}\boldsymbol{\alpha}}{2\sigma^2}\right\} \tag{2.19}$$

From 2.19 we obtain that the log-posterior distribution of the vector of parameters $\boldsymbol{\alpha}$ is

$$\log f(\boldsymbol{\alpha}|\boldsymbol{y}) = -\frac{1}{2\sigma^2}\left\{\|\boldsymbol{y}-\boldsymbol{B}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\alpha}'\boldsymbol{P}\boldsymbol{\alpha}\right\} + \text{constant}$$

which, up to a multiplicative constant, is the objective function 2.13 already mentioned. The marginal distribution of $\boldsymbol{Y}$ is the normalising constant for equation (2.19) and is given by

$$f(\boldsymbol{y}) \quad = \quad \int f(\boldsymbol{y}|\boldsymbol{\alpha})\ f(\boldsymbol{\alpha})\ d\boldsymbol{\alpha}$$

From the theory of normal distribution we know that this marginal distribution has to be normal, too. Hence, to specify it completely, it suffices to compute its expectation and its variance. It is

$$\mathbb{E}(\boldsymbol{Y}) \quad = \quad \mathbb{E}_{\boldsymbol{\alpha}}\left(\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{\alpha}}(\boldsymbol{Y})\right) = \mathbb{E}_{\boldsymbol{\alpha}}(\boldsymbol{B}\boldsymbol{\alpha}) = \boldsymbol{B}\,\mathbb{E}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \boldsymbol{0} \tag{2.20}$$

$$\begin{aligned} Var(\boldsymbol{Y}) \quad &= \quad Var_{\boldsymbol{\alpha}}\left(\mathbb{E}_{Y|\boldsymbol{\alpha}}(\boldsymbol{Y})\right) + \mathbb{E}_{\boldsymbol{\alpha}}\left(Var_{\boldsymbol{Y}|\boldsymbol{\alpha}}(\boldsymbol{Y})\right) = \\ &= \quad Var_{\boldsymbol{\alpha}}(\boldsymbol{B}\boldsymbol{\alpha}) + \mathbb{E}_{\boldsymbol{\alpha}}\left(\sigma^2\boldsymbol{I}_n\right) = \boldsymbol{B}\,Var_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})\,\boldsymbol{B}' + \sigma^2\boldsymbol{I}_n = \\ &= \quad \underbrace{\frac{\sigma^2}{\lambda}}_{\tau^2}\underbrace{\boldsymbol{B}\boldsymbol{P}^{-1}\boldsymbol{B}'}_{\boldsymbol{K}} + \sigma^2\boldsymbol{I}_n = \\ &= \quad \tau^2\,\boldsymbol{K} + \sigma^2\boldsymbol{I}_n \end{aligned} \tag{2.21}$$

Thus

$$\begin{aligned} \boldsymbol{Y} \quad &\sim \quad \mathcal{N}_n\left(\boldsymbol{0},\, \tau^2\boldsymbol{K} + \sigma^2\boldsymbol{I}_n\right) && \text{with} \\ \\ \boldsymbol{K}_{hk} \quad &= \quad \left(\boldsymbol{B}\boldsymbol{P}^{-1}\boldsymbol{B}'\right)_{hk} = \boldsymbol{b}_h'\boldsymbol{P}^{-1}\boldsymbol{b}_k && \text{and} \\ \boldsymbol{b}_h' \quad &= \quad (b_1(\boldsymbol{x}_h),\, \ldots,\, b_j(\boldsymbol{x}_h),\, \ldots,\, b_m(\boldsymbol{x}_h)) \end{aligned}$$

where $b_j(\boldsymbol{x}_h)$ represents the *j-th* basis function evaluated at the *h-th* vector of covariates. Now, let us assume that we want to predict the outcome of Y_0 for a given new vector of covariates $\boldsymbol{x}_0$. From 2.7, we have that

$$\begin{pmatrix} \boldsymbol{Y} \\ Y_0 \end{pmatrix} \sim \mathcal{N}_{n+1}\left(\begin{pmatrix} \boldsymbol{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2\boldsymbol{K} + \sigma^2\boldsymbol{I}_n & \tau^2\boldsymbol{k}_0 \\ \tau^2\boldsymbol{k}_0' & \tau^2 c_0 + \sigma^2 \end{pmatrix} \right) \quad \text{with}$$

$$\begin{aligned} (\boldsymbol{k}_0)_h &= \left(\boldsymbol{b}_0'\boldsymbol{P}^{-1}\boldsymbol{B}'\right)_h = \boldsymbol{b}_0'\boldsymbol{P}^{-1}\boldsymbol{b}_h \qquad \boldsymbol{k}_0 \in \mathbb{R}^n \quad \text{and} \\ c_0 &= \boldsymbol{b}_0'\boldsymbol{P}^{-1}\boldsymbol{b}_0 \end{aligned}$$

From 2.8, we obtain

$$Y_0|\boldsymbol{Y} \sim \mathcal{N}\left(\boldsymbol{k}_0'\left(\boldsymbol{K} + \frac{\sigma^2}{\tau^2}\boldsymbol{I}_n\right)^{-1}\boldsymbol{y},\ \tau^2\left[c_0 - \boldsymbol{k}_0'\left(\boldsymbol{K} + \frac{\sigma^2}{\tau^2}\boldsymbol{I}_n\right)^{-1}\boldsymbol{k}_0\right] + \sigma^2 \right) \tag{2.22}$$

The mean of the distribution in 2.22 is the point estimate of y_0 and the variance for a prediction interval for a new observation, is the same as the one in the aforementioned distribution.

2.7 Overview on the choice of the penalisation parameter

When trying to fit a smooth function to a given data set using the P-splines approach, the choice of the smoothing parameter λ is a crucial matter as it will determine the trade-off between smoothness and capturing the signal.

The smaller the value of λ, the less the overall curvature will be penalised; hence, the fitted function will follow the data points more closely. This is known as *undersmoothing* or *overfitting*. In this case the prediction error is underestimated and the fitted function would be very different for a new set of data from the same model.

Conversely we speak of *oversmoothing* or *underfitting* when the overall curvature is highly penalised. In this case most of the pattern due to the signal will be disregarded, leading to a flatter fitting function. Oversmoothing produces biased regression coefficients and inflation in the estimate of the variance.

2.7.1 Typical model selection criteria

The traditional approaches rely on finding the value of the penalisation parameter that minimises some model selection criteria. These criteria generally are made up of a trade-off between the complexity of the model and its goodness of fit. If we denote by p the number of free parameters in the model (or degrees of freedom), by n the number of observations, by $\hat{\sigma}^2 = \frac{1}{n-p}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$, by H the hat matrix and by L the value of the likelihood function for the estimated model, the most common selection criteria used are

- **Akaike's Information Criterion (AIC)** (Akaike, 1973)

$$AIC = -2\log(L) + 2p = n\log(\hat{\sigma}^2) + 2p$$

- **Akaike's Information Criterion (corrected) (AICc)** (see Hurvich et al., 1998)

$$AICc = \log(\hat{\sigma}^2) + 1 + \frac{2(p+1)}{n-p-2}$$

- **Bayesian Information Criterion (BIC)** (see Schwarz, 1978)

$$BIC = -2\log(L) + p\log(n) = n\log(\hat{\sigma}^2) + p\log(n)$$

- **(Ordinary) Cross-Validation (CV)** (see Wood, 2006)

$$CV = \frac{1}{n}\sum_i \left(\frac{y_i - \hat{y}_i}{1 - H_{ii}}\right)^2$$

- **Generalised Cross-Validation (GCV)** (see Wood, 2006)

$$GCV = \frac{n\hat{\sigma}^2}{n-p}$$

These definitions are up to an additive constant. In general, AIC tends to select more complex models than BIC, i.e. it is prone to overfitting/undersmoothing by favouring small values of the penalisation parameter λ. On the other side, BIC imposes a stronger penalty on the number of parameters. This leads to the choice of simpler models and hence BIC is prone to underfitting/oversmoothing by preferring large values for λ.

AICc is an improved version of AIC which aims at correcting its tendency to undersmooth.

The rationale behind CV is to leave out one data point at a time in the fitting process and use the resulting "reduced" regression function to predict the value of the omitted observation. The average of the square of the difference between the predicted and actual value of each observation is used as a measure of the goodness of fit for a particular model. Fortunately, as Wood (2006) shows, this computationally inefficient method is equivalent to the formula given above.

However, even using the stated formula can be expensive if the model depends on several smoothing parameters. This can be addressed by replacing the weights $1-H_{ii}$ by the mean weight $tr(\boldsymbol{I}-\boldsymbol{H})/n$ to obtain the GCV expression aforementioned.

Wood (2011) indicates that although asymptotically prediction error methods give better prediction error performance than likelihood-based methods, the convergence of smoothing parameters to their optimal values may be slow and hence are prone to occasional severe undersmoothing by yielding small values for the penalistation parameter λ.

Figure 2.5 shows a comparison among the different criteria for model selection. A one-dimensional model is fitted using P-splines to data corresponding to the concentration of a contaminant over time at a fixed location (see chapter 5 for details). In this case, the degrees of freedom rather than the value of the penalisation parameter λ are used as a reference.

It can be noticed that the smaller the effective dimension, the narrower is the resulting interval and the flatter the fitted curve. Whereas the first four cases

seem to produce a fairly similar fitting, the cross-validation technique tends to yield greater variability in the estimated curve.

2.7.2 Shortcomings

We have already mentioned that traditional model selection criteria are prone to overfitting except for BIC which, on the contrary, tends to yield values for the penalisation parameter which are larger than the optimal ones.

For model selection criteria prone to undersmoothing, there is another issue of importance to be considered. It refers to extremely high unexpected predicted values where there is no data supporting these outcomes. This undesired effect is known as *ballooning*.

Figure 2.6 provides an example of this situation. It corresponds to a simulation which will be described in detail in chapter 4. For the moment, we can say that we know the true values of the measurements over a certain region at a given point in time. The top-left plot of the figure depicts the actual situation.

After the addition of some noise, this surface can be estimated using AIC, AICc, CV and GCV for model selection criteria. The high peaks in the predictions are clearly inappropriate.

Because these predictions seem rather implausible, we must explore another method for choosing the penalisation parameter in order to avoid ballooning. In the next chapter, we are going to develop a method based on the Bayesian framework to try to solve the problem.

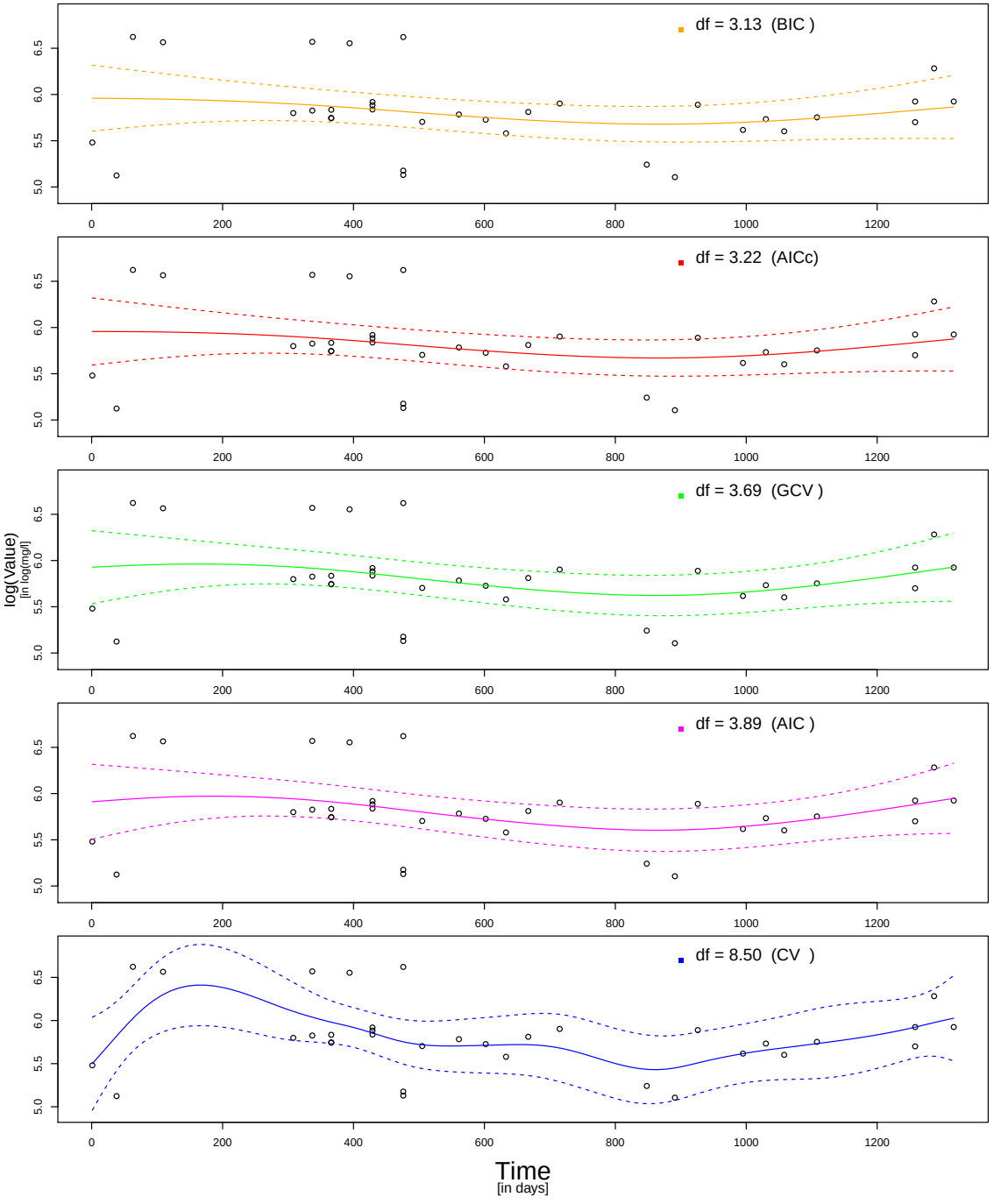
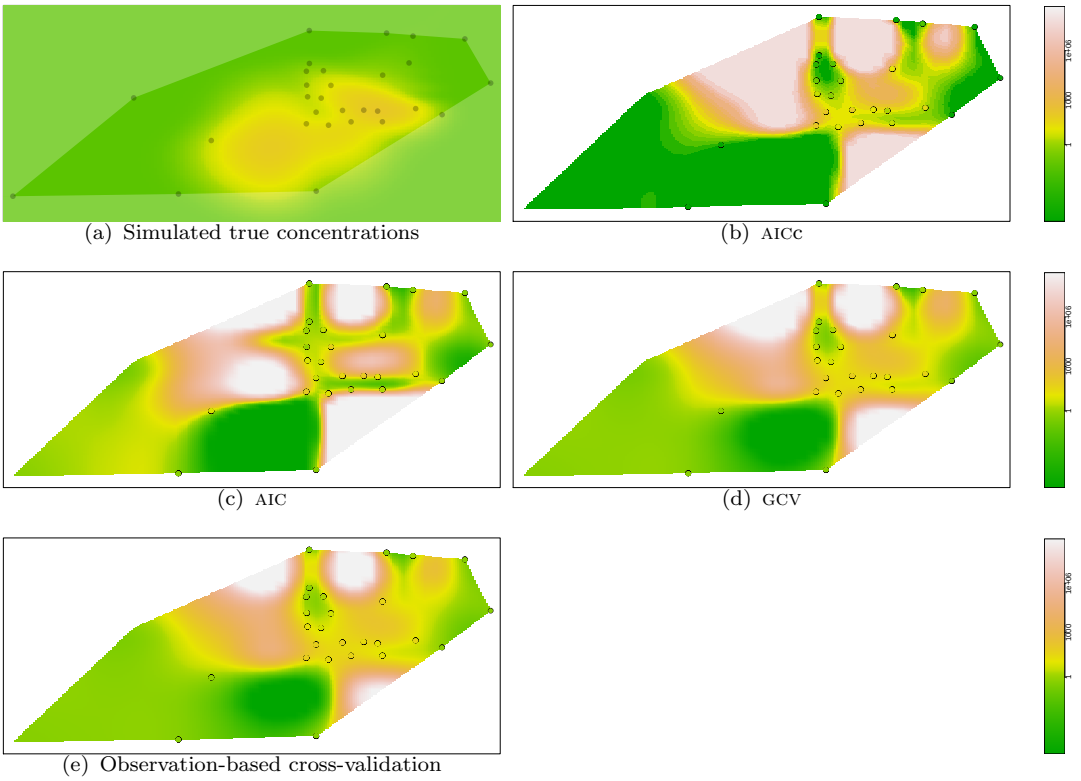


FIGURE 2.5: Model choice for one-dimensional P-spline fitting by minimising different criteria



(a) Simulated true concentrations (b) AICc

(c) AIC (d) GCV

(e) Observation-based cross-validation

FIGURE 2.6: Simulated true model (top left) and related predictions obtained using different model selection criteria for choosing the smoothing parameter

Chapter 3

Efficient Bayesian determination of the penalisation parameter

3.1 Introduction

In this chapter we will set out the Bayesian approach for selecting the smoothing parameter. Let M_λ be the model for a particular value of the penalisation parameter λ. We are interested in computing $f_{M_\lambda|\boldsymbol{Y}}$, the posterior distribution of the model M_λ given the data $\boldsymbol{Y} = \boldsymbol{y}$.

In our initial set-up, we assume that the observation model is $\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma^2$, $M_\lambda \sim \mathcal{N}_n(\boldsymbol{B\alpha}, \sigma^2 \boldsymbol{I_n})$ and hence

$$f_{\boldsymbol{Y}|\boldsymbol{\alpha},\sigma^2,M_\lambda} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{ -\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{B\alpha})'(\boldsymbol{y}-\boldsymbol{B\alpha})\Big\} \tag{3.1}$$

with $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{B} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{\alpha} \in \mathbb{R}^m$.

As for the **prior distribution** of the parameters $\boldsymbol{\alpha}, \sigma^2|M_\lambda$, a $\mathcal{NIG}_m\Big(\boldsymbol{\mu}, \boldsymbol{V}(\lambda), a, b\Big)$ is adopted:

$$f_{\boldsymbol{\alpha},\sigma^2|M_\lambda} = \frac{b^a}{(2\pi)^{m/2}\,\Gamma(a)\,|\boldsymbol{V}(\lambda)|^{1/2}}\,[\sigma^2]^{-(a+m/2+1)} \times \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\alpha}-\boldsymbol{\mu})'\boldsymbol{V}(\lambda)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu})+2b\right]\right\} \tag{3.2}$$

with $\boldsymbol{\mu} \in \mathbb{R}^m$ and the scalars a and b both in $\mathbb{R}_{>0}$. The hyperparameter $\boldsymbol{V}(\lambda)$ is a symmetric positive definite matrix of dimension $m \times m$ and is of full rank (we will relax this assumption in subsection 3.4). Finally, for the penalisation parameter λ, an improper uniform prior f_{M_λ} will be considered.

In section 3.3 we will establish that with the previous assumptions, $f_{M_\lambda|\boldsymbol{Y}}$ takes the form

$$f_{M_\lambda|\boldsymbol{Y}} \propto \frac{\Gamma(a^*)\,|\boldsymbol{V}^*(\lambda)|^{\frac{1}{2}}}{[b^*]^{a^*}\,|\boldsymbol{V}(\lambda)|^{\frac{1}{2}}}\; f_{M_\lambda}$$

with

$$\begin{aligned}
\boldsymbol{V}^* &= (\boldsymbol{B}'\boldsymbol{B}+\boldsymbol{V}^{-1})^{-1}\\
\boldsymbol{\mu}^* &= \boldsymbol{V}^*(\boldsymbol{B}'\boldsymbol{y}+\boldsymbol{V}^{-1}\boldsymbol{\mu})\\
a^* &= a+\frac{n}{2}\\
b^* &= b+\frac{1}{2}\left[\boldsymbol{y}'\boldsymbol{y}+\boldsymbol{\mu}'\boldsymbol{V}^{-1}\boldsymbol{\mu}-(\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}\boldsymbol{\mu}^*\right]
\end{aligned}$$

Subsequently, in section 3.4, we will deal with the choice of sensible hyperparameters for the prior distribution.

Appendix A contains a brief summary on concepts about (semi) positive definite matrices which will be used throughout the rest of the chapter.

3.2 Noninformative Priors

In the previous section, we have indicated that we will place a Normal-Inverse Gamma prior distribution on the parameters $\boldsymbol{\alpha}, \sigma^2 | M_\lambda$. This choice is motivated by a practical reason: the resulting posterior distribution $f_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda}$ can be expressed in a closed form also as a Normal-Inverse Gamma. Priors having the property of leading to a posterior in the same family, are said to be "*conjugate*" for the corresponding likelihood function.

The Bayesian models considered in scientific investigation generally assume that the likelihood dominates over the prior. One reason for this is that these kind of investigations are undertaken under the expectation of increasing the knowledge by a substantial amount and this would not be the case if the prior were very informative. Even if the scientist holds strong priors beliefs, using a "neutral" prior would lead to a posterior distribution which represented what someone should learn about the parameters if *a priori* the researcher knew very little about these parameters.

But how could we be more formal about this idea of an uninformative prior i.e. that we know very little *a priori* relative to what what the data has to tell us about the parameters ? For the sake of simplicity, let us assume that we are interested in a single parameter θ. If $\boldsymbol{y}$ represents the vector of observed values, we shall say that the likelihood is *data translated* if it has the form

$$L\left(\boldsymbol{y}|\phi\left(\theta\right)\right) = g\Big[\phi(\theta) - h(\boldsymbol{y})\Big]$$

where $\phi(\theta)$ is some one-to-one transformation of θ and $h(\boldsymbol{y})$ is a function of the data only. In such cases, the shape of the likelihood as a function of $\phi(\theta)$ is completely determined a priori except for its location which depends on the data yet to be observed. In other words, we know very little a priori relative to what the data is going to tell us and hence we are willing to accept one value of $\phi(\theta)$ as another. This state of indifference can be represented by an improper uniform prior distribution on ϕ, i.e. $p(\phi) \propto c$ which we shall call *noninformative* or *uninformative* for $\phi(\theta)$ with respect to the data.

Using the the theorem of change of variable, the noninformative prior for θ is given by

$$p(\theta) = p(\phi) \left| \frac{d\phi}{d\theta} \right| \propto \left| \frac{d\phi}{d\theta} \right| \tag{3.3}$$

As an example, let us consider the Normal distribution. If $Y \sim \mathcal{N}(\theta, \sigma^2)$ and assuming σ^2 known, it is

$$L(\boldsymbol{y}|\phi) \propto \exp\left\{ -\frac{(\phi - \hat{\phi})^2}{2\sigma^2} \right\} \tag{3.4}$$

with $\phi(\theta) = \theta$ and $h(\boldsymbol{y}) = \hat{\phi} = \bar{\boldsymbol{y}}$. Figure 3.1 represents the likelihood function for the Normal distribution (assuming σ^2 known) as a function of $\phi(\theta)$ for different values of $\hat{\phi}$. The improper uniform prior on $\phi(\theta)$ represents our state of indifference about this parametrization of the likelihood function.

The corresponding log-likelihood function $\ell(\boldsymbol{y}|\phi) = \log L(\boldsymbol{y}|\phi)$ for the Normal distribution has the form

$$\ell(\boldsymbol{y}|\phi) = constant - \frac{(\phi - \hat{\phi})^2}{2\sigma^2} \tag{3.5}$$

In order to find the parametrization $\phi(\theta)$ in the general case leading to a data translated likelihood function for which an improper uniform prior $p(\phi)$ can be considered, we shall approximate $\ell(\boldsymbol{y}|\phi)$ by its Taylor's expansion up to the second term around $\hat{\phi}$, the maximum-likelihood estimator of ϕ. It is

$$\ell(\boldsymbol{y}|\phi) \quad = \quad \ell(\boldsymbol{y}|\hat{\phi}) + \left[\frac{\partial \ell(\boldsymbol{y}|\phi)}{\partial \phi} \right]_{\theta=\hat{\phi}} (\phi - \hat{\phi}) - \frac{1}{2} \left[-\frac{\partial^2 \ell(\boldsymbol{y}|\phi)}{\partial \phi^2} \right]_{\phi=\hat{\phi}} (\phi - \hat{\phi})^2 \tag{3.6}$$

Taking into account that $\frac{\partial \ell(\boldsymbol{y}|\phi)}{\partial \phi}$ vanishes at $\phi = \hat{\phi}$ and comparing equations (3.5) and (3.6), we notice that we can make $\ell(\boldsymbol{y}|\phi)$ approximately data translated, by managing to make $\left[-\frac{\partial^2 \ell(\boldsymbol{y}|\phi)}{\partial \phi^2} \right]_{\phi=\hat{\phi}}$ roughly constant. Actually, this expression is a random variable, but due to the law of the large numbers,

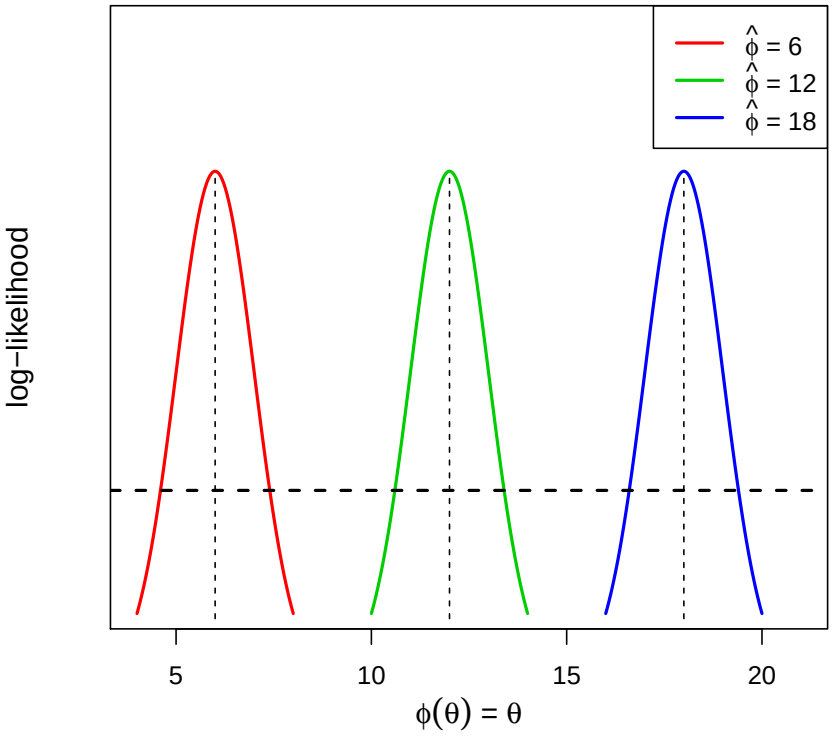


FIGURE 3.1: Likelihood for the Normal distribution as a function of θ for different values of $\hat{\theta}$ (σ^2 known). The dashed line represents the noninformative prior

$$\begin{aligned}
\left[-\frac{\partial^2 \ell(\boldsymbol{y}|\phi)}{\partial \phi^2}\right]_{\phi=\hat{\phi}} &= n\,\frac{\partial^2}{\partial \phi^2}\left[-\frac{1}{n}\sum_{i=1}^{n}\ell(y_i|\phi)\right]_{\phi=\hat{\phi}} \\
&\approx n\,\frac{\partial^2}{\partial \phi^2}\left[-\mathbb{E}\left(\ell(y|\phi)\right)\right] \\
&= n\,\mathbb{E}\left(-\frac{d^2}{d\phi^2}\ell(y|\phi)\right) \\
&\propto \mathbb{E}\left(-\frac{\partial^2}{\partial \theta^2}\ell(y|\phi)\right)\left(\frac{\partial \theta}{\partial \phi}\right)^2 \qquad (3.7)
\end{aligned}$$

Therefore in order to make $\ell(\boldsymbol{y}|\phi)$ approximately data translated, we must take

$$\left|\frac{d\phi}{d\theta}\right| \propto \mathcal{J}(\theta) = \left[\mathbb{E}\left(-\frac{\partial^2}{\partial \theta^2}\ell(y|\phi)\right)\right]^{\frac{1}{2}} \qquad (3.8)$$

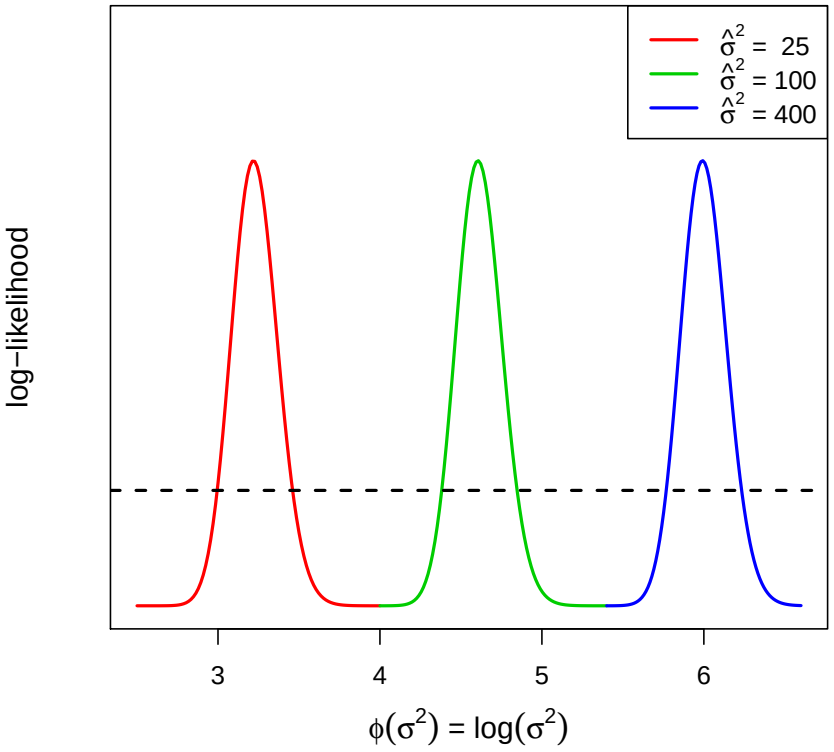


FIGURE 3.2: Likelihood for the Normal distribution as a function of $\log(\sigma^2)$ for different values of $\hat{\sigma}^2$ (θ known). The dashed line represents the improper uniform noninformative prior

$\mathcal{J}(\theta)$ is known as Jeffreys' prior and recalling equation (3.6) we see that an uninformative prior in the original parameter θ is given by

$$p(\theta) \quad \propto \quad \mathcal{J}(\theta) \tag{3.9}$$

Additionally, from equation (3.8) we obtain that the parametrization $\phi(\theta)$ that yields $\ell(\boldsymbol{y}|\phi)$ in the data translated form is

$$\phi(\theta) \quad = \quad \int^{\theta} \mathcal{J}(t)\, dt \tag{3.10}$$

Let us compute Jeffreys' prior $p(S)$ for a Normal distribution $\mathcal{N}(\mu,\, S)$, i.e. $S = \sigma^2$, assuming μ is known and the parametrization $\phi(\sigma^2)$. It is

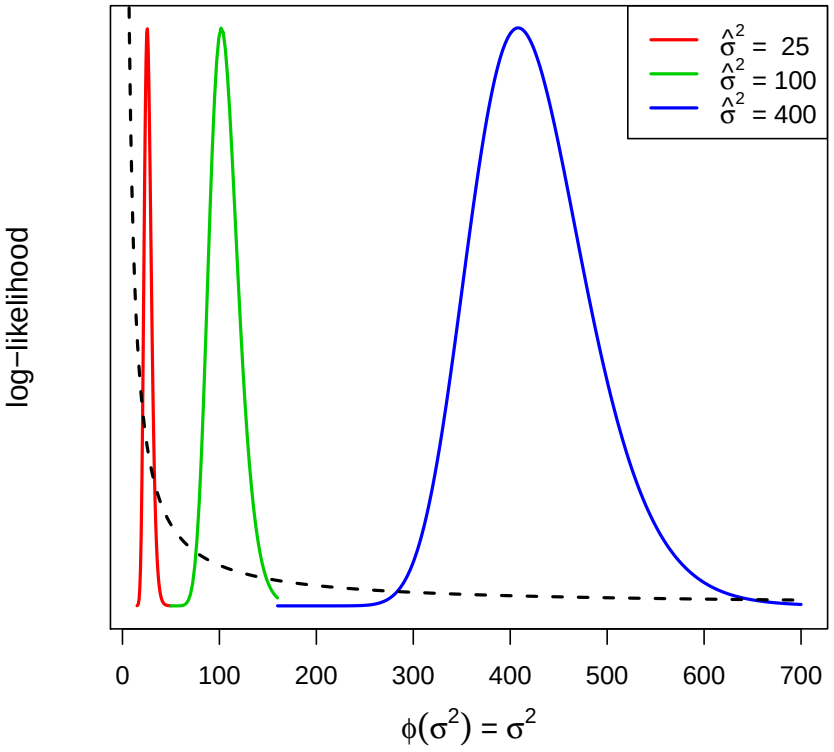


FIGURE 3.3: Likelihood for the Normal distribution as a function of σ^2 for different values of $\hat{\sigma}^2$ (θ known). The dashed line represents the noninformative prior

$$
\begin{aligned}
f(y|\mu,\, S) &\propto \frac{1}{S^{\frac{1}{2}}} \exp\left\{-\frac{(y-\mu)^2}{2S}\right\} \\
\ell(y|\mu,\, S) &= -\frac{1}{2}\log(S) - \frac{(y-\mu)^2}{2} S^{-1} \\
-\frac{\partial^2 \ell}{\partial S^2} &= -\frac{1}{2} S^{-2} + (y-\mu)^2 S^{-3} \\
\mathbb{E}\left(-\frac{\partial^2 \ell}{\partial S^2}\right) &= -\frac{1}{2} S^{-2} + S^{-3}\mathbb{E}\left[(y-\mu)^2\right] = -\frac{1}{2} S^{-2} + S^{-3} S \propto S^{-2} \\
\mathcal{J}(\sigma^2) &= \sigma^{-2}
\end{aligned}
\tag{3.11}
$$

$$\phi(\sigma^2) \quad = \quad \int^{\sigma^2} \frac{dt}{t} \quad = \quad \log(\sigma^2) \tag{3.12}$$

Figure 3.2 shows the likelihood in data translated form for different values of $\hat{\sigma}^2$ and Figure 3.3 represents the same likelihood functions in the original parameter σ^2. The corresponding noninformative priors are also included in these pictures.

The use of Jeffreys' prior for scale parameters has received some criticism in the literature (see e.g. Gelman, 2006; Gelman et al., 2013), with a (proper or improper) uniform distribution on σ^2 (rather than $\log(\sigma^2)$) often suggested as a better alternative, though this distribution is not uninformative in the sense of Jeffreys' prior discussed in this section (see e.g. Box and Tiao, 1992).

3.3 The derivation of the posterior density of the penalisation parameter λ

Bayes' Theorem tells us that $f_{M_\lambda|\boldsymbol{Y}} \propto f_{\boldsymbol{Y}|M_\lambda} f_{M_\lambda}$, where $f_{\boldsymbol{Y}|M_\lambda}$ corresponds to the distribution of the data for the particular model indexed by λ and f_{M_λ} reflects our prior beliefs regarding the distribution of the penalisation parameter. Because we have no reason to support any particular preference about λ, as mentioned earlier, f_{M_λ} will correspond to a vague improper uniform distribution.

We will obtain the density $f_{\boldsymbol{Y}|M_\lambda}$ as a by-product in the computation of the joint posterior density of the parameters $\boldsymbol{\alpha}$, $\sigma^2|\boldsymbol{Y}, M_\lambda$. Combining the densities (3.1) and (3.2) using Bayes' Rule, yields

$$\begin{aligned} f_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda} &= \frac{f_{\boldsymbol{\alpha},\sigma^2,\boldsymbol{Y},M_\lambda}}{f_{\boldsymbol{Y},M_\lambda}} \\ &= \frac{f_{\boldsymbol{Y}|\boldsymbol{\alpha},\sigma^2,M_\lambda} \quad f_{\boldsymbol{\alpha},\sigma^2|M_\lambda} \quad f_{M_\lambda}}{f_{\boldsymbol{Y}|M_\lambda} \quad f_{M_\lambda}} \end{aligned}$$

$$
\begin{aligned}
&= \frac{f_{\boldsymbol{Y}|\boldsymbol{\alpha},\sigma^2,M_\lambda} \quad f_{\boldsymbol{\alpha},\sigma^2|M_\lambda}}{f_{\boldsymbol{Y}|M_\lambda}} \\
&= \frac{b^a}{(2\pi)^{\frac{m+n}{2}}\,\Gamma(a)\,|\boldsymbol{V}|^{1/2}\,f_{\boldsymbol{Y}|M_\lambda}}\,\Lambda(\boldsymbol{\alpha},\sigma^2) \qquad (3.13)
\end{aligned}
$$

where

$$
\Lambda(\boldsymbol{\alpha},\sigma^2) = [\sigma^2]^{-(a+\frac{m+n}{2}+1)}\,\exp\left\{-\frac{1}{2\sigma^2}\Sigma(\boldsymbol{\alpha})\right\}
$$

$$
\begin{aligned}
\Sigma(\boldsymbol{\alpha}) &= (\boldsymbol{y}-\boldsymbol{B\alpha})'\,(\boldsymbol{y}-\boldsymbol{B\alpha}) + (\boldsymbol{\alpha}-\boldsymbol{\mu})'\,\boldsymbol{V}^{-1}\,(\boldsymbol{\alpha}-\boldsymbol{\mu}) + 2b \\
&= \boldsymbol{y'y} - \boldsymbol{y'B\alpha} - \boldsymbol{\alpha'B'y} + \boldsymbol{\alpha'B'B\alpha} + \boldsymbol{\alpha'V^{-1}\alpha} - \boldsymbol{\alpha'V^{-1}\mu} \\
&\quad - \boldsymbol{\mu'V^{-1}\alpha} + \boldsymbol{\mu'V^{-1}\mu} + 2b \\
&= \boldsymbol{\alpha'}(\boldsymbol{B'B}+\boldsymbol{V^{-1}})\boldsymbol{\alpha} - 2\boldsymbol{\alpha'}(\boldsymbol{B'y}+\boldsymbol{V^{-1}\mu}) + (2b + \boldsymbol{y'y} + \boldsymbol{\mu'V^{-1}}\mu) \qquad (3.14)
\end{aligned}
$$

The previous computation used the fact that $\boldsymbol{y'B\alpha}$ is a scalar and hence $\boldsymbol{y'B\alpha} = (\boldsymbol{y'B\alpha})' = \boldsymbol{\alpha'B'y}$. The same argument applies to $\boldsymbol{\alpha'V^{-1}\mu}$ and $\boldsymbol{\mu'V^{-1}\alpha}$, recalling in addition that $\boldsymbol{V}$ is symmetric (and therefore $\boldsymbol{V}^{-1}$ is symmetric, too).

The expansion of the exponent in the kernel of a $\mathcal{NIG}_m\,(\boldsymbol{\mu}^*,\,\boldsymbol{V}^*,\,a^*,\,b^*)$ distribution yields

$$
(\boldsymbol{\alpha}-\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu}^*)+2b^* = \boldsymbol{\alpha}'(\boldsymbol{V}^*)^{-1}\boldsymbol{\alpha}-2\boldsymbol{\alpha}'(\boldsymbol{V}^*)^{-1}\boldsymbol{\mu}^*+(\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}\boldsymbol{\mu}^*+2b^*
$$

By comparing the left hand side of the above expression with the expansion of $\Sigma(\boldsymbol{\alpha})$ in (3.14), it can be observed that $f_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda}$ takes on the form of a Normal-Inverse Gamma distribution with parameters $\boldsymbol{V}^*$ and $\boldsymbol{\mu}^*$ given by

$$\boldsymbol{V}^* = (\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{V}^{-1})^{-1} \qquad\qquad \boldsymbol{\mu}^* = \boldsymbol{V}^*(\boldsymbol{B}'\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{\mu})$$

Notice that $\boldsymbol{V}^*$ is also symmetric. The value of b^* can be established using $(\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}\boldsymbol{\mu}^* + 2b^* = 2b + \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\mu}'\boldsymbol{V}^{-1}\boldsymbol{\mu}$, which leads to

$$\begin{aligned} b^* &= \frac{1}{2}\Big[2b + \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\mu}'\boldsymbol{V}^{-1}\boldsymbol{\mu} - (\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}\boldsymbol{\mu}^*\Big] \\ &= \frac{1}{2}\Big[2b + \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\mu}'\boldsymbol{V}^{-1}\boldsymbol{\mu} - (\boldsymbol{y}'\boldsymbol{B} + \boldsymbol{\mu}'\boldsymbol{V}^{-1})(\boldsymbol{V}^*)'(\boldsymbol{V}^*)^{-1}\boldsymbol{V}^*(\boldsymbol{B}'\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{\mu})\Big] \\ &= \frac{1}{2}\Big[2b + \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\mu}'\boldsymbol{V}^{-1}\boldsymbol{\mu} - (\boldsymbol{y}'\boldsymbol{B} + \boldsymbol{\mu}'\boldsymbol{V}^{-1})\boldsymbol{V}^*(\boldsymbol{B}'\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{\mu})\Big] \end{aligned}$$

Finally, by setting

$$a^* = a + \frac{n}{2}$$

we have that

$$\begin{aligned} \Lambda(\boldsymbol{\alpha}, \sigma^2) &= [\sigma^2]^{-(a + \frac{m+n}{2} + 1)} \exp\left\{ -\frac{1}{2\sigma^2}\Sigma(\boldsymbol{\alpha}) \right\} \\ &= [\sigma^2]^{-(a^* + \frac{m}{2} + 1)} \exp\left\{ -\frac{1}{2\sigma^2}\Big[(\boldsymbol{\alpha} - \boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}(\boldsymbol{\alpha} - \boldsymbol{\mu}^*) + 2b^*\Big] \right\} \end{aligned}$$

which represents the kernel of a Normal-Inverse Gamma distribution $f^*_{\boldsymbol{\alpha}, \sigma^2|\boldsymbol{Y}, M_\lambda}$ with parameters $\boldsymbol{\mu}^*$, $\boldsymbol{V}^*$, a^*, b^*. Hence, it holds that

$$1 = \int\int f^*_{\boldsymbol{\alpha}, \sigma^2|\boldsymbol{Y}, M_\lambda}\, d\boldsymbol{\alpha}\, d\sigma^2 = \int\int \frac{(b^*)^{a^*}}{(2\pi)^{\frac{m}{2}}\, |\boldsymbol{V}^*|^{\frac{m}{2}}\, \Gamma(a^*)}\, \Lambda(\boldsymbol{\alpha},\, \sigma^2)\, d\boldsymbol{\alpha}\, d\sigma^2$$

where

$$\Lambda(\boldsymbol{\alpha},\, \sigma^2) = \frac{(2\pi)^{\frac{m}{2}}\, |\boldsymbol{V}^*|^{\frac{1}{2}}\, \Gamma(a^*)}{(b^*)^{a^*}}\; f^*_{\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{Y},M_\lambda}$$

Substituting this into the expression for the posterior density $f_{\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{Y},M_\lambda}$ in equation (3.13),

$$\begin{aligned} f_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda} &= \frac{b^a}{(2\pi)^{\frac{m+n}{2}}\, \Gamma(a)\, |\boldsymbol{V}|^{1/2}\, f_{\boldsymbol{Y}|M_\lambda}}\, \Lambda(\boldsymbol{\alpha}, \sigma^2) \\ &= \frac{b^a}{(2\pi)^{\frac{m+n}{2}}\, \Gamma(a)\, |\boldsymbol{V}|^{1/2}\, f_{\boldsymbol{Y}|M_\lambda}}\, \frac{(2\pi)^{\frac{m}{2}}\, |\boldsymbol{V}^*|^{\frac{1}{2}}\, \Gamma(a^*)}{(b^*)^{a^*}}\; f^*_{\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{Y},M_\lambda} \\ &= \frac{b^a\, \Gamma(a^*)\, |\boldsymbol{V}^*|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}\, (b^*)^{a^*}\, \Gamma(a)\, |\boldsymbol{V}|^{\frac{1}{2}}\, f_{\boldsymbol{Y}|M_\lambda}}\; f^*_{\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{Y},M_\lambda} \end{aligned}$$

Because $f_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda}$ is also a density, it must hold that

$$\int\int f_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda} d\boldsymbol{\alpha} d\sigma^2 \;=\; 1 \;=\; \int\int f^*_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda} d\boldsymbol{\alpha} d\sigma^2$$

and consequently, it necessarily follows that

$$\frac{b^a\, \Gamma(a^*)\, |\boldsymbol{V}^*|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}\, (b^*)^{a^*}\, \Gamma(a)\, |\boldsymbol{V}|^{\frac{1}{2}}\, f_{\boldsymbol{Y}|M_\lambda}} \;=\; 1$$

yielding $f^*_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda} = f_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda}$; hence $\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{Y}, M_\lambda \propto \mathcal{NIG}_m\,(\boldsymbol{\mu}^*,\, \boldsymbol{V}^*,\, a^*,\, b^*)$ with

$$\boldsymbol{V}^* = (\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{V}^{-1})^{-1} \tag{3.15}$$

$$\boldsymbol{\mu}^* = \boldsymbol{V}^*(\boldsymbol{B}'\boldsymbol{y} + \boldsymbol{V}^{-1}\boldsymbol{\mu}) \tag{3.16}$$

$$a^* = a + \frac{n}{2} \tag{3.17}$$

$$\begin{aligned} b^* &= \frac{1}{2}\Big[2b + \boldsymbol{y'y} + \boldsymbol{\mu' V^{-1} \mu} - (\boldsymbol{y'B} + \boldsymbol{\mu' V^{-1}})\boldsymbol{V}^*(\boldsymbol{B'y} + \boldsymbol{V^{-1}\mu})\Big] \\ &= b + \frac{1}{2}\Big[\boldsymbol{y'y} + \boldsymbol{\mu' V^{-1} \mu} - (\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}\boldsymbol{\mu}^*\Big] \end{aligned} \tag{3.18}$$

In addition, it follows that

$$f_{\boldsymbol{Y}|M_\lambda} = \frac{b^a\,\Gamma(a^*)\,|\boldsymbol{V}^*|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}\,(b^*)^{a^*}\,\Gamma(a)\,|\boldsymbol{V}|^{\frac{1}{2}}}$$

Recalling that only the hyperparameter $\boldsymbol{V}$ of the prior distribution $\boldsymbol{\alpha}, \sigma^2|M_\lambda$, is a function of λ, we have that

$$f_{M_\lambda|\boldsymbol{Y}} \quad \propto \quad f_{\boldsymbol{Y}|M_\lambda} f_{M_\lambda} \quad \propto \quad \frac{\Gamma(a^*)\left|\boldsymbol{V}^*(\lambda)\right|^{\frac{1}{2}}}{\Big[b^*(\lambda)\Big]^{a^*}\left|\boldsymbol{V}(\lambda)\right|^{\frac{1}{2}}}\; f_{M_\lambda}$$

by dropping all the constants not depending either on λ or on the posterior parameters. In practice, for computation efficiency, we are going to use a discrete version of $f_{M_\lambda|\boldsymbol{Y}}$. Recalling that we have assumed that f_{M_λ} has the form of an improper uniform, we can write

$$f_{M_\lambda|\boldsymbol{Y}} = \frac{G(\lambda)\left|\boldsymbol{V}(\lambda)\right|^{-\frac{1}{2}}}{\sum_\lambda G(\lambda)\left|\boldsymbol{V}(\lambda)\right|^{-\frac{1}{2}}} \qquad \text{where} \qquad G(\lambda) = \frac{\Gamma(a^*)\left|\boldsymbol{V}^*(\lambda)\right|^{\frac{1}{2}}}{\Big[b^*(\lambda)\Big]^{a^*}} \tag{3.19}$$

3.4 On the choice of the hyperparameters and its consequences

From the fact that the prior distribution for the parameters $\boldsymbol{\alpha}, \sigma^2 | M_\lambda$ was chosen to be $\mathcal{NIG}_m\Big(\boldsymbol{\mu},\, \boldsymbol{V}(\lambda),\, a,\, b\Big)$, it follows that $\boldsymbol{\alpha}|\sigma^2, M_\lambda \sim \mathcal{N}_m\Big(\boldsymbol{\mu},\, \sigma^2\boldsymbol{V}(\lambda)\Big)$ and $\sigma^2 \sim \mathcal{IG}(a, b)$ as the distribution of σ^2 does not depend on λ.

As mentioned at the beginning of section 3.2, the Normal-Inverse Gamma distribution is the conjugate of the multivariate Normal distribution and hence leads to a posterior distribution on the parameters in closed form which is also Normal-Inverse Gamma. We based the choice of the prior on these grounds. But we want also to select the corresponding hyperparameters in such a way that it will not dominate over the likelihood.

Recalling in addition that for the likelihood function it was assumed that $\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma^2, M_\lambda \sim \mathcal{N}_n\left(\boldsymbol{B\alpha}, \sigma^2\boldsymbol{I_n}\right)$, we have that for the joint distribution

$$
\begin{aligned}
f_{\boldsymbol{\alpha},\sigma^2|\boldsymbol{Y},M_\lambda} &= f_{\boldsymbol{Y}|\boldsymbol{\alpha},\sigma^2,M_\lambda} \times f_{\boldsymbol{\alpha}|\sigma^2,M_\lambda} \times f_{\sigma^2} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{B\alpha})'(\boldsymbol{y}-\boldsymbol{B\alpha}) \right\} \\
&\quad \times \frac{1}{(2\pi)^{m/2}\left|\sigma^2\boldsymbol{V}(\lambda)\right|^{1/2}} \exp\left\{ -\frac{1}{2}\Big(\boldsymbol{\alpha}-\boldsymbol{\mu}\Big)'\Big[\sigma^2\boldsymbol{V}(\lambda)\Big]^{-1}\Big(\boldsymbol{\alpha}-\boldsymbol{\mu}\Big) \right\} \\
&\quad \times f_{\sigma^2} \\
&\propto \exp\Big(-\frac{1}{2\sigma^2}\Big)\Big\{(\boldsymbol{y}-\boldsymbol{B\alpha})'(\boldsymbol{y}-\boldsymbol{B\alpha}) + (\boldsymbol{\alpha}-\boldsymbol{\mu})'\boldsymbol{V}(\lambda)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu})\Big\} \\
&\quad \times f_{\sigma^2} \qquad (3.20)
\end{aligned}
$$

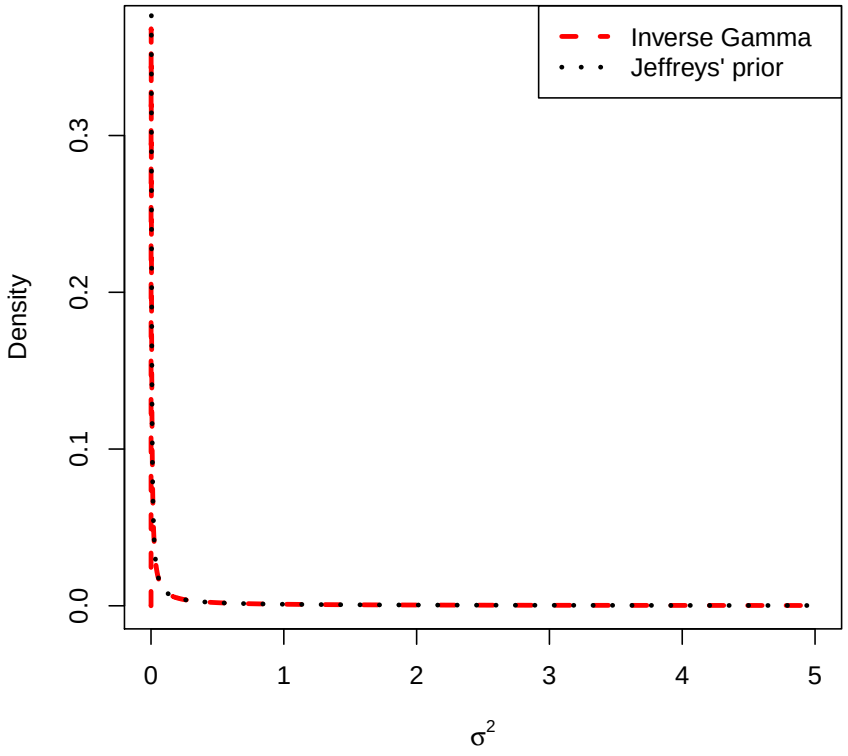


FIGURE 3.4: Inverse Gamma distribution for hyperparameters $a = b = 0.001$ and Jeffreys' prior for σ^2 corresponding to a Normal likelihood. Jeffreys' prior has been rescaled using the normalising constant of the Inverse Gamma distribution

The Inverse Gamma distribution has the form

$$f_{\sigma^2} \propto \left[\sigma^2\right]^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\}$$

Figure 3.4 shows that for very small values of the hyperparameters a and b, the Inverse Gamma distribution approaches the uninformative Jeffreys' prior for σ^2 in the case of the Normal likelihood, i.e. $p(\sigma^2) \propto \sigma^{-2}$ (in the figure Jeffreys' prior has been rescaled using the normalising constant of the Inverse Gamma distribution, namely $\frac{b^a}{\Gamma(a)}$). Based on this rationale, we shall choose the values $a = b = 0.001$ for these hyperparameters of the Inverse Gamma prior on σ^2.

$\boldsymbol{\mu} = \mathbf{0}$ is a sensible choice for the mean of the coefficients to account for our lack of preference on their sign. At this point, the expression between braces in equation

(3.20) simplifies to $(\boldsymbol{y} - \boldsymbol{B\alpha})'(\boldsymbol{y} - \boldsymbol{B\alpha}) + \boldsymbol{\alpha}'\boldsymbol{V}(\lambda)^{-1}\boldsymbol{\alpha}$. If we set set

$$\boldsymbol{V}(\lambda)^{-1} \quad = \quad \lambda \boldsymbol{D}'\boldsymbol{D} \tag{3.21}$$

where $\boldsymbol{D}$ is the second order difference matrix

$$\boldsymbol{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}$$

this expression corresponds precisely to the objective function 2.13 to be optimised when the non-parametric technique of P-splines is used for regression.

Recalling equation (2.14), the effect of such a choice is that the sequence of coordinates of the fitted parameter $\hat{\boldsymbol{\alpha}}$ does not change abruptly, yielding therefore a smooth regression function.

But unfortunately, it gives also rise to an issue of indetermination because $\boldsymbol{D}'\boldsymbol{D}$ is an $m \times m$ semi-positive definite symmetric matrix of rank $m - 2$ and hence it is not invertible. In order to by-pass this mishap we could consider instead

$$\boldsymbol{V}(\lambda,\, \tau)^{-1} = \lambda \boldsymbol{D}'\boldsymbol{D} + \tau \boldsymbol{I_m} \tag{3.22}$$

We will see that this matrix has full rank m. Consequently we can define $f_{M_{\lambda|\boldsymbol{Y}}}$ to be the limit of the resulting expression when $\tau \to 0$, provided that a limit exists. Recalling that we have set $\boldsymbol{\mu} = \boldsymbol{0}$ for the mean of the coefficients and equations (3.15), (3.18) and (3.22), we can rewrite the expression of $f_{M_{\lambda|\boldsymbol{Y}}}$ (3.19) as

$$f_{M_\lambda|\boldsymbol{Y}} = \lim_{\tau \to 0} \frac{G(\lambda,\tau)\left|\boldsymbol{V}(\lambda,\tau)\right|^{-\frac{1}{2}}}{\sum_\lambda G(\lambda,\tau)\left|\boldsymbol{V}(\lambda,\tau)\right|^{-\frac{1}{2}}} \quad \text{where} \quad G(\lambda,\tau) = \frac{\Gamma(a^*)\left|\boldsymbol{V}^*(\lambda,\tau)\right|^{\frac{1}{2}}}{\left[b^*(\lambda,\tau)\right]^{a^*}} \quad \text{and}$$

$$\begin{aligned}
\Big|\boldsymbol{V}^*(\lambda,\tau)\Big|^{\frac{1}{2}} &= \Big|\Big[\boldsymbol{B'B}+\boldsymbol{V}(\lambda,\tau)^{-1}\Big]^{-1}\Big|^{\frac{1}{2}} = |\boldsymbol{B'B}+\boldsymbol{V}(\lambda,\tau)^{-1}|^{-\frac{1}{2}} \\
&= |\boldsymbol{B'B}+\lambda\,\boldsymbol{D'D}+\tau\boldsymbol{I_m}|^{-\frac{1}{2}}
\end{aligned}$$

$$\Big|\boldsymbol{V}(\lambda,\tau)\Big|^{-\frac{1}{2}} = \Big|\boldsymbol{V}(\lambda,\tau)^{-1}\Big|^{\frac{1}{2}} = \Big|\lambda\boldsymbol{D'D}+\tau\boldsymbol{I_m}\Big|^{\frac{1}{2}} \tag{3.23}$$

$$\begin{aligned}
b^*(\lambda,\tau) &= b+\frac{1}{2}\Big[\boldsymbol{y'y}-\boldsymbol{y'B}\,\boldsymbol{V}^*(\lambda,\tau)\,\boldsymbol{B'y}\Big] \\
&= b+\frac{1}{2}\boldsymbol{y'}\Big[\boldsymbol{I_n}-\boldsymbol{B}(\boldsymbol{B'B}+\lambda\,\boldsymbol{D'D}+\tau\boldsymbol{I_m})^{-1}\boldsymbol{B'}\Big]\boldsymbol{y}
\end{aligned}$$

Notice that $G(\lambda,\,\tau)$ is continuous at $\tau=0$ and therefore $\lim_{\tau\to 0} G(\lambda,\,\tau) = G(\lambda,\,0)$. As mentioned earlier, $\boldsymbol{D'D}$ is an $m\times m$ semi-positive definite symmetric matrix of rank $m-2$. Hence there exists an orthogonal matrix $\boldsymbol{P}^{m\times m}$ such that

$$\boldsymbol{P}\,(\boldsymbol{D'D})\,\boldsymbol{P'} = \boldsymbol{\Delta} = \begin{pmatrix}
\delta_1 & 0 & \cdots & 0 & 0 & 0 \\
0 & \delta_2 & \cdots & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & \delta_{m-2} & 0 & 0 \\
0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & \cdots & 0 & 0 & 0
\end{pmatrix}$$

where $\delta_1,\,\ldots,\,\delta_{m-2}$ are the positive eigenvalues of $\boldsymbol{D'D}$. Thus

$$\begin{aligned}
\left|\boldsymbol{V}(\lambda,\,\tau)\right|^{-1} &= \left|\boldsymbol{V}(\lambda,\,\tau)^{-1}\right| = \left|\,\lambda\boldsymbol{D}'\boldsymbol{D} + \tau\boldsymbol{I_m}\,\right| = \lambda^m \left|\,\boldsymbol{D}'\boldsymbol{D} + \frac{\tau}{\lambda}\boldsymbol{I_m}\,\right| \\
&= \lambda^m \left|\,\boldsymbol{P}'\boldsymbol{\Delta}\boldsymbol{P} + \tfrac{\tau}{\lambda}\boldsymbol{I_m}\,\right| = \lambda^m \left|\,\boldsymbol{P}'(\boldsymbol{\Delta} + \frac{\tau}{\lambda}\boldsymbol{I_m})\boldsymbol{P}\,\right| \\
&= \lambda^m \left|\,\boldsymbol{P}\boldsymbol{P}'(\boldsymbol{\Delta} + \tfrac{\tau}{\lambda}\boldsymbol{I_m})\,\right| = \lambda^m \left|\,\boldsymbol{\Delta} + \tfrac{\tau}{\lambda}\boldsymbol{I_m}\,\right| \\
&= \lambda^{m-2}\,\tau^2 \prod_{i=1}^{m-2}\left(\delta_i + \frac{\tau}{\lambda}\right)
\end{aligned}$$

Hence, $\left|\boldsymbol{V}(\lambda,\,\tau)^{-1}\right| \neq 0$ implying that $\boldsymbol{V}(\lambda,\,\tau)^{-1}$ has full rank n when $\tau \neq 0$. The previous equation together with (3.23) yields

$$\left|\boldsymbol{V}(\lambda,\,\tau)\right|^{-\frac{1}{2}} = \lambda^{\frac{m-2}{2}}\,\tau\left[\prod_{i=1}^{m-2}\left(\delta_i + \frac{\tau}{\lambda}\right)\right]^{\frac{1}{2}}$$

and consequently

$$\begin{aligned}
f_{M_\lambda|\boldsymbol{Y}} &= \lim_{\tau\to 0} \frac{G(\lambda,\tau)\left|\boldsymbol{V}(\lambda,\tau)\right|^{-\frac{1}{2}}}{\sum_\lambda G(\lambda,\tau)\left|\boldsymbol{V}(\lambda,\tau)\right|^{-\frac{1}{2}}} \\
&= \lim_{\tau\to 0} \frac{G(\lambda,\tau)\,\lambda^{\frac{m-2}{2}}\,\tau\left[\prod_{i=1}^{m-2}(\delta_i + \frac{\tau}{\lambda})\right]^{\frac{1}{2}}}{\sum_\lambda\, G(\lambda,\tau)\,\lambda^{\frac{m-2}{2}}\,\tau\left[\prod_{i=1}^{m-2}(\delta_i + \frac{\tau}{\lambda})\right]^{\frac{1}{2}}} \\
&= \frac{G(\lambda,0)\,\lambda^{\frac{m-2}{2}}\left[\prod_{i=1}^{n-2}\delta_i\right]^{\frac{1}{2}}}{\sum_\lambda\, G(\lambda,0)\,\lambda^{\frac{m-2}{2}}\left[\prod_{i=1}^{m-2}\delta_i\right]^{\frac{1}{2}}} \\
&= \frac{\lambda^{\frac{m-2}{2}}\,\tilde{G}(\lambda)}{\sum_\lambda\,\lambda^{\frac{m-2}{2}}\,\tilde{G}(\lambda)}
\end{aligned}$$

where

$$\begin{aligned}\tilde{G}(\lambda) &= G(\lambda,0) = \frac{\Gamma(a^*)\left|\boldsymbol{V}^*(\lambda,0)\right|^{\frac{1}{2}}}{\left[b^*(\lambda,0)\right]^{a^*}} \\ &= \frac{\Gamma(a^*)\left|\boldsymbol{B}'\boldsymbol{B}+\lambda\,\boldsymbol{D}'\boldsymbol{D}\right|^{-\frac{1}{2}}}{\left\{b+\frac{1}{2}\boldsymbol{y}'\left[\boldsymbol{I_n}-\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B}+\lambda\,\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\right]\boldsymbol{y}\right\}^{a^*}} \qquad (3.24)\end{aligned}$$

In general,

$$f_{M_\lambda|\boldsymbol{Y}} \propto \lambda^{\frac{\text{rank}(D'D)}{2}} \times \frac{\Gamma\left(a^*\right)\left|V^*(\lambda)\right|^{\frac{1}{2}}}{\left[b^*(\lambda)\right]^{a^*}} \qquad (3.25)$$

It should be noticed that the choice of the number of basis functions is a crucial issue in the spatio-temporal setting. This is due to the number of matrix operations that need to be carried out to compute the determinant and the inverse of $\boldsymbol{H} = \boldsymbol{B}'\boldsymbol{B}+\lambda\,\boldsymbol{D}'\boldsymbol{D}$ in equation (3.24). These tasks have to be performed for every value of λ and therefore they may be very time consuming. Let us recall that the inversion of an $m \times m$ matrix involves around m^3 operations. In three dimensions, it is $m = p^3$ where p is roughly the one-dimensional number of basis functions. Therefore the computation of H^{-1} involves around $m^3 = p^9$ operations for each value of λ considered in the discrete approximation of the posterior distribution of the model M_λ given that $\boldsymbol{Y} = \boldsymbol{y}$. We will address these issues in section 3.8.

3.5 P-splines and Linear Mixed Models

There is a very close connection between P-splines and Linear Mixed Models which is worth mentioning. Let us recall that in the classical linear mixed models formulation

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \qquad \text{with} \quad \boldsymbol{\gamma} \sim \mathcal{N}_q\left(\boldsymbol{0}, \tau^2\boldsymbol{I}\right) \quad \text{and} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right) \tag{3.26}$$

where $\boldsymbol{Y} \in \mathbb{R}^n$ is the random vector of observations, $\boldsymbol{Z} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{U} \in \mathbb{R}^{n\times q}$ are design matrices, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of fixed effects, $\boldsymbol{\gamma} \in \mathbb{R}^q$ is the vector of random effects, τ^2 corresponds to the random effects variance and σ^2 represents the error variance.

It can be shown (see e.g. Fahrmeir et al., 2013) that the estimation of the parameters of the model is performed by minimising the objective function

$$S\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \sigma^2\right) = \left(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma}\right)'\left(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma}\right) + \frac{\sigma^2}{\tau^2}\boldsymbol{\gamma}'\boldsymbol{\gamma} \tag{3.27}$$

Generally the procedures for estimating these parameters rely on REML methods to obtain unbiased estimators for σ^2 and τ^2. Comparing equations (2.13) and (3.27), we notice that we can take advantage of such procedures to estimate the smoothing parameter λ. If we manage to rewrite the parameters defining the P-splines model in such a way that

$$\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} = \boldsymbol{B}\boldsymbol{\alpha} \tag{3.28}$$

$$\boldsymbol{\alpha}'\boldsymbol{D}'\boldsymbol{D}\boldsymbol{\alpha} = \boldsymbol{\gamma}'\boldsymbol{\gamma} \tag{3.29}$$

then we obtain the estimate of the penalisation parameter as $\hat{\lambda} = \dfrac{\hat{\sigma}^2}{\hat{\tau}^2}$. Let us recall that $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, $\boldsymbol{\alpha} \in \mathbb{R}^m$ and $\boldsymbol{D} \in \mathbb{R}^{r\times m}$. We start by proposing the decomposition

$$\boldsymbol{\alpha} = \tilde{\boldsymbol{Z}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\boldsymbol{\gamma} \tag{3.30}$$

with $\boldsymbol{\alpha} \in \mathbb{R}^m$, $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{m \times p}$ and $\tilde{\boldsymbol{U}} \in \mathbb{R}^{m \times q}$. The decomposition (3.30) has to meet the conditions

$$\begin{aligned} \boldsymbol{D}\tilde{\boldsymbol{Z}} &= \boldsymbol{0}_{r \times p} && (3.31) \\ \tilde{\boldsymbol{U}}' \boldsymbol{D}' \boldsymbol{D} \tilde{\boldsymbol{U}} &= \boldsymbol{I}_q && (3.32) \end{aligned}$$

Equation (3.31) imposes no penalisation on the vector of fixed effects $\boldsymbol{\beta}$ whereas equation (3.32) says that the components of the vector of random effects $\boldsymbol{\gamma}$ are independent and identically distributed. In addition, the matrix $[\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{U}}]$ must have full rank m to yield a one-to-one transformation.

The condition (3.29) is met as a consequence of the restrictions imposed by (3.31) and (3.32) because

$$\begin{aligned} \boldsymbol{\alpha}' \boldsymbol{D}' \boldsymbol{D} \boldsymbol{\alpha} &= \left(\tilde{\boldsymbol{Z}} \boldsymbol{\beta} + \tilde{\boldsymbol{U}} \boldsymbol{\gamma} \right)' \boldsymbol{D}' \boldsymbol{D} \left(\tilde{\boldsymbol{Z}} \boldsymbol{\beta} + \tilde{\boldsymbol{U}} \boldsymbol{\gamma} \right) \\ &= \left(\underbrace{\boldsymbol{D}\tilde{\boldsymbol{Z}}}_{\boldsymbol{0}_{r \times p}} \boldsymbol{\beta} + \boldsymbol{D}\tilde{\boldsymbol{U}} \boldsymbol{\gamma} \right)' \left(\underbrace{\boldsymbol{D}\tilde{\boldsymbol{Z}}}_{\boldsymbol{0}_{r \times p}} \boldsymbol{\beta} + \boldsymbol{D}\tilde{\boldsymbol{U}} \boldsymbol{\gamma} \right) \\ &= \boldsymbol{\gamma}' \underbrace{\tilde{\boldsymbol{U}}' \boldsymbol{D}' \boldsymbol{D} \tilde{\boldsymbol{U}}}_{\boldsymbol{I}_q} \boldsymbol{\gamma} \\ &= \boldsymbol{\gamma}' \boldsymbol{\gamma} \end{aligned}$$

If we manage to construct the matrices $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{U}}$ satisfying (3.31) and (3.32) such that $\operatorname{rank}\left([\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{U}}] \right) = m$, then the equivalence between the P-spline model (2.12) and the linear mixed model (3.26) is stated because

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{B} \left(\tilde{\boldsymbol{Z}} \boldsymbol{\beta} + \tilde{\boldsymbol{U}} \boldsymbol{\gamma} \right) + \boldsymbol{\varepsilon} \\ &= \left(\boldsymbol{B}\tilde{\boldsymbol{Z}} \right) \boldsymbol{\beta} + \left(\boldsymbol{B}\tilde{\boldsymbol{U}} \right) \boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\ &= \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \end{aligned}$$

Equation (3.31) suggests that the matrix $\hat{\boldsymbol{Z}}$ can be obtained by considering a basis of the null space for the linear application $f : \mathbb{R}^m \to \mathbb{R}^r$, $f(\boldsymbol{z}) = \boldsymbol{D}\boldsymbol{z}$ which has dimension $d = m - r$. It is easy to identify $\hat{\boldsymbol{Z}}$ if $\boldsymbol{D}$ is the difference matrix of order d. In this case, the columns of $\hat{\boldsymbol{Z}}$ are

$$\hat{\boldsymbol{z}}_j = \sum_{k=1}^{m} k^{j-1}\boldsymbol{e}_k \qquad j = 1, \ldots, d \tag{3.33}$$

where the vectors $\boldsymbol{e}_k$, $k = 1, \ldots, m$, correspond to the canonical basis of $\mathbb{R}^m$.

The $m \times m$ matrix $\boldsymbol{D}'\boldsymbol{D}$ is symmetric and semi-positive defined with rank r. By virtue of the Theorem of Spectral Decomposition (appendix A, item 11) we have that $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{P}'$ with $\boldsymbol{P} \in \mathbb{R}^{m\times m}$ orthogonal and $\boldsymbol{\Delta} \in \mathbb{R}^{m\times m}$ diagonal made up with the r non-null eigenvalues and the $m - r$ null eigenvalues of $\boldsymbol{D}'\boldsymbol{D}$. If we consider the matrix of orthogonal eigenvectors $\boldsymbol{P}_+ \in \mathbb{R}^{m\times r}$ corresponding to the non-null eigenvalues and the matrix $\boldsymbol{\Delta}_+ \in \mathbb{R}^{r\times r}$ constructed with these eigenvalues, it also holds that $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{P}_+\boldsymbol{\Delta}_+\boldsymbol{P}'_+$. If we set $\tilde{\boldsymbol{U}} = \boldsymbol{P}_+\boldsymbol{\Delta}_+^{-\frac{1}{2}}$

$$\begin{aligned}\tilde{\boldsymbol{U}}'(\boldsymbol{D}'\boldsymbol{D})\tilde{\boldsymbol{U}} &= \left(\boldsymbol{\Delta}_+^{-\frac{1}{2}}\boldsymbol{P}'_+\right)\left(\boldsymbol{P}_+\boldsymbol{\Delta}_+\boldsymbol{P}'_+\right)\left(\boldsymbol{P}_+\boldsymbol{\Delta}_+^{-\frac{1}{2}}\right)\\ &= \boldsymbol{I}_r\end{aligned}$$

and hence the condition (3.32) is satisfied achieving therefore the desired decomposition. Notice that as a corollary of the construction of $\tilde{\boldsymbol{U}}$ we obtain that $q = r$.

3.6 Model Averaging

Under the Bayesian approach $w_\lambda = f_{M_\lambda|\boldsymbol{Y}}$ represents the probability that the optimal value of the penalisation parameter takes on the value of λ given that we have observed the data $\boldsymbol{Y}$ and hence the obvious choice for λ is the one maximising w_λ. This value is known as the *Maximum a Posteriori* (MAP) estimator of λ.

Strictly adhering to this approach we would need to proceed by *model averaging*. If we denote by $\hat{y}_\lambda$ the posterior expectation of a particular fitted value for a given value of the penalisation parameter λ, the observations are estimated by means of $\tilde{\boldsymbol{y}} = \sum_\lambda w_\lambda \hat{\boldsymbol{y}}_\lambda$. In practice, not all the values of λ with posterior positive density are used in this averaging process. First, for computation efficiency, we select a subset from the overall possible values and then we consider those such that $w_\lambda \geq \frac{1}{K} \max\{w_\lambda\}$. A typical value for K suggested in the literature is $K = 20$ (see Raftery et al., 1997). This reduced case is known as Occam's model averaging.

Considering now that λ is a random variable, we have that for the variance of the predictive posterior distribution using model averaging it is

$$\begin{aligned} Var(\hat{y}_i) &= \mathbb{E}_\lambda\big(Var(\hat{y}_i|\lambda)\big) + Var_\lambda\big(\mathbb{E}(\hat{y}_i|\lambda)\big) \\ &= \sum_\lambda w_\lambda Var(\hat{y}_i|\lambda) + \sum_\lambda w_\lambda \big((\hat{y}_{\lambda,i}) - \tilde{y}_i\big)^2 \end{aligned}$$

where the first term represents the variance *within groups* or variance for a given value of λ and the second term is the variance *between groups* or variance due to different values of the penalisation parameter.

3.7 Ballooning

The issue of ballooning was outlined in subsection 2.7.2. An "intuitive" explanation for these unexpected high predicted concentration values can be easily visualised in the one-dimensional simulations shown in Figures 3.5, 3.6, 3.7, 3.8, 3.9 and 3.10. In this case we have a "gap" or "hole" in the data with a large gradient for the observations in the neighborhood. Traditional methods for the choice of the penalisation parameter tend to produce high predictions in the "uncertain" area.

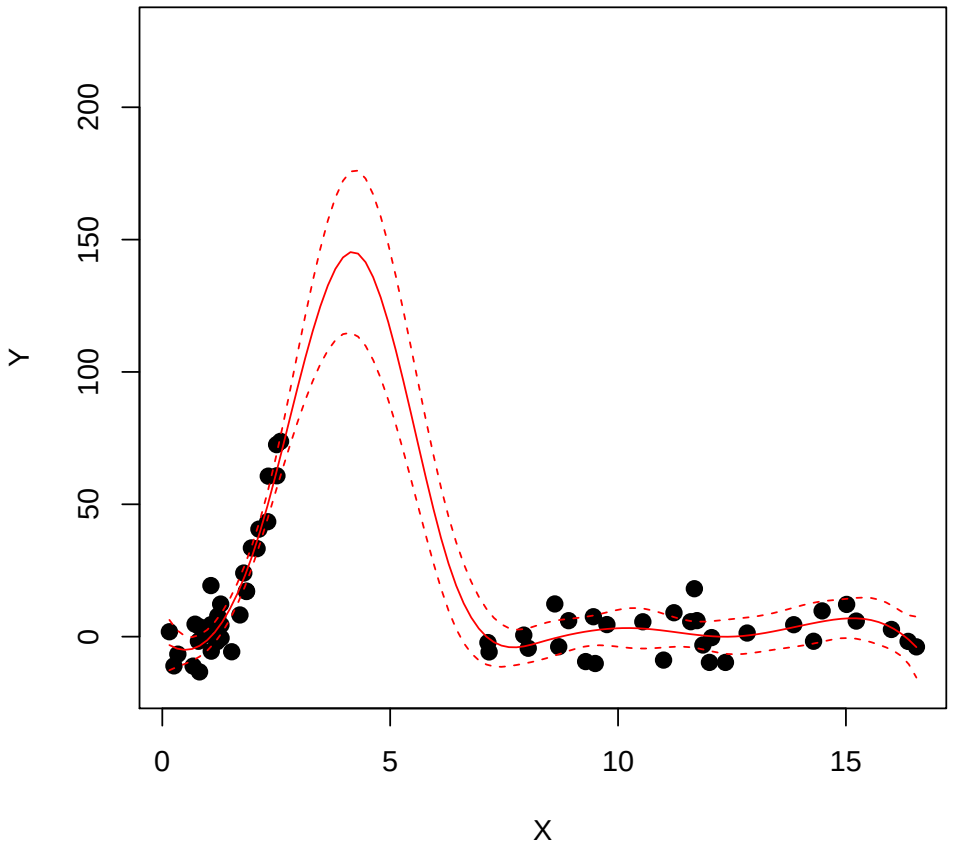


FIGURE 3.5: Predictions for one-dimensional simulation - Optimal MAP penalisation parameter determination with 95% confidence intervals

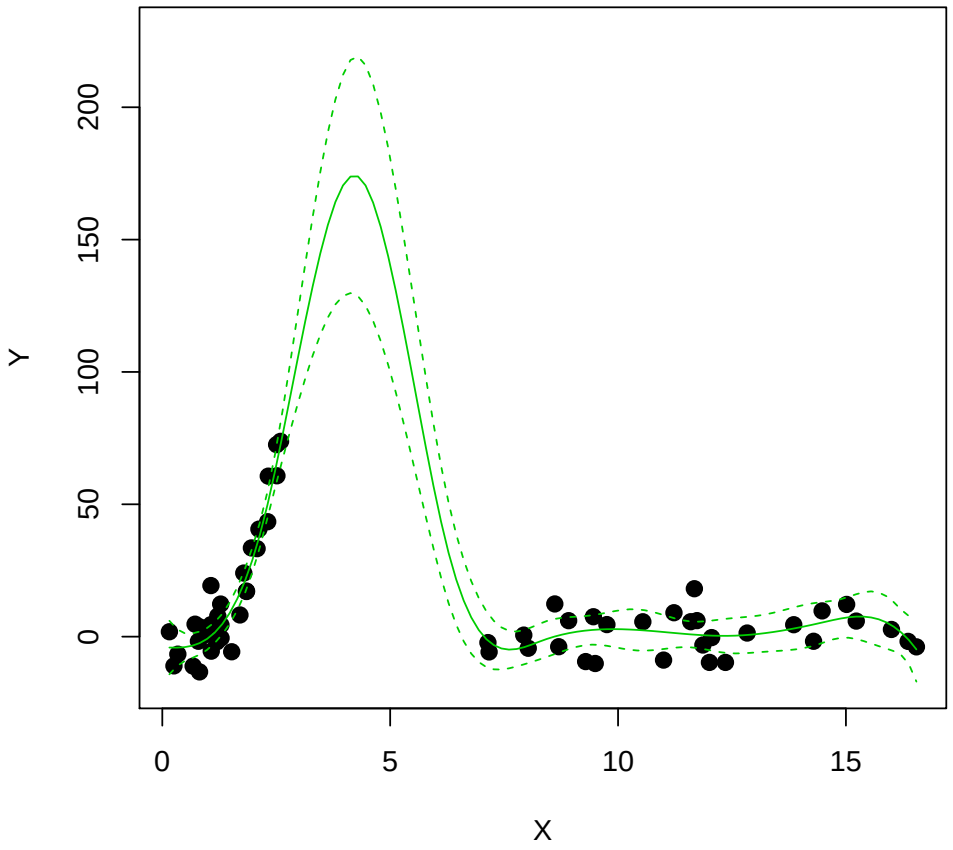


FIGURE 3.6: Predictions for one-dimensional simulation - Optimal CV penalisation parameter determination with 95% confidence intervals

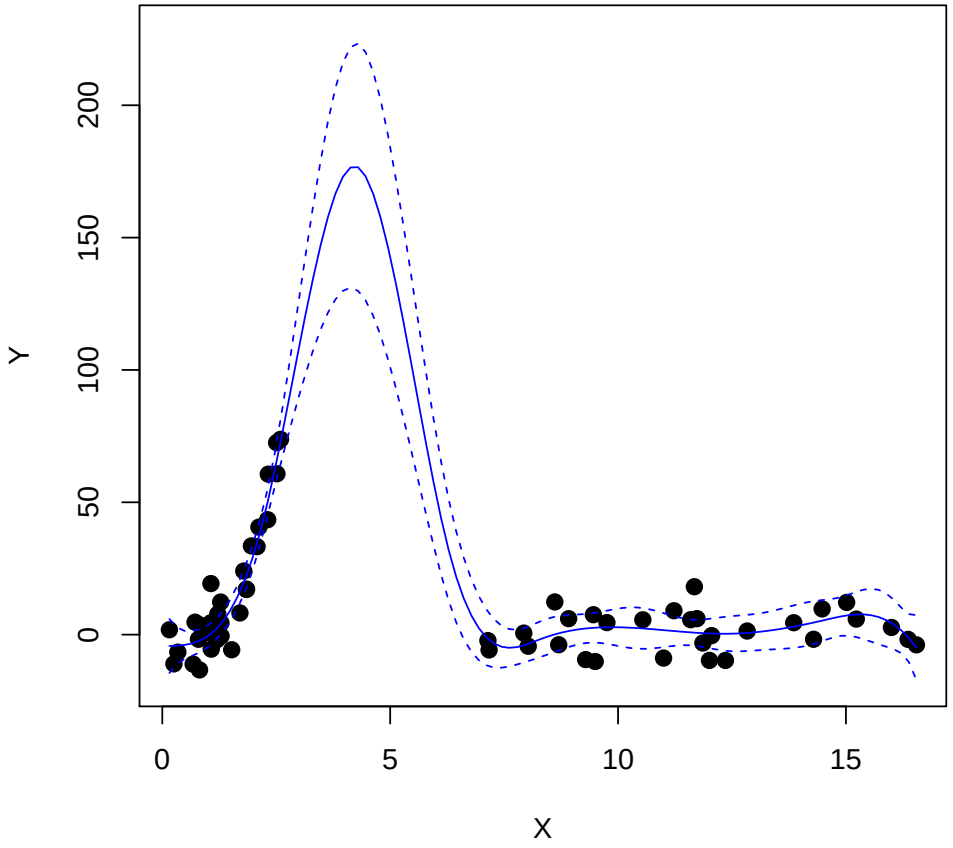


FIGURE 3.7: Predictions for one-dimensional simulation - Optimal GCV penalisation parameter determination with 95% confidence intervals

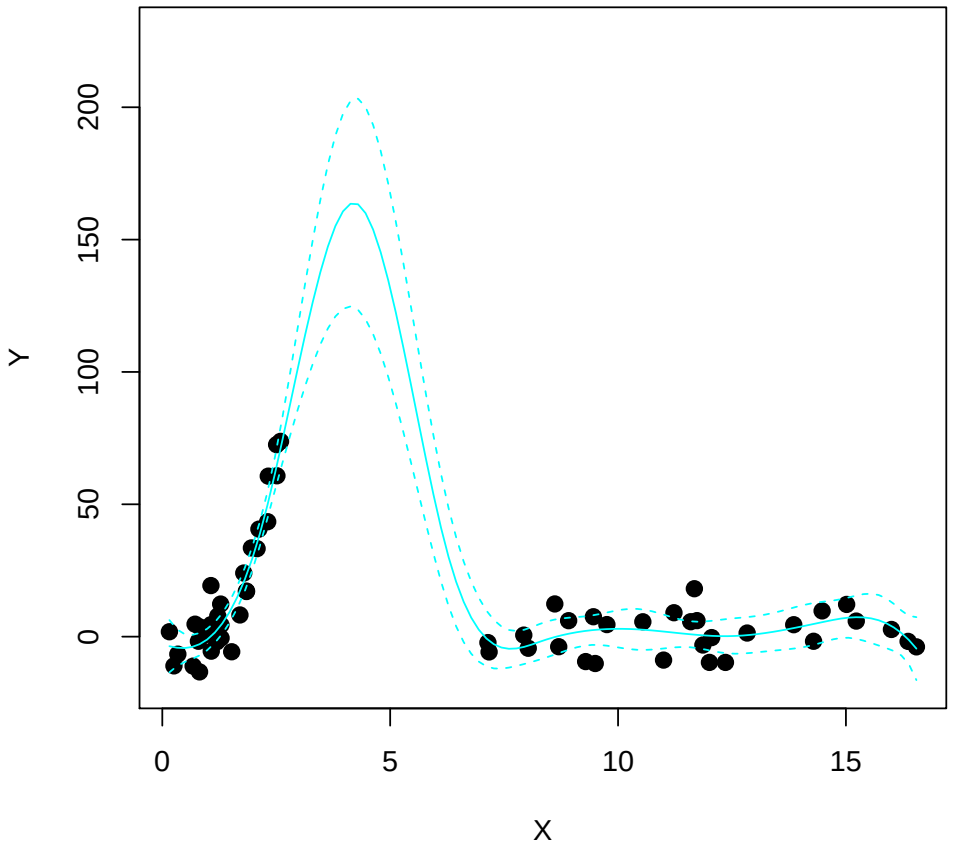


FIGURE 3.8: Predictions for one-dimensional simulation - Optimal BIC penalisation parameter determination with 95% confidence intervals

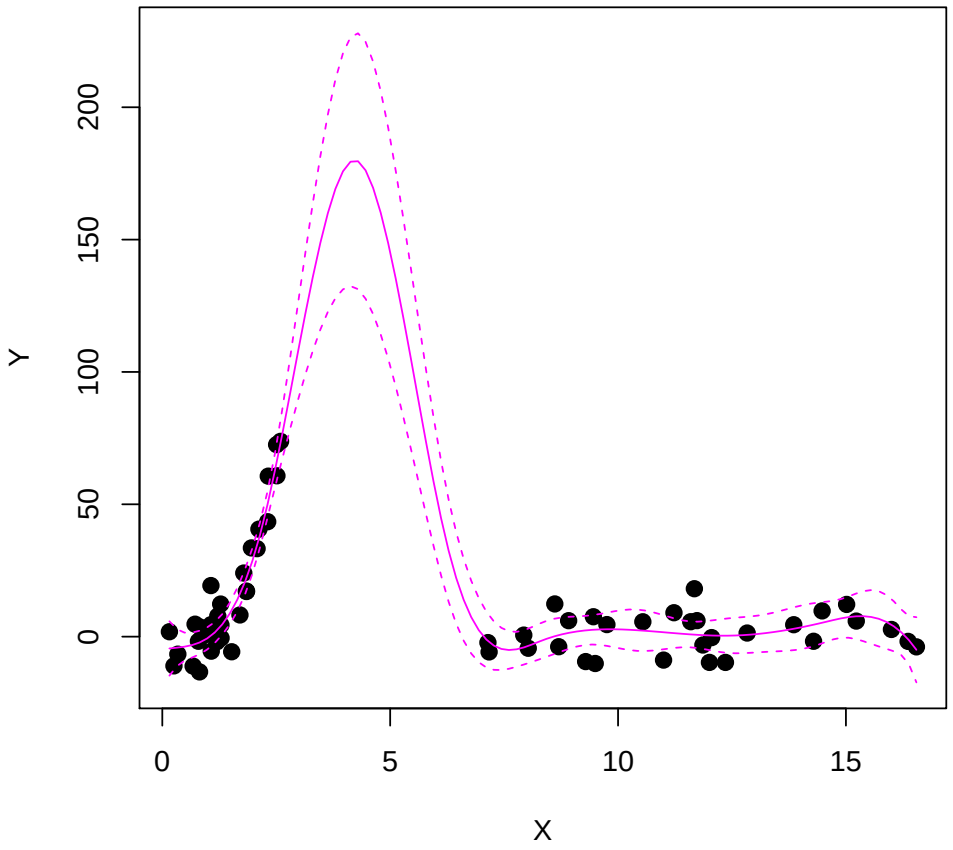


FIGURE 3.9: Predictions for one-dimensional simulation - Optimal AIC penalisation parameter determination with 95% confidence intervals

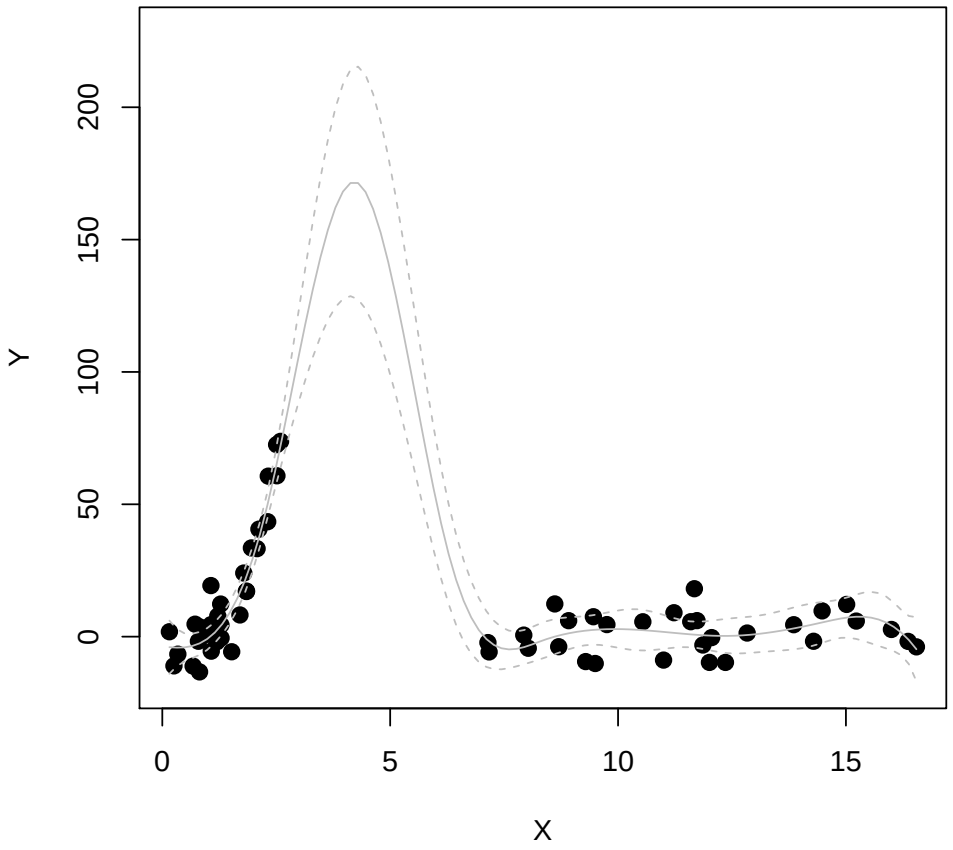


FIGURE 3.10: Predictions for one-dimensional simulation - Optimal AICC penalisation parameter determination with 95% confidence intervals

The MAP criterion yields a smoother fitting function. Figure 3.11 shows how in this example, the MAP penalises overfit more severely while the other methods are prone to severe undersmoothing.

Nevertheless, in some cases, even the MAP criterion fails to solve the problem of ballooning. This is due to the fact that in such cases, the assumption of a smooth change in the signal present in the data falls down.

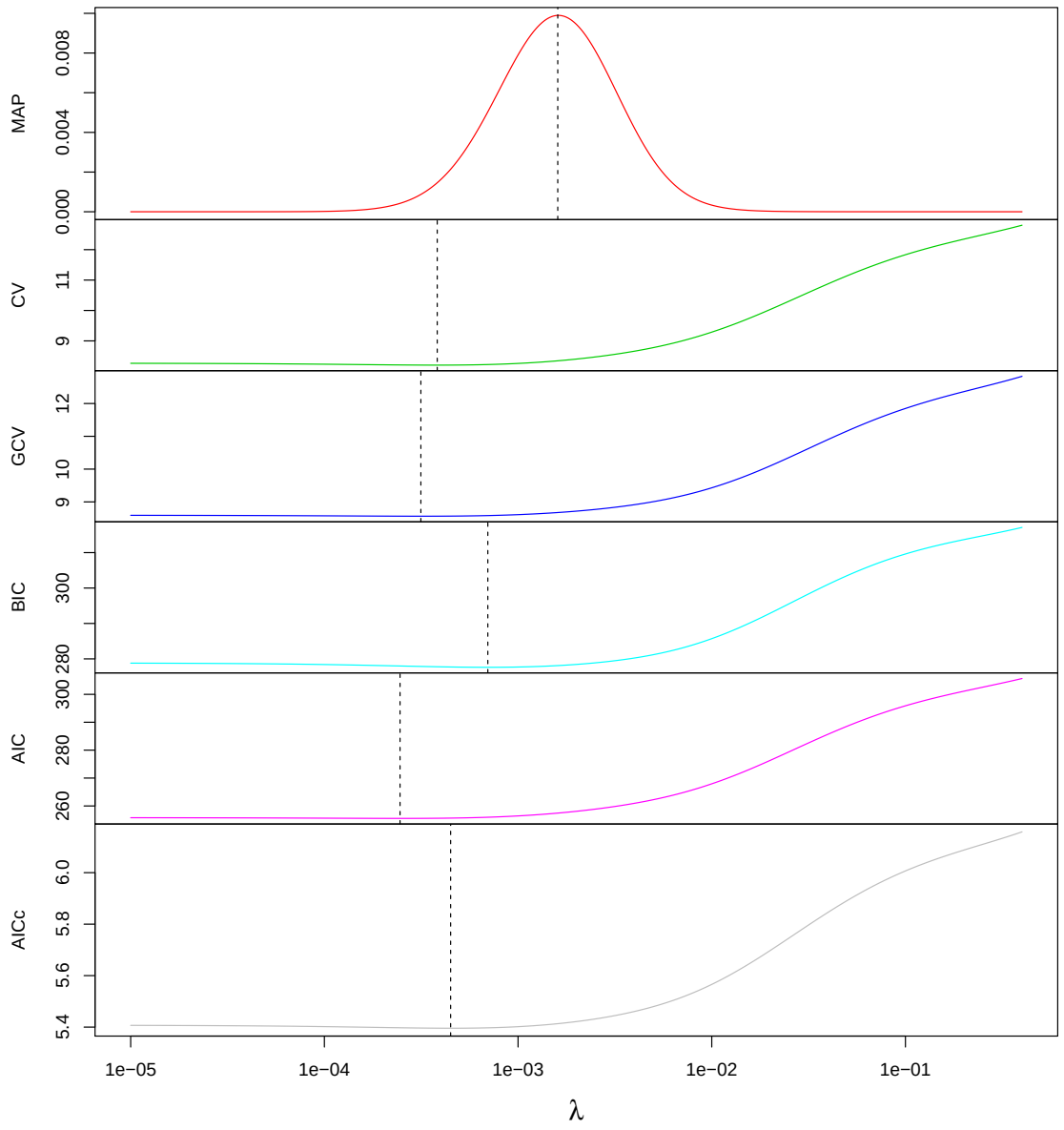


FIGURE 3.11: Predictions for one-dimensional simulation - Optimal penalisation parameter determination by optimising different criteria (λ in log scale). The vertical dashed lines indicate the optimal value of λ

To overcome this issue, we propose to modify the P-splines standard assumptions by

1. **Use penalty based on first rather than second order differences**: With the use of second order differences, the penalty shrinks towards low curvature, slowly changing the gradient. By using first order differences, the

penalty shrinks towards low gradient and allows for a more quickly changing curvature.

2. **Use quadratic rather than cubic P-splines**: We impose fewer smoothness constraints at the knots.

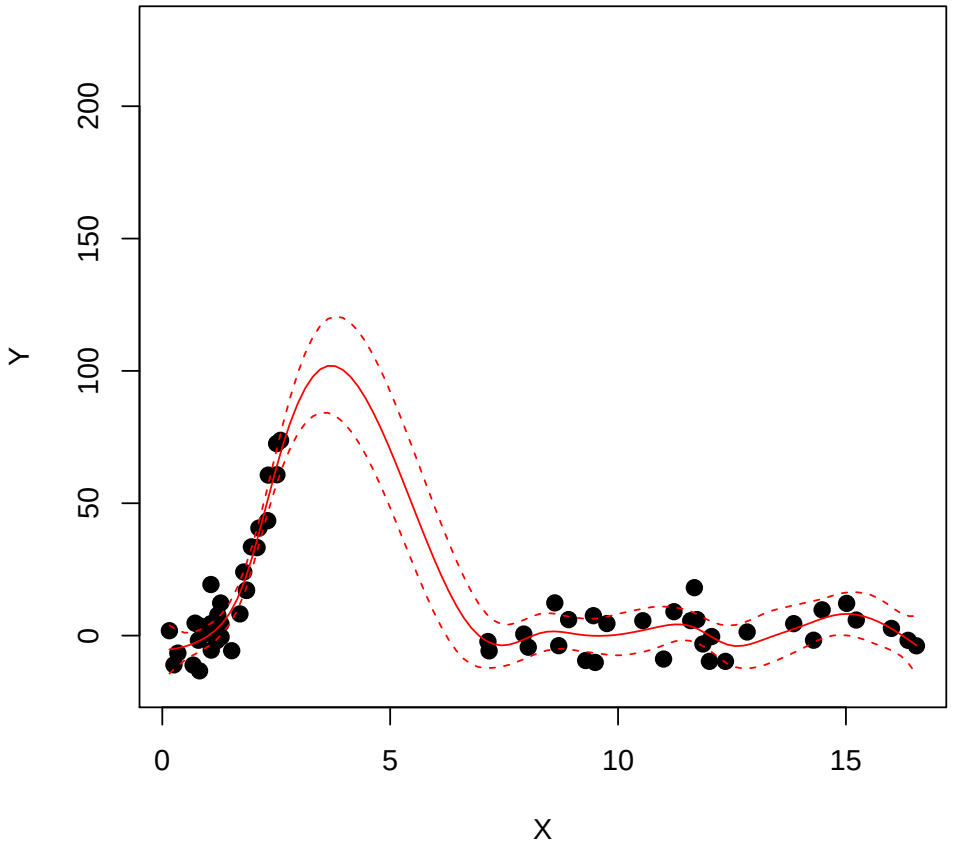


FIGURE 3.12: Predictions for one-dimensional simulation - Optimal MAP penalisation parameter determination using relaxed assumptions with 95% confidence intervals

3. **Increase the number of basis functions**: A low number of basis functions is an unrealistic assumption as it does not allow to carry over the effect of the penalisation. The two previous relaxing conditions have very little effect in modeling the lack of smoothness if we use a very low number of basis functions, as splines have a discontinuous higher order derivative at the knots. As we have mentioned earlier, this is a critical issue from the computational prospective and we will deal with it in the next subsection. In the spatio-temporal examples that follow in the subsequent chapters, we

will use 7 basis functions for easting and northing and 4 for time under the standard assumptions whereas these figures will be 14, 8 and 5 in the relaxed framework (see section 4.2).

Figure 3.12 shows the effect of relaxing the smoothness assumptions using the same illustrative example presented previously in Figure 3.5. It takes less to recover from the steep gradient in the data, because we favour flatter rather than smoother fitting curves.

It is worth noticing that the number of basis functions, the type of basis functions and especially the type of penalty used influences the way in which the gap of missing data is filled. In the case of the standard assumptions, the penalty based on second order differences yields lower bands which are higher than expected because recovering from a steep gradient is not immediate, as the penalty encourages constant derivative. On the other hand, the relaxed assumptions encourage the function itself to be constant so it can recover from a steep gradient more quickly. There is very little overlap in the confidence bands in Figures 3.5 and 3.12, showing that the choice of the penalty and the degree of the polynomials making up the spline functions can lead to quite different results in the peaks.

In the next chapters, we will see practical applications of the MAP technique using the standard and the relaxed assumptions.

3.8 Computational speed

We have seen that in order to evaluate $f_{M_\lambda|\boldsymbol{Y}}$, according to equations (3.24) and (3.25), we must compute the determinant and the inverse of the $m \times m$ matrix $\boldsymbol{B'B} + \lambda \boldsymbol{D'D}$ for every value of the penalisation parameter λ considered, with $m = p^3$ where p stands roughly for the number of one-dimensional basis functions being used.

As mentioned at the end of section 3.4 these tasks are very time consuming, in particular in our spatio-temporal setting. In this section we will explore an efficient method to perform these computations.

In principle, the sparseness of $\boldsymbol{B}$ and $\boldsymbol{D}$ could have been exploited to achieve efficiency in the use of memory (see Bowman et al., 2013, for details). But due to the fact that memory is not an issue in our context, we have gone for an implementation which does not use sparse matrix methods for $\boldsymbol{B}$ and $\boldsymbol{D}$.

For every value of λ, the naïve approach needs $\mathcal{O}(m^3)$ calculations. The method set out below reduces this to a one-off computation of complexity $\mathcal{O}(m^3)$ and a computation of complexity $\mathcal{O}(m)$ for each value of λ.

This technique jointly diagonalises the matrices $\boldsymbol{B}'\boldsymbol{B}$ and $\boldsymbol{D}'\boldsymbol{D}$ (see Golub and Van Loan, 1996) and is similar to the method first proposed by Eldén (1977) and also used by Wood (2000).

If l different values of λ are to be compared the naïve approach is of complexity $\mathcal{O}(l \times m^3)$. Our alternative approach is only of complexity $\mathcal{O}(m^3 + l \times m)$.

Without loss of generality we can consider that the symmetric matrix $\boldsymbol{\Omega}_0 \in \mathbb{R}^{m\times m}$ defined by $\boldsymbol{\Omega}_0 = \boldsymbol{B}'\boldsymbol{B} + \boldsymbol{D}'\boldsymbol{D}$ will be always strictly positive definite. Hence according to the Theorem of Spectral Decomposition (appendix A, item 11) it holds that

$$\boldsymbol{\Omega}_0 = \boldsymbol{B}'\boldsymbol{B} + \boldsymbol{D}'\boldsymbol{D} = \boldsymbol{P}_0\boldsymbol{\Delta}_0\boldsymbol{P}'_0 \tag{3.34}$$

with $\boldsymbol{P}_0$ orthogonal and $\boldsymbol{\Delta}_0$ diagonal and invertible. Hence the matrix $\boldsymbol{\Omega}_D \in \mathbb{R}^{m\times m}$, $\boldsymbol{\Omega}_D = \left(\boldsymbol{D}\boldsymbol{P}_0\boldsymbol{\Delta}_0^{-\frac{1}{2}}\right)'\left(\boldsymbol{D}\boldsymbol{P}_0\boldsymbol{\Delta}_0^{-\frac{1}{2}}\right)$ is well defined. Besides, it is symmetric and semi-positive definite (because $\text{rank}(\boldsymbol{D}) < m$). Hence, again by virtue of the Theorem of Spectral Decomposition, we have that

$$\boldsymbol{\Omega}_D = \left(\boldsymbol{D}\boldsymbol{P}_0\boldsymbol{\Delta}_0^{-\frac{1}{2}}\right)'\left(\boldsymbol{D}\boldsymbol{P}_0\boldsymbol{\Delta}_0^{-\frac{1}{2}}\right) = \boldsymbol{\Delta}_0^{-\frac{1}{2}}\boldsymbol{P'}_0\boldsymbol{D}'\boldsymbol{D}\boldsymbol{P}_0\boldsymbol{\Delta}_0^{-\frac{1}{2}} = \boldsymbol{P}_D\boldsymbol{\Delta}_D\boldsymbol{P'}_D \quad (3.35)$$

with $\boldsymbol{P}_D$ orthogonal and $\boldsymbol{\Delta}_D$ diagonal. Solving for $\boldsymbol{D}'\boldsymbol{D}$ from equation (3.35)

$$\boldsymbol{D}'\boldsymbol{D} = \Big(\underbrace{\boldsymbol{P}_0\boldsymbol{\Delta}_0^{\frac{1}{2}}\boldsymbol{P}_D}_{=\boldsymbol{U}}\Big)\,\boldsymbol{\Delta}_D\,\Big(\underbrace{\boldsymbol{P'}_D\boldsymbol{\Delta}_0^{\frac{1}{2}}\boldsymbol{P'}_0}_{=\boldsymbol{U}'}\Big) = \boldsymbol{U}\,\boldsymbol{\Delta}_D\,\boldsymbol{U}' \quad (3.36)$$

with $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ clearly invertible. Hence, from equation (3.34) we have that

$$\begin{aligned}\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{D}'\boldsymbol{D} &= \boldsymbol{U}\boldsymbol{U}^{-1}\left[\boldsymbol{P}_0\boldsymbol{\Delta}_0\boldsymbol{P'}_0\right]\left(\boldsymbol{U'}\right)^{-1}\boldsymbol{U}' \\ &= \boldsymbol{U}\left[\boldsymbol{P'}_D\boldsymbol{\Delta}_0^{-\frac{1}{2}}\boldsymbol{P'}_0\right]\left[\boldsymbol{P}_0\boldsymbol{\Delta}_0\boldsymbol{P'}_0\right]\left[\boldsymbol{P}_0\boldsymbol{\Delta}_0^{-\frac{1}{2}}\boldsymbol{P}_D\right]\boldsymbol{U}' \\ &= \boldsymbol{U}\boldsymbol{U}' \end{aligned} \quad (3.37)$$

yielding

$$\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{U}\boldsymbol{U}' - \boldsymbol{D}'\boldsymbol{D} = \boldsymbol{U}\boldsymbol{U}' - \boldsymbol{U}\boldsymbol{\Delta}_D\boldsymbol{U}' \quad (3.38)$$

Therefore, from equations (3.36) and (3.38)

$$\begin{aligned}\boldsymbol{B}'\boldsymbol{B} + \lambda\boldsymbol{D}'\boldsymbol{D} &= \left[\boldsymbol{U}\boldsymbol{U}' - \boldsymbol{U}\boldsymbol{\Delta}_D\boldsymbol{U}'\right] + \lambda\boldsymbol{U}\boldsymbol{\Delta}_D\boldsymbol{U}' \\ &= \boldsymbol{U}\left[\boldsymbol{I}_m - \boldsymbol{\Delta}_D + \lambda\boldsymbol{\Delta}_D\right]\boldsymbol{U}' \\ &= \boldsymbol{U}\Big[\underbrace{\boldsymbol{I}_m + (\lambda-1)\boldsymbol{\Delta}_D}_{=\boldsymbol{\Delta}_\lambda}\Big]\boldsymbol{U}' \\ &= \boldsymbol{U}\,\boldsymbol{\Delta}_\lambda\,\boldsymbol{U}' \end{aligned} \quad (3.39)$$

with $\boldsymbol{\Delta}_\lambda \in \mathbb{R}^{m\times m}$ diagonal. Finally equation (3.39) produces

$$\begin{aligned} \left|\boldsymbol{B}'\boldsymbol{B}+\lambda\boldsymbol{D}'\boldsymbol{D}\right| &= \left|\boldsymbol{U}\right|\left|\boldsymbol{\Delta}_\lambda\right|\left|\boldsymbol{U}'\right| \\ &= \left|\boldsymbol{U}\right|^2\left|\boldsymbol{\Delta}_\lambda\right| \end{aligned} \tag{3.40}$$

and

$$\begin{aligned} \left(\boldsymbol{B}'\boldsymbol{B}+\lambda\boldsymbol{D}'\boldsymbol{D}\right)^{-1} &= \left(\boldsymbol{U}'\right)^{-1}\boldsymbol{\Delta}_\lambda^{-1}\boldsymbol{U}^{-1} \\ &= \left(\boldsymbol{U}^{-1}\right)'\boldsymbol{\Delta}_\lambda^{-1}\boldsymbol{U}^{-1} \end{aligned} \tag{3.41}$$

Taking into account that for every orthogonal matrix $\boldsymbol{P}$ is $|\boldsymbol{P}|^2 = 1$ (see appendix A, item 13),

$$\begin{aligned} |\boldsymbol{U}|^2 &= \left|\boldsymbol{P}_0\boldsymbol{\Delta}_0^{\frac{1}{2}}\boldsymbol{P}_D\right|^2 \\ &= \left|\boldsymbol{P}_0\right|^2\left|\boldsymbol{\Delta}_0^{\frac{1}{2}}\right|^2\left|\boldsymbol{P}_D\right|^2 = \left|\boldsymbol{\Delta}_0\right| \end{aligned} \tag{3.42}$$

leading to

$$\left|\boldsymbol{B}'\boldsymbol{B}+\lambda\boldsymbol{D}'\boldsymbol{D}\right| = \left|\boldsymbol{\Delta}_0\right|\left|\boldsymbol{\Delta}_\lambda\right| \tag{3.43}$$

Besides, from equations (3.24) and (3.25) we have that

$$b(\lambda) = b+\frac{1}{2}\boldsymbol{y}'\left[\boldsymbol{I_n}-\boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B}+\lambda\,\boldsymbol{D}'\boldsymbol{D}\right)^{-1}\boldsymbol{B}'\right]\boldsymbol{y}$$

$$= \quad b + \frac{1}{2}\boldsymbol{y}'\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}'\boldsymbol{B}\Big(\boldsymbol{B}'\boldsymbol{B} + \lambda\,\boldsymbol{D}'\boldsymbol{D}\Big)^{-1}\boldsymbol{B}'\boldsymbol{y} \tag{3.44}$$

and hence recalling the definition of $\boldsymbol{U}$ from equations (3.36), and (3.41) it is

$$\begin{aligned}
\boldsymbol{y}'\boldsymbol{B}\Big(\boldsymbol{B}'\boldsymbol{B} + \lambda\,\boldsymbol{D}'\boldsymbol{D}\Big)^{-1}\boldsymbol{B}'\boldsymbol{y} &= \boldsymbol{y}'\boldsymbol{B}\left(\boldsymbol{U}^{-1}\right)'\boldsymbol{\Delta}_\lambda^{-1}\boldsymbol{U}^{-1}\,\boldsymbol{B}'\boldsymbol{y} \\
&= \boldsymbol{y}'\boldsymbol{B}\left(\boldsymbol{P}_0\boldsymbol{\Delta}_0^{-\frac{1}{2}}\boldsymbol{P}_D\right)\boldsymbol{\Delta}_\lambda^{-1}\left(\boldsymbol{P'}_D\boldsymbol{\Delta}_0^{-\frac{1}{2}}\boldsymbol{P'}_0\right)\boldsymbol{B}'\boldsymbol{y} \\
&= \Big(\underbrace{\boldsymbol{y}'\boldsymbol{B}\boldsymbol{P}_0\boldsymbol{\Delta}_0^{-\frac{1}{2}}\boldsymbol{P}_D}_{=\boldsymbol{w}'}\Big)\boldsymbol{\Delta}_\lambda^{-1}\Big(\underbrace{\boldsymbol{P'}_D\boldsymbol{\Delta}_0^{-\frac{1}{2}}\boldsymbol{P'}_0\boldsymbol{B}'\boldsymbol{y}}_{=\boldsymbol{w}}\Big) \\
&= \boldsymbol{w}'\,\boldsymbol{\Delta}_\lambda^{-1}\,\boldsymbol{w}
\end{aligned} \tag{3.45}$$

with $\boldsymbol{w} \in \mathbb{R}^m$. Combining equations (3.43), (3.45) and (3.43) with equations (3.24) and (3.25) we finally obtain

$$f_{M_\lambda|\boldsymbol{Y}} \quad \propto \quad \lambda^{\frac{\operatorname{rank}(D'D)}{2}} \times \frac{\left[\Gamma\big(a^*\big)\left|\boldsymbol{\Delta}_0\right|^{-\frac{1}{2}}\right]\left|\boldsymbol{\Delta}_\lambda\right|^{-\frac{1}{2}}}{\left\{\left[b + \frac{1}{2}\|\boldsymbol{y}\|^2\right] - \frac{1}{2}\boldsymbol{w}'\boldsymbol{\Delta}_\lambda^{-1}\boldsymbol{w}\right\}^{a^*}} \tag{3.46}$$

Note that we do not need to compute $\hat{\boldsymbol{\alpha}}$ which would be of complexity $\mathcal{O}\big(m^2\big)$, so the overall complexity for each value of λ is only $\mathcal{O}\big(m\big)$, which is very fast.

Equation (3.46) depends on λ only through the determinant and the inverse of $\boldsymbol{\Delta}_\lambda$ which is diagonal (recall equation (3.39)) and hence they are very fast to compute. The expressions within square brackets as well as $\boldsymbol{w}$ do not depend on λ and therefore they need to be evaluated only once.

As mentioned earlier, we choose the optimal value of the penalisation parameter by picking-up the one maximising the posterior distribution $f_{M_\lambda|\boldsymbol{Y}}$ among a certain number of plausible candidates.

For the sake of a practical comparison of the execution times involved in the computation of the optimal value of λ using the aforementioned discrete procedure, we have considered the set-up described in more detail in section 5.2 based on actual data provided by Shell.

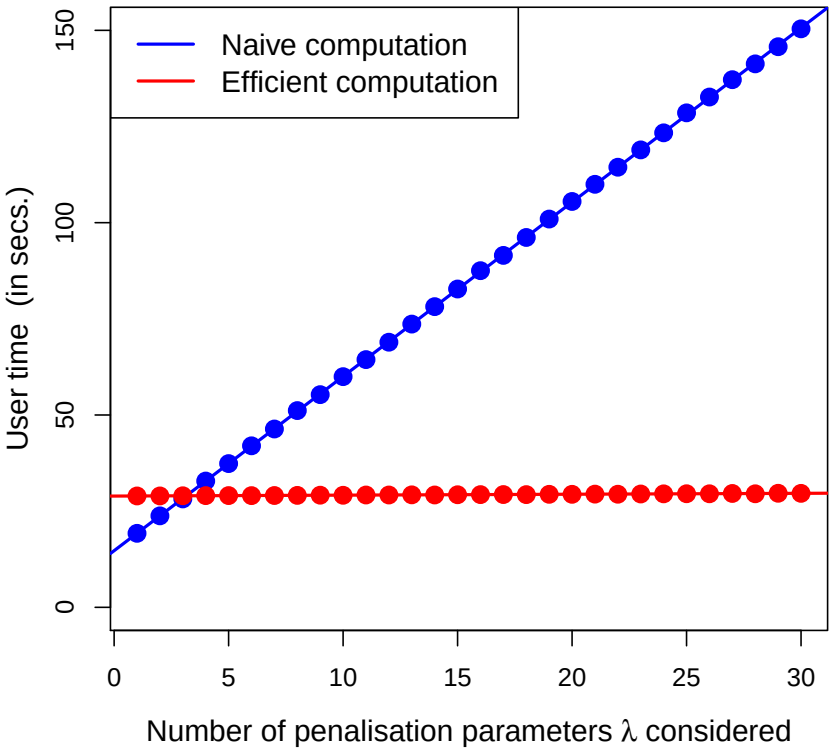


FIGURE 3.13: Comparison of the execution times of the posterior density for different numbers of the penalisation parameter λ

This comparison of the execution times between the naïve computation with equation (3.25) and the more efficient method using equation (3.46), was carried out under the framework of the relaxed assumptions.

Figure 3.13 pictures the total user time (in seconds) used to compute the posterior distributions of λ for different numbers of the penalisation parameter. It can be noticed that whereas this time increases very quickly with the number of values considered for λ using the naïve approach, it is practically constant if the optimised method is employed. Recall that as mentioned at the beginning of this section, if l different values of λ are to be compared, the naïve approach is of complexity $\mathcal{O}\big(l \times m^3\big)$ whereas the optimised approach is only of complexity $\mathcal{O}\big(m^3 + l \times m\big)$.

The almost constant figure corresponding to the second method, is due to the one-off computation of the decompositions of $\boldsymbol{\Omega}_0$ and $\boldsymbol{\Omega}_D$ (equations (3.34) and (3.35)) and the evaluation of $\boldsymbol{w}$ as indicated in equation (3.45). Both complexities increase linearly with l (m is fixed) but the slope under the naïve approach (m^3) is very high compared to the slope using the optimised method (m), in particular if a fairly large number of basis functions is used.

Similarly Figure 3.14 shows the comparison in the total execution times involved using both methods, considering a fixed number of candidates (30) for the penalisation parameter λ but different dimensions for the vector of parameters $\hat{\alpha}$. The increasing number of parameters corresponds to a higher number of basis functions on each of the spatio-temporal dimensions.

Now the complexity of both methods increases in a cubic fashion with m (l is fixed); however, under the naïve approach the coefficient of the higher order term in the expression of the complexity is l whereas in the optimised method such coefficient is 1. This is the reason for the steep increase of the total execution time using equation (3.25) to compute the posterior densities for all the values of the penalisation parameter λ considered.

The values used to construct Figures 3.13 and 3.14 can be found in the appendix B (Tables B.1 and B.2 respectively).

A comparison between execution times was performed using the function *gam* of the *R* package *mgcv* (see Wood, 2006). For moderately sized problems the approach proposed here is ten times faster. We expect this difference to be even bigger for larger data sets.

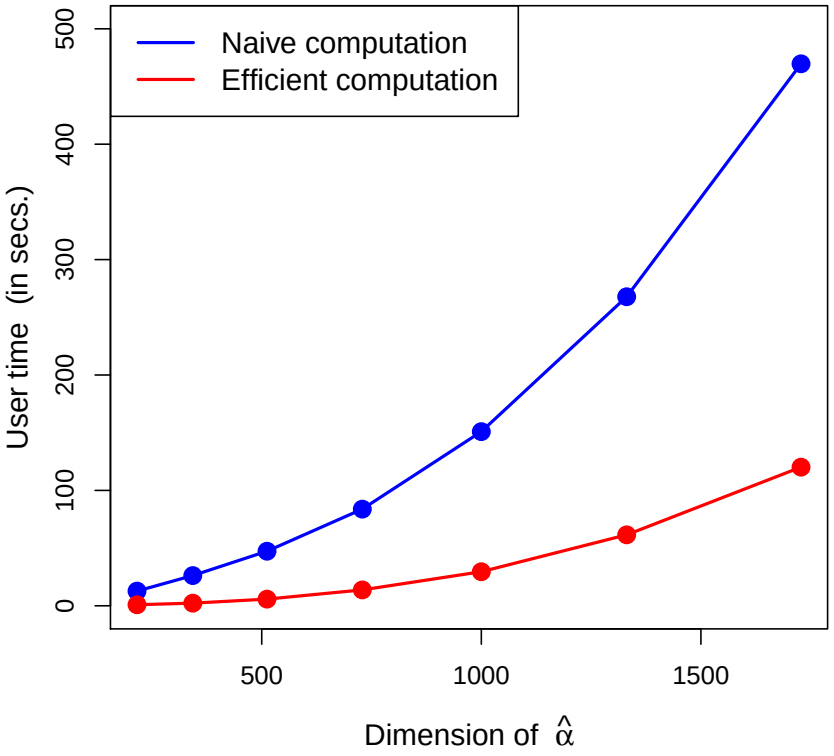


FIGURE 3.14: Comparison of the execution times of the posterior density for 30 values of the penalisation parameter λ for different dimensions of $\hat{\alpha}$

3.9 On the choice of a second smoothing parameter

Ideally, we should use two smoothing parameters: one for space and another one for time. This could be done in a more flexible context such as Morrisey et al. (2011) which resorts to posterior sampling using MCMC. However, given the time constraints imposed in this work, such an approach is infeasible.

In principle, a second smoothing parameter can be added to the model but the efficient linear algebra described in section 3.8 can be used to tune only one of such smoothing parameters. We have informally tackled the trade-off of smoothness between space and time by using a different number of basis functions. Because

smoothness between space and time seems to be always on the same relative scale, we can use information from past experience to guide the choice of the corresponding ratio of the number of basis functions. Given the choice of the number of basis functions, we expect the relative ratio of smoothing parameters between space and time to be estimated close to 1.

We will set out below how a second smoothing parameter can be handled more formally. Suppose that we want to minimise

$$(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}) + \lambda_1 \boldsymbol{\alpha}' \boldsymbol{D}_1' \boldsymbol{D}_1 \boldsymbol{\alpha} + \lambda_2 \boldsymbol{\alpha}' \boldsymbol{D}_2' \boldsymbol{D}_2 \boldsymbol{\alpha}, \tag{3.47}$$

where first penalty, say, corresponds to smoothness in space and the second penalty corresponds to smoothness in time. As explained above, the efficient linear algebra cannot be applied directly to this problem. We will briefly describe two methods that still harness the power of the efficient linear algebra, but require one parameter to be tuned manually using a grid search.

One approach is incorporate one penalty into the data, i.e. consider augmented data

$$\tilde{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \qquad\qquad \tilde{\boldsymbol{B}} = \begin{bmatrix} \boldsymbol{B} \\ \sqrt{\lambda_1}\boldsymbol{D}_1 \end{bmatrix}$$

We can then rewrite (3.47) as

$$(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{B}}\boldsymbol{\alpha})'(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{B}}\boldsymbol{\alpha}) + \lambda_2 \boldsymbol{\alpha}' \boldsymbol{D}_2' \boldsymbol{D}_2 \boldsymbol{\alpha},$$

In this formulation λ_1 would need to be adjusted "manually" (e.g. using a grid search), but given λ_1, λ_2 can be tuned efficiently.

Of course, one can also choose to incorporate $\boldsymbol{D}_2$, rather than $\boldsymbol{D}_1$, into the augmented data, or even switch between the two as part of the tuning algorithm, thus avoiding the need for grid search. Similar approaches have been taken in the literature (see e.g. Zou and Hastie, 2005).

Al alternative approach better suited to the smoothing nature of the problem and used e.g. by R package *mgcv*, is to set $\kappa = \lambda_2/\lambda_1$ and rewrite (3.47) as

$$(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}) + \lambda_1 \boldsymbol{\alpha}' \left(\boldsymbol{D}_1'\boldsymbol{D}_1 + \kappa \boldsymbol{D}_2'\boldsymbol{D}_2\right) \boldsymbol{\alpha}.$$

The parameter κ would need to be adjusted manually, but, given κ, the parameter λ_1, which controls the overall smoothness, can be tuned using the efficient linear algebra.

Chapter 4

Simulation Study

4.1 The proposed "true" model

The objective of this chapter is to carry out a simulation study to compare the different methods of selecting the smoothing parameter in a systematic way. As mentioned in chapter 1, we will use the data from Bowman et al. (2013).

The data were simulated from a highly idealised model for the spread of a solute in water, based on the partial differential equation

$$\frac{\partial y}{\partial t} = D \cdot \left(\frac{\partial^2 y}{\partial x_1^2} + \frac{\partial^2 y}{\partial x_2^2} \right) + \psi_1(x_1, x_2) \frac{\partial y}{\partial x_1} + \psi_2(x_1, x_2) \frac{\partial y}{\partial x_2}.$$

Here y denotes the concentration of the solute, x_1 and x_2 denote the spatial co-ordinates and $t \in [0, 1]$ denotes time. The first term describes the spread of the solute in the groundwater by diffusion, with the constant D controlling how fast the solute spreads. The two further advection terms describe how the solute is affected by groundwater flow, whose direction and velocity is represented by the functions ψ_1 and ψ_2. These functions were chosen to correspond to the observed groundwater levels in the benzene example discussed in section 5.2.

Figure 4.1(a) shows the assumed groundwater levels and flow which, in the simulations, for simplicity, are considered to be constant over time.

The assumed initial spread of the solute is given in Figure 4.1(b). Figures 4.1(c), 4.1(d) and 4.2(a) show the spread at time $t \in \{0.25, 0.5, 0.7\}$ (in years). The "true" concentrations were obtained by interpolating the numerical solution to the differential equation, computed over a $N \times N \times N$ regular grid with $N = 100$.

Observed measurement data were generated by multiplicative Gaussian error terms because the uncertainty in the measured concentrations can reasonably be expected to be proportional to the magnitude of the value (e.g. the uncertainty around a measured value of $10\mu g/\ell$ would be expected to be very much less than the uncertainty surrounding a measured value of $10000\mu g/\ell$), with standard deviation chosen to give a signal-to-noise ratio on the log-scale of $10:1$. This reflects the fact that measurements of the solutes are usually quite accurate. A very small value of 0.05 was used for within-well correlation of the data, while the between-well correlation was assumed to be 0. Before the data were analysed they were transformed using the function $\log(y+1)$. The additive term was introduced because the simulations can produce concentrations of exactly 0. All model fitting and evaluation was performed on the transformed scale.

Three different designs were used. The first scenario uses exactly the same well coordinates and sample dates as the benzene example discussed in section 5.2. It consists of 1402 observations sampled at 29 well locations. The second scenario uses a much larger number of 280 randomly placed wells which are sampled much less frequently, resulting in the same number of observations. The second scenario is a much better design from a statistical point of view but is, of course, much more expensive, as establishing a new well is considerably most costly than collecting a sample from an existing one. The third scenario uses the same wells as the first scenario, but only has 100 observations in total, with each well sampled only about four times on average.

4.2 The fitted model

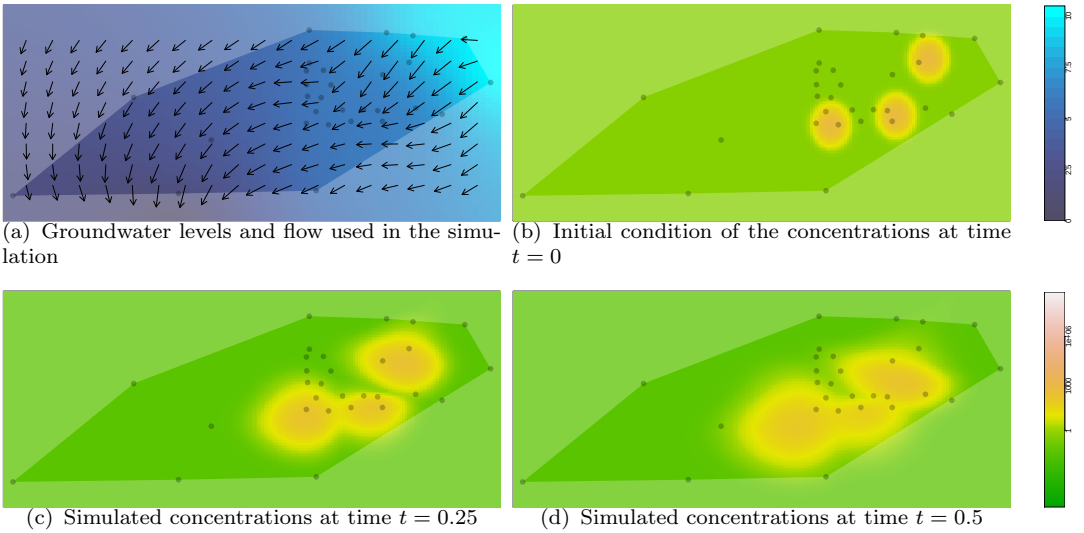
A P-spline model with relaxed assumptions was used (see section 3.7), i.e. a P-spline model with second order basis functions, a first order penalty and 14 basis functions for easting, 8 for northing and 5 for time. The different number of basis functions for space match the different extents of the monitored region in easting and northing in the guiding example, while the reduced number of basis functions for time was chosen to reflect the fact that concentrations vary more quickly in space than in time. Addressing these issues through the basis functions allows a single smoothing parameter to be used in the model in order to achieve computational speed, making it much faster than *mgcv* as mentioned at the end of section 3.8. Where little *a priori* information on solute behaviour is available, a natural default would be to choose a common number of basis function in each dimension. The overall number of basis functions is deliberately chosen to be rather low to allow fast computations. Experimentation has shown these numbers of basis functions to be effective from this perspective, in addition to preventing ballooning or overfitting and producing good estimates of the underlying solute patterns. Under the previously mentioned assumptions, the vector of parameters $Bh\alpha$ has dimension 560 for the three scenarios.

$$\text{MSE} = \frac{1}{N^3} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \left[f(x_{1i}, x_{2j}, t_k) - \hat{f}(x_{1i}, x_{2j}, t_k) \right]^2$$

At each iteration, the optimal value of the smoothing parameter was chosen using AICc, GCV, BIC (see subsection 2.7.1) and Bayesian MAP (see section 3.4) as model selection criteria.

Also 10-fold cross-validation was used in the simulations and in chapter 5. Cross-validation was performed in two different fashions: either by removing entire wells (well-based cross-validation) or by removing single observations (observation-based cross-validation) ignoring the well structure.

Table 4.2 shows the results obtained from 500 replications for all three scenarios. From the table it is immediately clear that no one method outperforms all other methods for all three scenarios.



(a) Groundwater levels and flow used in the simulation

(b) Initial condition of the concentrations at time $t = 0$

(c) Simulated concentrations at time $t = 0.25$

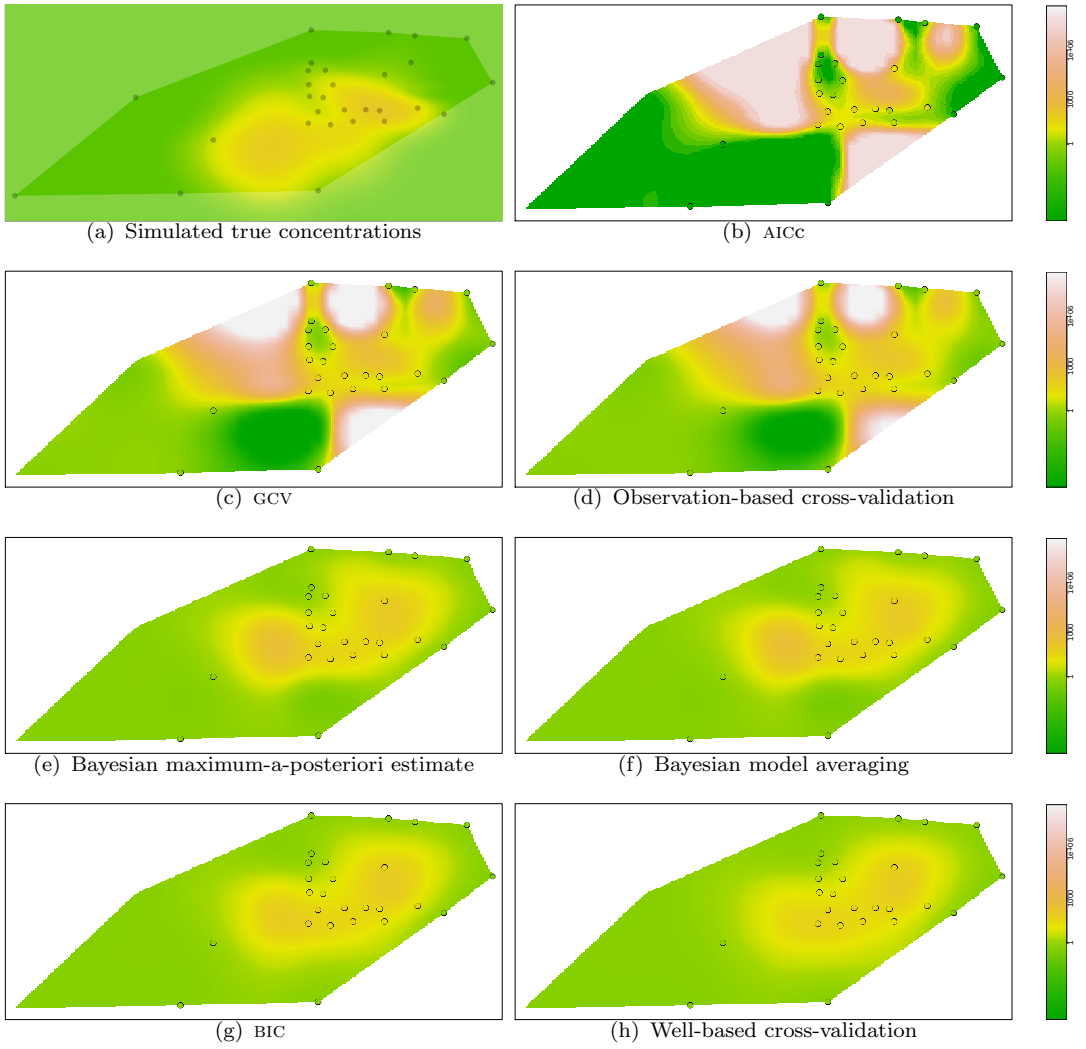
(d) Simulated concentrations at time $t = 0.5$

FIGURE 4.1: Flow model, initial concentrations and simulated concentrations for $t \in \{0, 0.25, 0.5\}$ used in the simulation study. The simulated concentrations for t=0.7 are shown in Figure 4.2(a)

Out of the three scenarios presented, only scenario one is more prone to ballooning, although the figures for GCV under scenario three provide strong suspicion of high unexpected predictions, at least for some iterations. Back to scenario one, AICc and GCV show poor performance. Figures 4.2 (b), (c) and (d) show the reason for the poor performance of these and observation-based cross-validation, as all three lead to severe ballooning. The Bayesian approaches (MAP, Model Averaging and BIC) as well as well-based cross-validation, give much better performance as suggested in Figures 4.2 (e), (f), (g) and (h) where no evidence of ballooning is depicted. All the snapshots in Figure 4.2 are taken at time t=0.7 (in years). The values of the penalisation parameter λ used to produce the plots in Figure 4.2 are listed in Table 4.1.

Figure 4.3 (a) shows density strip plots of the distribution of the smoothing parameter λ for each method. This shows that the Bayesian approaches and well-based cross-validation select values of the smoothing parameter λ which are large enough to prevent ballooning. The problems with other methods are caused by values of λ which are too low.

Though BIC performs very well if the focus in on preventing ballooning, it is prone to underfitting. In the second scenario, which provides the "best" data for estimating the concentrations, BIC performs significantly worse than the other



(a) Simulated true concentrations (b) AICc

(c) GCV (d) Observation-based cross-validation

(e) Bayesian maximum-a-posteriori estimate (f) Bayesian model averaging

(g) BIC (h) Well-based cross-validation

FIGURE 4.2: Simulated true model (top left) as well as predictions obtained in one iteration of the simulation at time t=0.7 using the wells from scenario 1. Each panel corresponds to the use of a different criterion for selecting the smoothing parameter

methods. As Figure 4.3 (b) shows, this is due to selecting a value for λ which is too large.

In all three scenarios, the MAP and the fully Bayesian approach give good results, being the best method in the second and the third scenarios.

Cross-validation is, by far, the most computationally demanding method and the results depend on how it is carried out: Well-based cross-validation favours very large values of the penalty parameter whereas observation-based cross-validation tends to undersmooth by selecting very small values of λ. The reason for the

difference is that, in this data set, ballooning occurs only in space and not in time. There is a relatively small number of wells and these are sampled very frequently in time. Omitting observations individually typically does not create gaps in time which are large enough to allow ballooning at individual wells. Cross-validation can therefore address ballooning only if a well is omitted entirely. The difference between the two variants is much less pronounced in the second and third scenario.

In order to consider the influence of the choice of the number of basis functions, further simulations were carried out. These simulations were all based on the first scenario which, as mentioned, seems to be more prone to ballooning and also corresponds to the real design discussed in section 5.2 in the benzene example.

Tables 4.3, 4.4 4.5 and 4.6 present the mean squared errors and standard errors for the different model selection criteria, based on 500 simulations using the relaxed assumptions (second order basis functions and first order penalty) but varying the number of basis functions (the corresponding values are mentioned in the caption of each table).

As earlier, the different numbers of basis functions in northing and easting aim to reflect the different extents of the monitored region whereas a smaller value is chosen for time as concentrations vary more quickly in space than in time.

The number of basis functions were chosen in such a way that the dimension of $\hat{\boldsymbol{\alpha}}$ roughly doubled for each simulation. Table 4.4 reproduces again the figures corresponding to the first scenario from Table 4.2.

If we consider the Bayesian approaches, we see that MAP and Model Averaging benefit by increasing the number of parameters, whereas it seems to be no noticeable improvement for BIC. The mean squared errors tend to stabilise by increasing the number of basis functions, suggesting that at a certain point these numbers do not yield further enhancement in the model.

Table 4.7 shows the the mean squared errors and standard errors for the different model selection criteria, based on 500 simulations using the standard assumptions

Criterion used to select smoothness	Penalisation parameter λ
AICc	8.521e-8
GCV	1.137e-6
Obs.-based CV	7.969e-3
Bayesian MAP	3.806e-3
BIC	3.577e-1
Well-based CV	4.293e-1

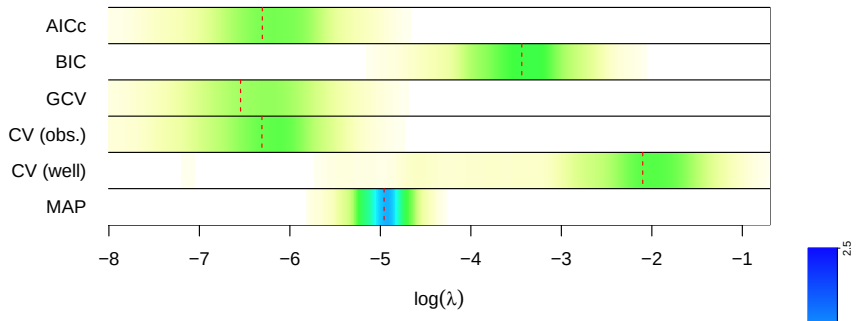
TABLE 4.1: Values of the penalisation parameter λ used to produce the plots in Figure 4.2

Criterion used to select smoothness	Scenario 1		Scenario 2		Scenario 3	
	Mean	(S.E.)	Mean	(S.E.)	Mean	(S.E.)
AICc	214.668	(44.519)	0.221	(0.001)	0.991	(0.006)
GCV	231.481	(26.727)	0.220	(0.002)	8.125	(6.805)
Obs.-based CV	10.829	(1.676)	0.222	(0.001)	1.111	(0.022)
Bayesian MAP	1.304	(0.028)	0.218	(0.001)	0.980	(0.006)
Bayesian model avg.	1.280	(0.027)	0.218	(0.001)	0.979	(0.006)
BIC	0.854	(0.007)	0.317	(0.001)	1.105	(0.005)
Well-based CV	0.870	(0.007)	0.221	(0.001)	1.017	(0.007)

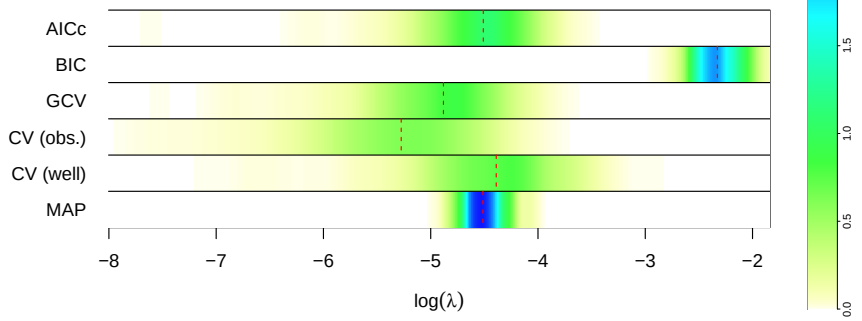
TABLE 4.2: Mean squared errors of the predictions averaged over the convex hull of the data for the three well scenarios

(third order basis functions and second order penalty) but with the same number of basis functions as in Table 4.4 (and Table 4.2). We notice that in this case the ballooning produced using AICc and GCV is less pronounced than in the case of relaxed assumptions, where the Bayesian criteria as well as well-based cross-validation perform better. In other words, relaxing the assumptions improves the methods that already work quite well, but it makes those criteria which are prone to ballooning perform even worse.

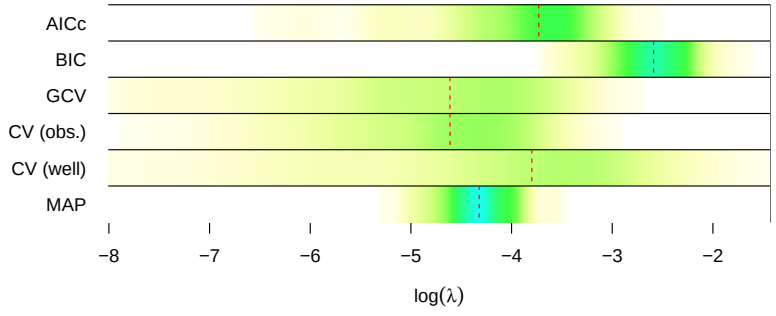
We may conclude that the number of basis functions is, ideally, a technical rather than a smoothing parameter that we try to set as low as possible due to computational effort. Ideally, setting the number of basis functions beyond this threshold would not produce any further benefit.



(a) Well scenario 1



(b) Well scenario 2



(c) Well scenario 3

FIGURE 4.3: Density strip plots of the smoothing parameters chosen by the different methods for both scenarios. The dashed red line indicates the median

Criterion used to select smoothness	Scenario 1	
	Mean	(S.E.)
AICc	1.620e+10	(1.300e+10)
GCV	7.612e+16	(7.611e+16)
Obs.-based CV	74.348	(2.739)
Bayesian MAP	1.512	(0.009)
Bayesian model avg.	1.483	(0.009)
BIC	0.836	(0.005)
Well-based CV	0.857	(0.006)

TABLE 4.3: Mean squared errors of the predictions averaged over the convex hull of the data for Scenario 1 under relaxed assumptions. The number of basis functions used are 10 for easting, 6 for northing and 3 for time. The dimension of the vector $\hat{\boldsymbol{\alpha}}$ is 270

Criterion used to select smoothness	Scenario 1	
	Mean	(S.E.)
AICc	214.668	(44.519)
GCV	231.481	(26.727)
Obs.-based CV	10.829	(1.676)
Bayesian MAP	1.304	(0.028)
Bayesian model avg.	1.280	(0.027)
BIC	0.854	(0.007)
Well-based CV	0.870	(0.007)

TABLE 4.4: Mean squared errors of the predictions averaged over the convex hull of the data for Scenario 1 under relaxed assumptions. The number of basis functions used are 14 for easting, 8 for northing and 5 for time. The dimension of the vector $\hat{\boldsymbol{\alpha}}$ is 560

Criterion used to select smoothness	Scenario 1	
	Mean	(S.E.)
AICc	2.825	(0.780)
GCV	4.100	(0.980)
Obs.-based CV	2.053	(0.452)
Bayesian MAP	0.876	(0.006)
Bayesian model avg.	0.874	(0.006)
BIC	0.880	(0.003)
Well-based CV	0.886	(0.007)

TABLE 4.5: Mean squared errors of the predictions averaged over the convex hull of the data for Scenario 1 under relaxed assumptions. The number of basis functions used are 17 for easting, 10 for northing and 7 for time. The dimension of the vector $\hat{\boldsymbol{\alpha}}$ is 1120

Criterion used to select smoothness	Scenario 1	
	Mean	(S.E.)
AICc	0.881	(0.003)
GCV	0.901	(0.009)
Obs.-based CV	0.886	(0.004)
Bayesian MAP	0.879	(0.003)
Bayesian model avg.	0.879	(0.003)
BIC	0.901	(0.002)
Well-based CV	0.899	(0.004)

TABLE 4.6: Mean squared errors of the predictions averaged over the convex hull of the data for Scenario 1 under relaxed assumptions. The number of basis functions used are 19 for easting, 12 for northing and 10 for time. The dimension of the vector $\hat{\boldsymbol{\alpha}}$ is 2280

Criterion used to select smoothness	Scenario 1	
	Mean	(S.E.)
AICc	28.499	(6.813)
GCV	37.444	(8.014)
Obs.-based CV	5.607	(0.885)
Bayesian MAP	1.577	(0.022)
Bayesian model avg.	1.552	(0.021)
BIC	1.187	(0.012)
Well-based CV	1.100	(0.009)

TABLE 4.7: Mean squared errors of the predictions averaged over the convex hull of the data for Scenario 1 under standard assumptions. The number of basis functions used are 14 for easting, 8 for northing and 5 for time. The dimension of the vector $\hat{\boldsymbol{\alpha}}$ is 756

Chapter 5

Application to Shell data

5.1 Background

Oil companies are compelled by law to control the level of soil contamination produced by their industrial processes. Constituents of crude oil and refined petrol such as benzene, toluene or ethylbenzene can have serious adverse health and ecological effects if released into the environment. A leaking process of an underground container taking place throughout a certain period of time may result in the solute contaminating the groundwater below the storage tank system. After such releases, networks of wells are set up to monitor possible groundwater contamination.

Environmental monitoring data typically has both a spatial structure determined by the location of the monitors, and a temporal one, determined by the frequency with which observations are taken at these locations.

Shell accomplishes this control task by means of an interactive user-friendly graphical software developed by its own staff. This software, called GWSDAT (*Ground-Water Spatio-Temporal Data Analysis Tool*), is aimed at end-users who are not statistically trained practitioners. Its main objective is to add value (cost savings and reduction in environmental liabilities) through improved risk-based decision

making and response. For example by early evaluation of increasing trends over time and space, reduction in the number of sites in long-term monitoring or active remediations through simple demonstrations of groundwater data and trends and efficient evaluation and reporting via standardised plots and tables created at a "mouse click".

Essentially, GWSDAT provides a graphical representation regarding the evolution of a particular spatio-temporal data set of observations. This software uses Microsoft Excel as the primary user interface and data entry platform whereas the underlying statistical calculations and graphical output were developed using the open source statistical program *R* and the package *rpanel* (see Bowman et al., 2007) for plots.

The statistical issues were initially tackled by means of the **Support Vector Machines (SVM)** approach, introduced by Vapnik and Chervonenkis in 1964. SVM (see Schölkopf and Smola, 2002; Smola and Schölkopf, 2004) comprise a set of algorithms whose main features are the usage of kernels, convex optimisation, sparseness of the solution and the possibility of influencing such solution by means of the so-called support vectors. This techniques were originally aimed at solving classification problems; subsequently, these ideas were extended to the case of smooth regression.

But the SVM approach does not allow to produce any confidence intervals. In addition, SVMs are prone to the undesired effect of ballooning, in our experience more so than spline-based models. Changing the kernel to a kernel less smooth than the Gaussian kernel might make SVMs more resilient to ballooning.

Shell decided consequently to endow GWSDAT with more robust functionalities by means of a model relying on strongly backed-up theoretical foundations, which best describes the evolution of groundwater solutes in time and space in the context of data collected by Shell and/or its affiliates. In addition, this model should meet both the following conditions: be fast enough to run interactively (i.e. in a few seconds) even with large data sets and yield a reliable level of uncertainty in the computed predictions so as to include them in the graphical interface with the end-user.

Criterion used to select smoothness	Penalisation parameter λ
GCV	7.961e-8
AICc	1.034e-7
Obs.-based CV	1.670e-5
Bayesian MAP	1.220e-3
BIC	7.394e-3
Well-based CV	2.630e-1

TABLE 5.1: Values of the penalisation parameter λ computed under standard assumptions for the different criteria for selecting the smoothing parameter

5.2 Case Study

This case study corresponds to the same set-up described in the first scenario used in chapter 4 but with actual data related to the solute benzene, a constituent of crude oil and refined petrol, for which "the truth" is unknown. As described in the aforementioned chapter, the design consists of 1402 observations out of which 362 are below the detection threshold and hence they were replaced by one-half the detection limit. These observations were sampled at 29 wells locations between October 15th, 1987 and November 25th, 2009 with observations recorded at irregular time intervals. The observed values correspond to the concentration of the solute measured in $\mu g/\ell$ (modelled on a log-scale).

We start by fitting the data using a P-spline model with the standard assumptions and by replacing the non-detects by one-half the reported detection limit.

Table 5.1 reports the value of the penalisation parameter λ computed using the different criteria for selecting the smoothing parameter. Figures 5.1, 5.2, 5.3, 5.4, 5.6 and 5.7 show the effect of the smoothing at the same point in time (t=16.44, in years) for the criteria in this table. Figure 5.5 corresponds to the Bayesian model averaging criterion. The triangles in the figures represent the wells in which non-detects were recorded.

The entries in Table 5.1 are listed in increasing order of the penalisation parameter λ. Similarly, the plots are presented in the same order.

As expected from the results obtained in chapter 4 for the first scenario, the first three model selection criteria (GCV, AICc and observation-based cross-validation) tend to overfit the data producing the undesired effect of high unexpected predicted values; BIC and well-based cross-validation avoid such effect by picking up large values for the penalisation parameter λ.

For the Bayesian criteria (MAP and model averaging) we notice that they perform much better than GCV, AICc and observation-based cross-validation although they do worse than BIC and well-based cross-validation because ballooning cannot be completely eliminated.

Under this context of standard assumptions, the optimal value for the penalisation parameter appears to be between those corresponding to the MAP and BIC criteria. It should be noticed that the value of λ for the BIC criterion (7.394e-3) is almost 6 times larger than the one corresponding to the MAP criterion (1.220e-3) suggesting that ballooning is avoided at expense of oversmoothing.

The uneven design of this data set is responsible for these inappropiate high predicted values: we have equally spaced basis functions but not equally spaced data. In addition, non-detects create artificial signals which are not properly dealt with.

We might have thought of a P-splines design with unequally spaced basis functions giving more flexibility (i.e. a greater number of greater functions) where there is more data. But such approach would have made very difficult to deal with the penalty.

We have proposed to tackle the issue of ballooning by using the relaxed assumptions approach (see section 3.7). The next section covers more in detail the issue of ballooning using other scenarios provided by Shell, and we revisit our case study in section 5.4 under the framework of the relaxed assumptions.

Chapter 6 addresses the issue of non-detects and again we reconsider our case study under the approach proposed in that chapter.

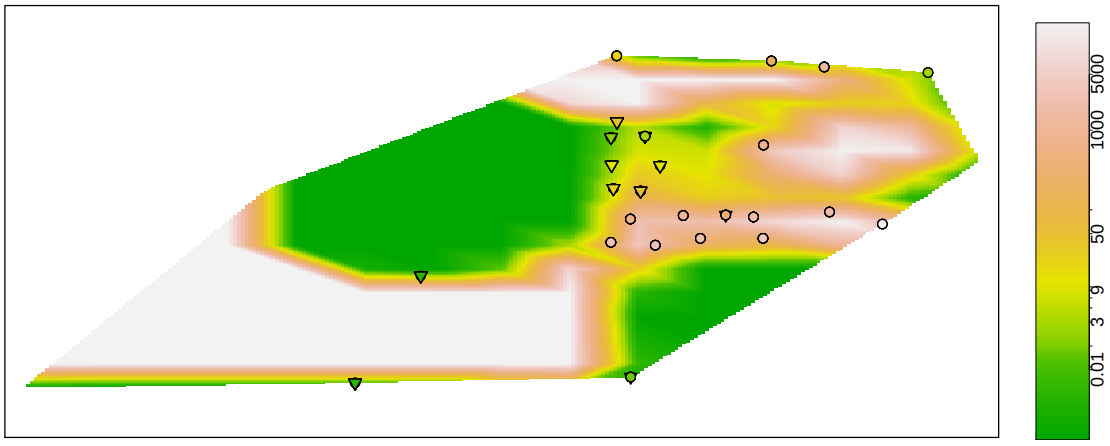


FIGURE 5.1: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=7.961e-8 was computed using the GCV criterion under the standard assumptions (triangles represent non-detects and circles correspond to observed data)

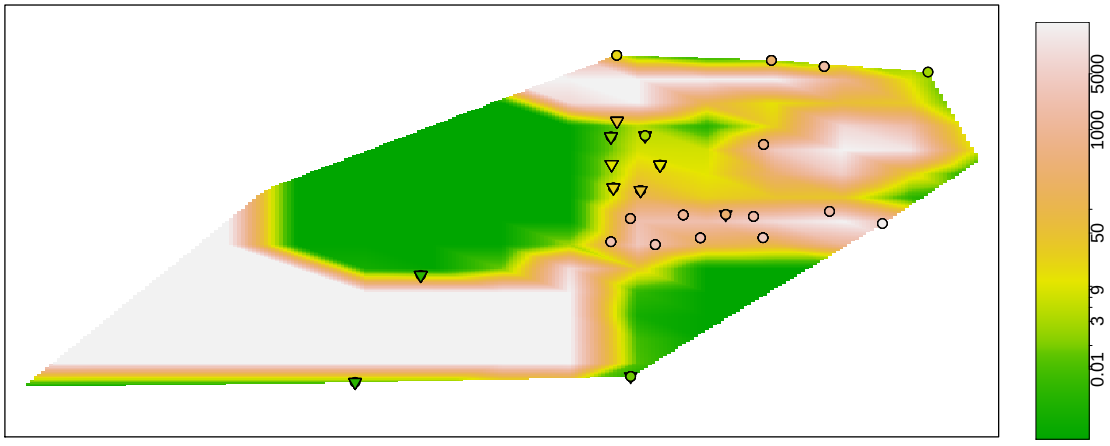


FIGURE 5.2: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=1.034e-7 was computed using the AICc criterion under the standard assumptions (triangles represent non-detects and circles correspond to observed data)

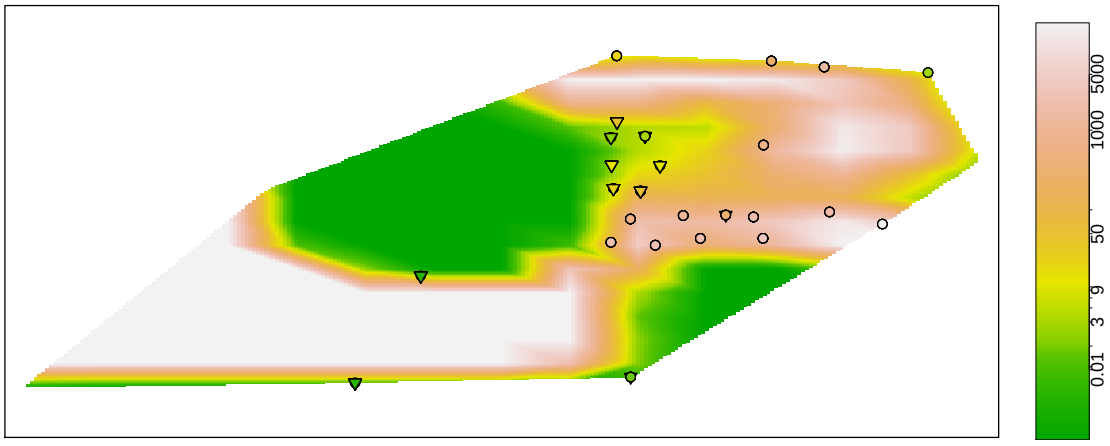


FIGURE 5.3: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=1.670e-5 was computed using the observation-based CV criterion under the standard assumptions (triangles represent non-detects and circles correspond to observed data)

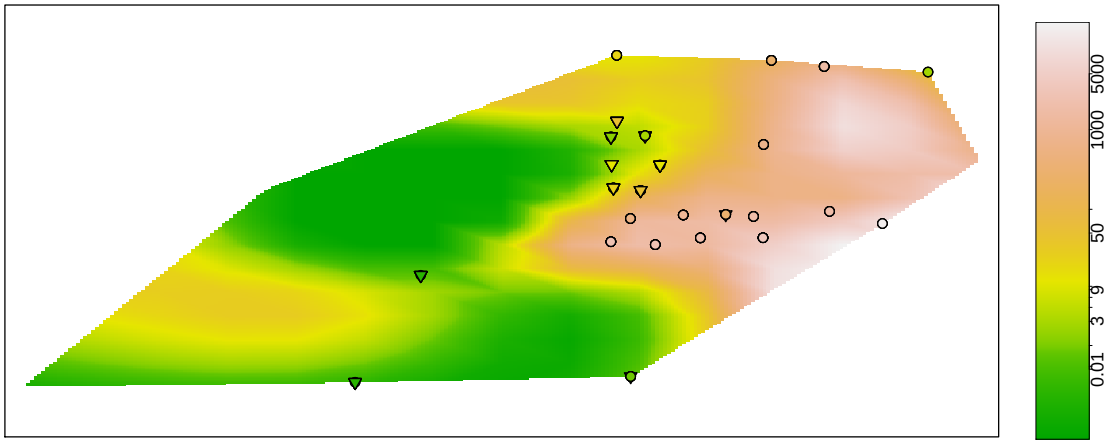


FIGURE 5.4: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=1.220e-3 was computed using the Bayesian MAP criterion under the standard assumptions (triangles represent non-detects and circles correspond to observed data)

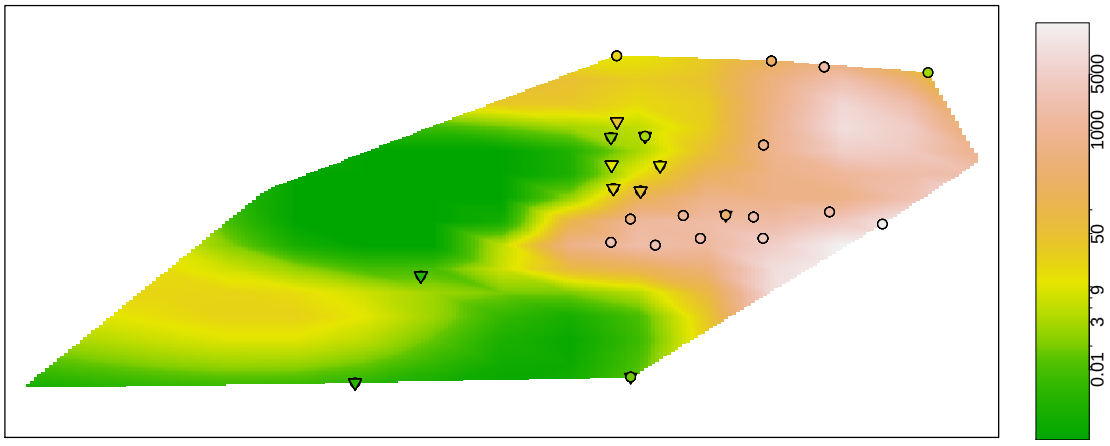


FIGURE 5.5: Predictions obtained for the real case study at time t=16.44. It corresponds to the Bayesian model averaging smoothing criterion under the standard assumptions (triangles represent non-detects and circles correspond to observed data)

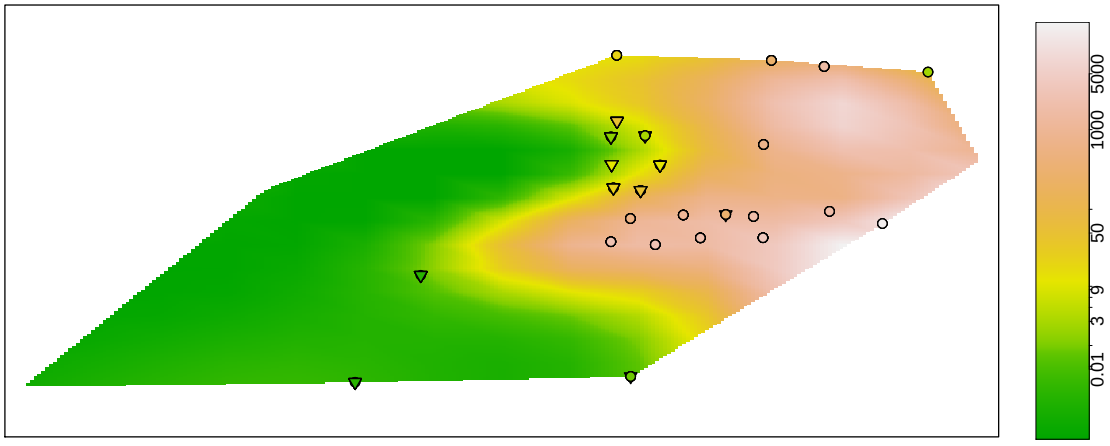


FIGURE 5.6: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=7.394e-3 was computed using the BIC criterion under the standard assumptions (triangles represent non-detects and circles correspond to observed data)

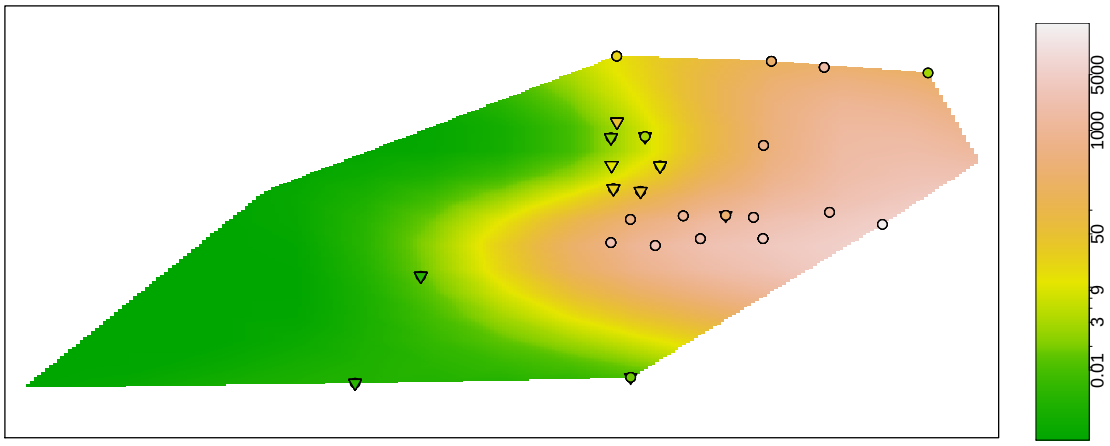


FIGURE 5.7: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=2.630e-1 was computed using the well-based CV criterion under the standard assumptions (triangles represent non-detects and circles correspond to observed data)

5.3 Ballooning

In section 3.7 we discussed the problem of ballooning and proposed a strategy to tackle it. In this section we will try to find out where ballooning comes from and present the results of the application of this strategy on some concrete spatio-temporal data in connection with the contamination process described in the first section of this chapter.

Only in this particular section, we will use two new different designs that we will call *Scenario A* and *Scenario B* (which correspond to two different data sets), to avoid any possible confusion with other scenarios mentioned in this thesis.

Scenario A has 26 wells and spans over time from July 18th, 1977 to October 6th, 2011 with observations recorded at irregular time intervals. As usual, the observed values correspond to the concentration of the solute measured in $\mu g/\ell$ modelled on a log-scale.

Two different contaminants will be considered under Scenario A: benzene (637 observations with 301 non-detects) and MTBE (637 observations with 163 non-detects). The plots for this scenario correspond to time t=13.41 (in years).

Figure 5.8 (top) shows the result of the predictions of benzene under Scenario A by selecting the optimal MAP value for the penalisation parameter under the standard assumptions.

The mismatch between the values reported at the wells and the predictions is evident. There are areas with not only extremely high unexpected values but also with extremely low predicted concentrations. Although, due to the log-transform used, the first case is the one with the worse practical implications, it is clear that the standard MAP technique presents serious flaws for some particular data sets.

Figure 5.8 (bottom) depicts a noticeable improvement in the predictions by manually selecting a higher value of the penalisation parameter λ. The picture suggests that although there might exist an optimal value for the penalisation parameter, due to some reason, the MAP procedure fails to pick it up correctly.

A more detailed analysis in Figure 5.9 (top) shows a group of three wells which consistently get "too low" concentrations in comparison with the values observed in their neighborhood. Figure 5.9 (bottom) pictures the predicted values using the MAP technique after having removed these three wells: the improvement in the expected predictions seems to suggest that these unusual observations were at the root of the problem of ballooning for this data set.

As mentioned at the beginning of subsection 2.5.2, standard P-splines models assume that the underlying signal in the data changes very slowly. In this particular setting, this assumption implies that depending on the network design, wells might be "trusted" to give a reliable indication of the gradient/curvature variation in their vicinity. In other words, as we indicated in advance in section 3.7, the root cause of ballooning stems in the mismatch between the smoothness in the signal assumed by the P-splines model and the actual data.

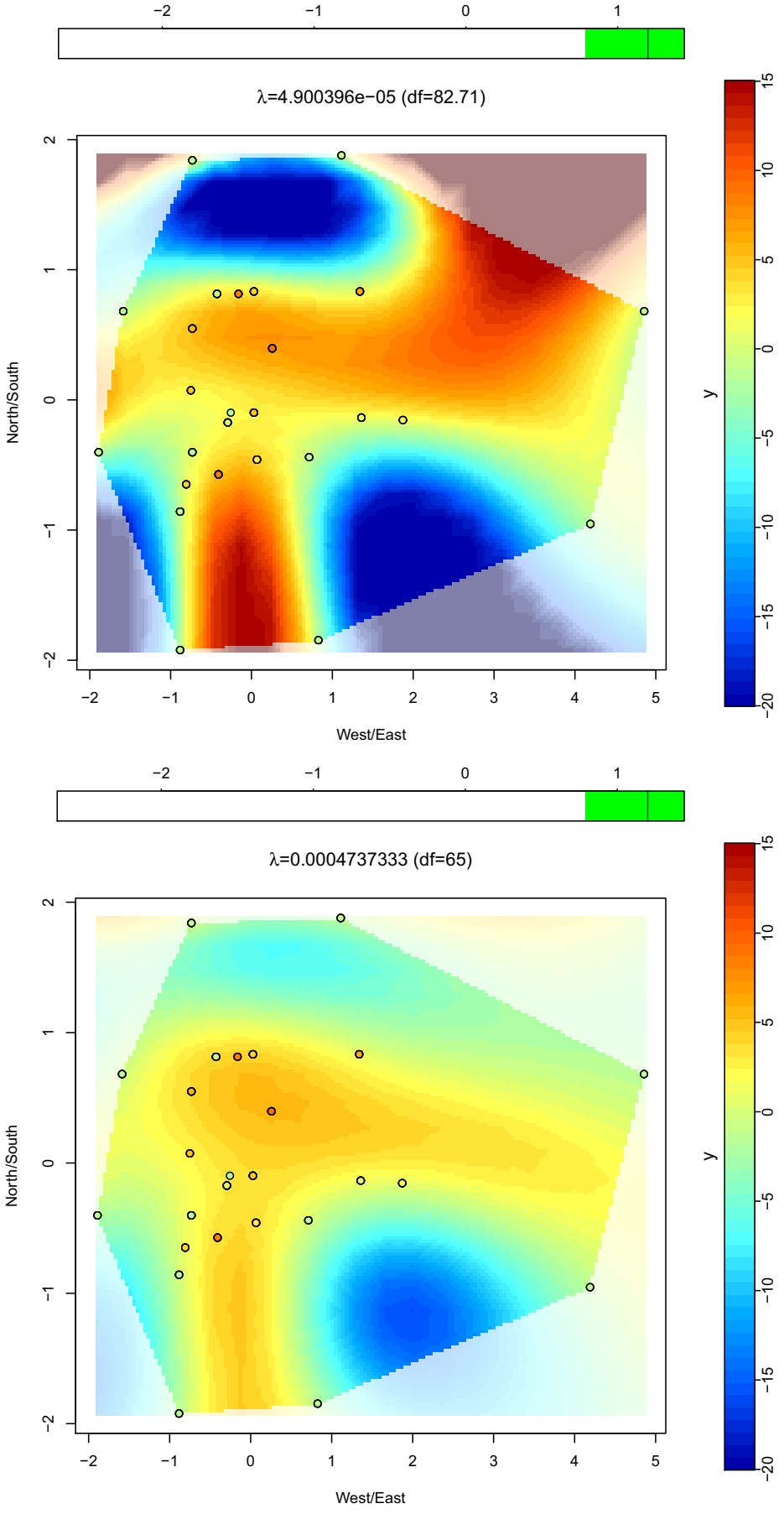


FIGURE 5.8: Benzene (Scenario A) - Standard Assumptions - λ selected automatically (top) - λ tuned manually (bottom) - (top bar indicates time)

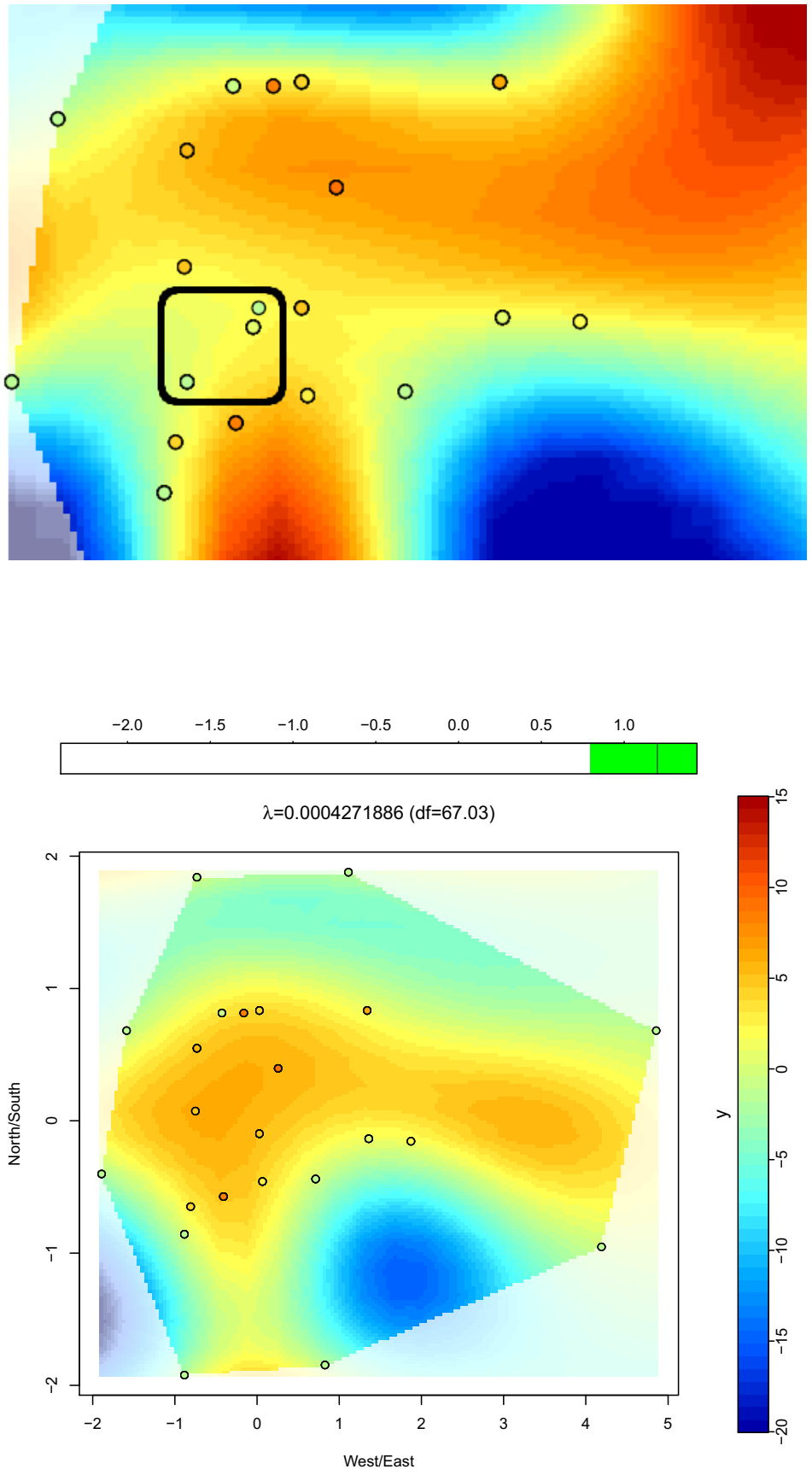


FIGURE 5.9: Benzene (Scenario A) - Standard Assumptions - Wells with too low concentrations (top) - λ tuned automatically after deleting the "problematic" wells (bottom) - (top bar indicates time)

As a first approach we might tackle the problem by trying to use a less complex model, i.e. by decreasing the degrees of freedom given by the MAP technique. But this action can cause the model to be too inflexible and to underestimate peak concentrations. In addition, shifting the value of λ might require a very informative prior, which would generalise poorly to other data sets.

Instead of dealing with the "symptoms" of the problem, we propose an intervention at the underlying cause of ballooning by relaxing the smoothness assumptions of the P-splines model. As described in section 3.7, in practice this means changing the "defaults settings" used under the standard model.

Figure 5.10 (top) shows the result of the MAP technique by relaxing only the conditions on the order of the differences applied to the penalty and the order of the polynomials making up the splines but with a low number of functions in the basis. The same figure at the bottom pictures the predictions using cubic P-splines and a quadratic penalty at the same degrees of freedom. It can be noticed that relaxing the first two conditions has little effect in this case.

Figure 5.11 displays the same situation but using a larger number of basis functions. A comparison between the pictures suggests a remarkable improvement if the penalisation parameter is chosen with the MAP approach by jointly relaxing the three conditions as proposed in section 3.7.

Two additional examples similar to Figure 5.11 are provided supporting the methodology proposed to tackle the issue of ballooning. Figure 5.12 shows the effect of using the standard and relaxed assumptions in fitting the data corresponding to a different contaminant (MTBE) under Scenario A.

Figure 5.13 displays the same comparison for benzene data under Scenario B. This scenario is made up of 27 wells extending from May 25th, 1999 to February 15th, 2011 with observations recorded also at irregular time intervals. Only one solute (benzene) is used in this scenario comprising 602 observations with 482 non-detects and the snapshots taken at time t=5.74 (in years). As usual, the concentration of the solute is measured in $\mu g/\ell$ modelled on a log-scale.

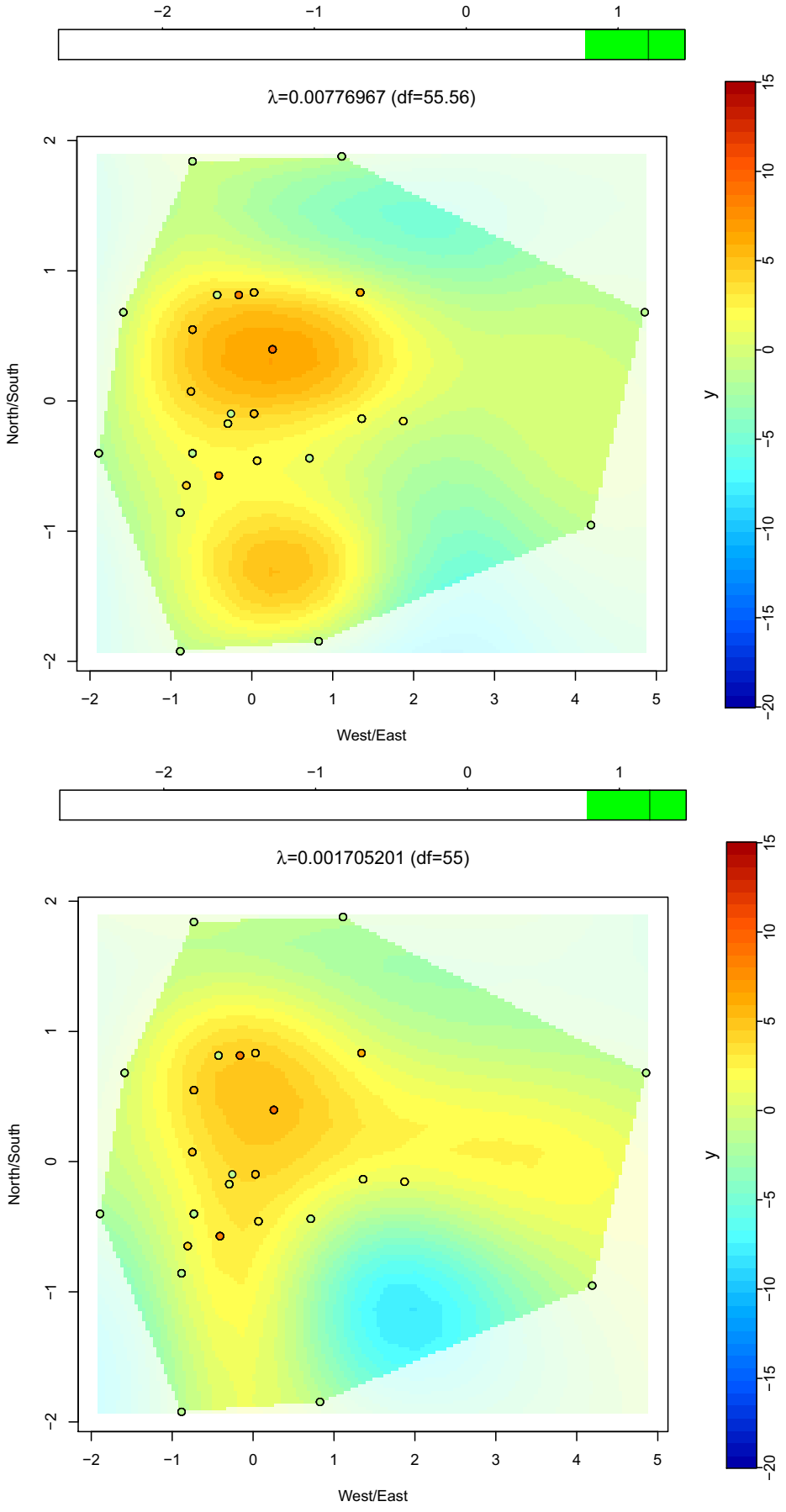


FIGURE 5.10: Benzene (Scenario A) - By relaxing assumptions 1) and 2) only (top) - Standard assumptions using the same df (bottom) - (top bar indicates time)

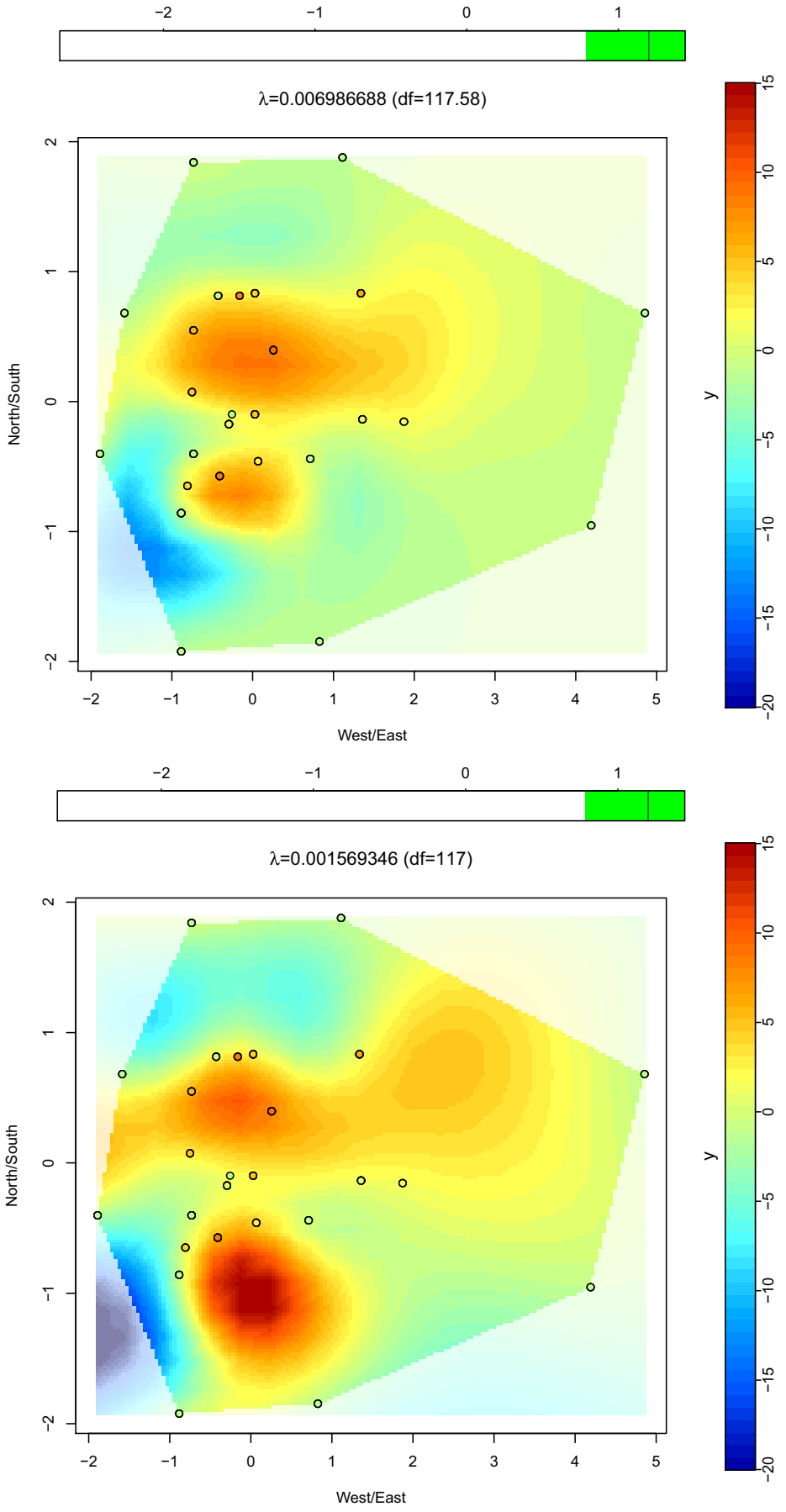


FIGURE 5.11: Benzene (Scenario A) - Relaxed assumptions (top) - Standard assumptions 1) and 2) only, using the same number of basis functions (bottom) - (top bar indicates time)

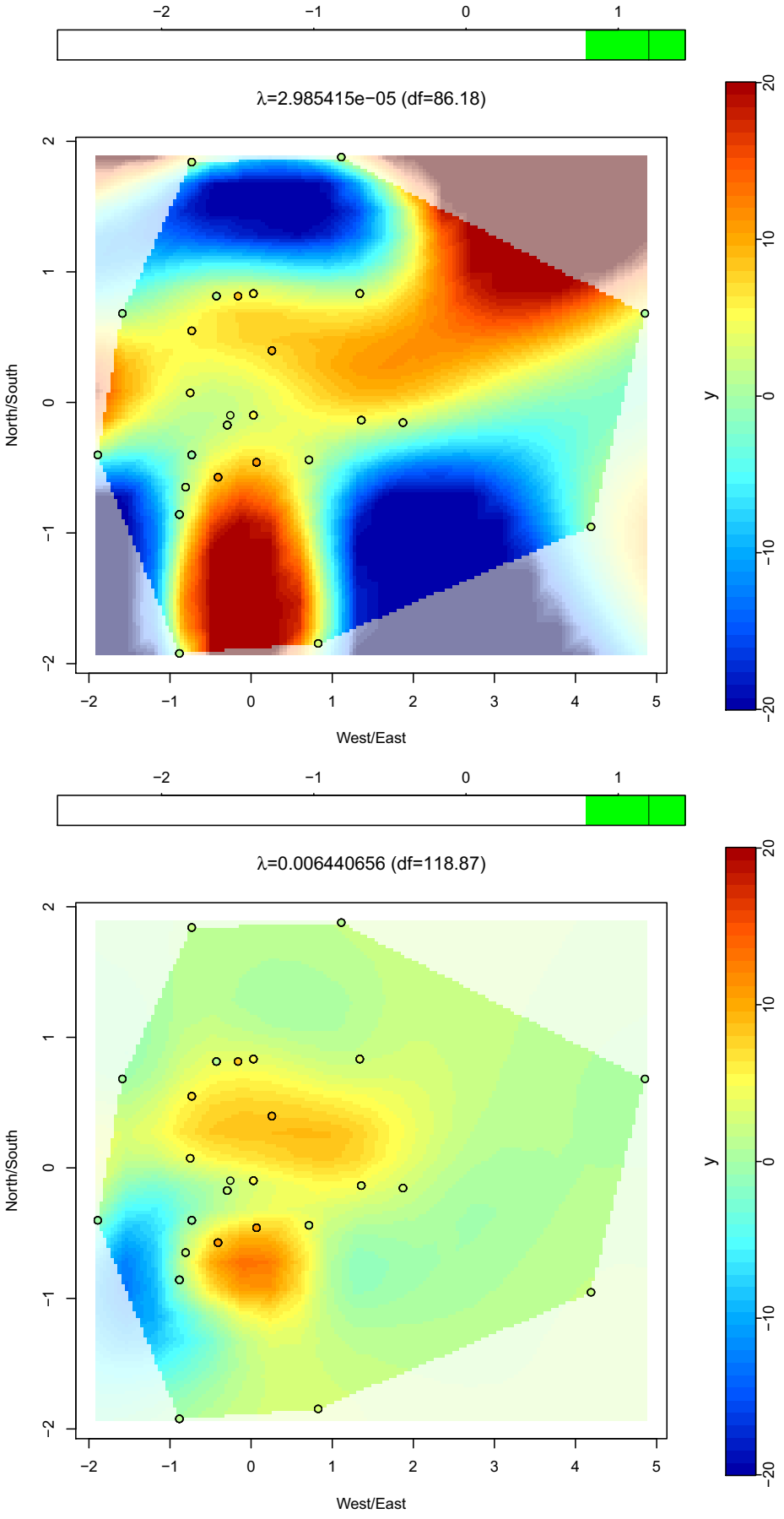


FIGURE 5.12: MTBE (Scenario A) - Standard assumptions (top) - Relaxed assumptions (bottom) - (top bar indicates time)

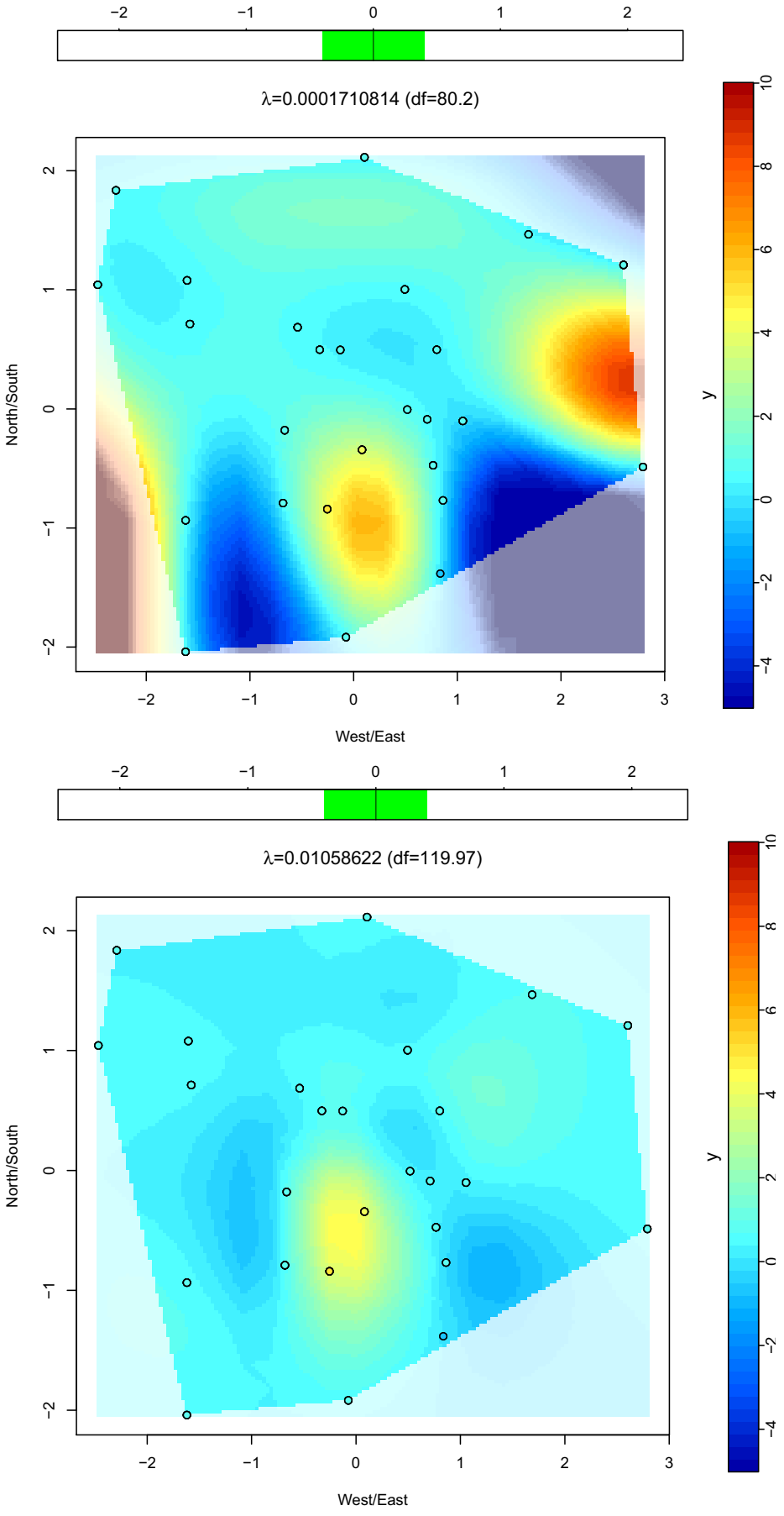


FIGURE 5.13: Benzene (Scenario B) - Standard assumptions (top) - Relaxed assumptions (bottom) - (top bar indicates time)

It is worth mentioning that another way of addressing the problem of ballooning is by fitting a mixed model with a random effect for each well. But this approach would not be feasible under the time constraints imposed as the performance would have been much more slow. Additionally, a mixed model assumes that the random effect is constant over time, which might not be true in this setting.

5.4 Case Study Revisited

In this section we reconsider our case study from section 5.2 under the framework of relaxed assumptions for our P-spline model, but still replacing non-detects by one-half the reported detection limit.

Table 5.2 reports the value of the penalisation parameter λ computed using the different criteria for selecting the smoothing parameter. Figures 5.15, 5.14, 5.16, 5.17, 5.19 and 5.20 show the effect of the smoothing at the same point in time (t=16.44, in years) for the criteria in the mentioned table. Figure 5.18 corresponds to the Bayesian model averaging criterion.

As earlier, the entries in Table 5.2 are listed in increasing order of the penalisation parameter λ, the plots are presented in the same order and the triangles in figures represent the wells in which non-detects were recorded.

As expected, under this framework of relaxed assumptions, the Bayesian MAP and Bayesian model averaging criteria for model selection, perform very well without producing the effect of ballooning in the area where little data are present. Notice also that the value of λ for the Bayesian MAP criterion (4.108e-3) is almost 2.5 times smaller than the one corresponding to the well-based cross-validation criterion (9.812e-3). It should be also noticed the improvement for the AICc, GCV and observation-based cross-validation under this framework, although they do not manage to completely avoid the extremely high unexpected values.

Criterion used to select smoothness	Penalisation parameter λ
GCV	3.473e-7
AICc	4.935e-7
Obs.-based CV	2.043e-5
Bayesian MAP	4.108e-3
Well-based CV	9.812e-3
BIC	1.455e-2

TABLE 5.2: Values of the penalisation parameter λ computed under relaxed assumptions for the different criteria for selecting the smoothing parameter

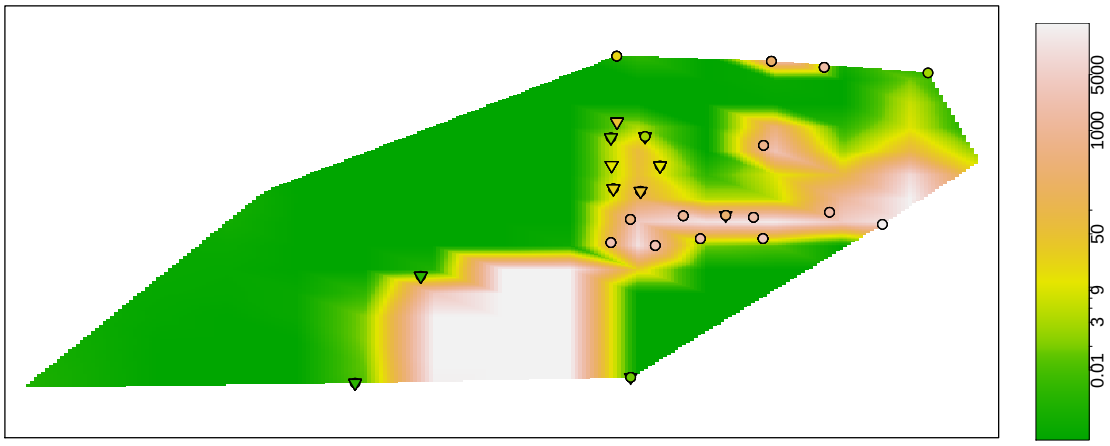


FIGURE 5.14: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=3.473e-7 was computed using the GCV criterion under the relaxed assumptions (triangles represent non-detects and circles correspond to observed data)

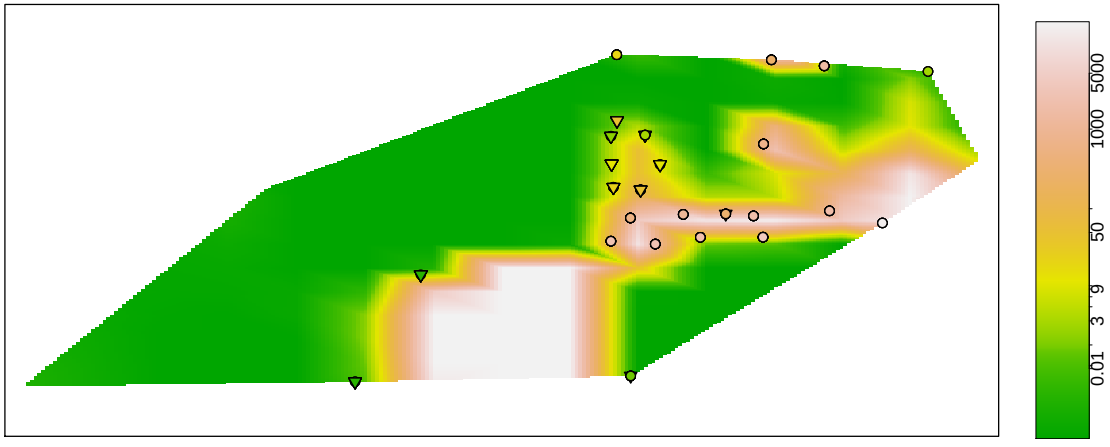


FIGURE 5.15: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=4.935e-7 was computed using the AICc criterion under the relaxed assumptions (triangles represent non-detects and circles correspond to observed data)

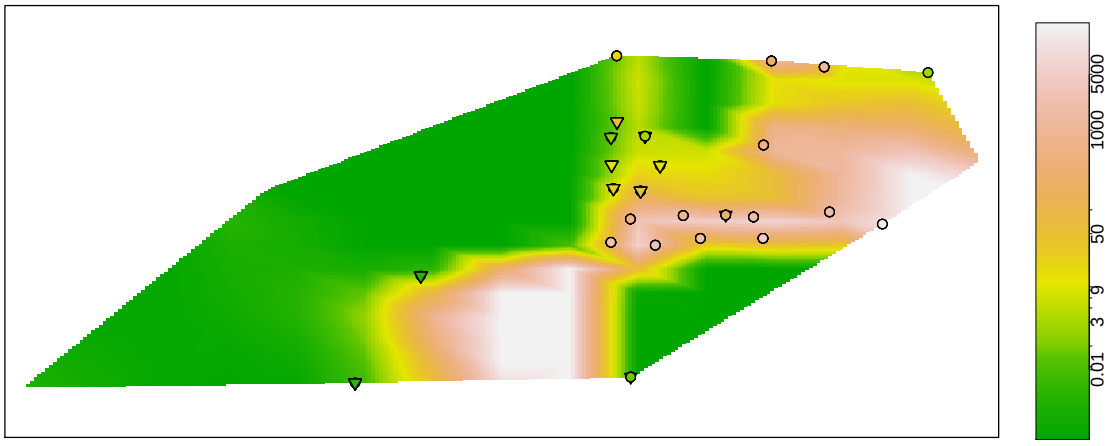


FIGURE 5.16: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=2.043e-5 was computed using the observation-based CV criterion under the relaxed assumptions (triangles represent non-detects and circles correspond to observed data)

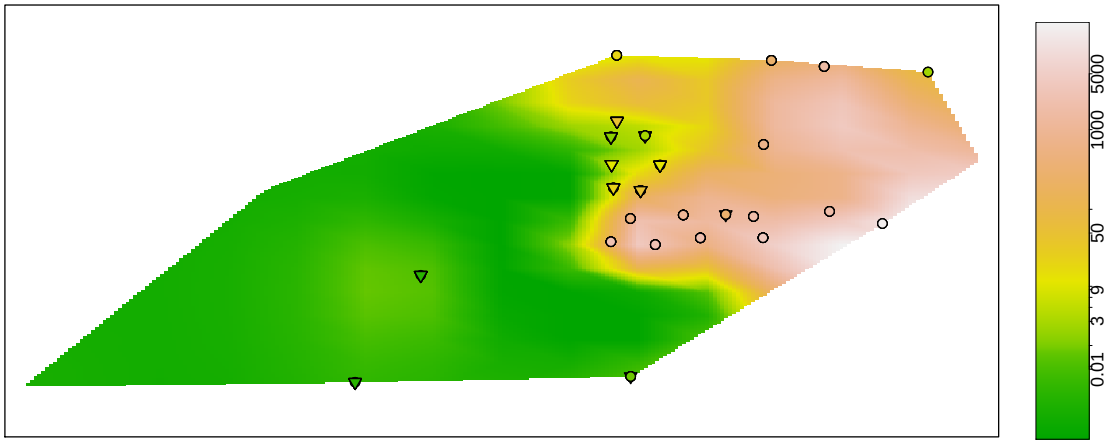


FIGURE 5.17: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=4.108e-3 was computed using the Bayesian MAP criterion under the relaxed assumptions (triangles represent non-detects and circles correspond to observed data)

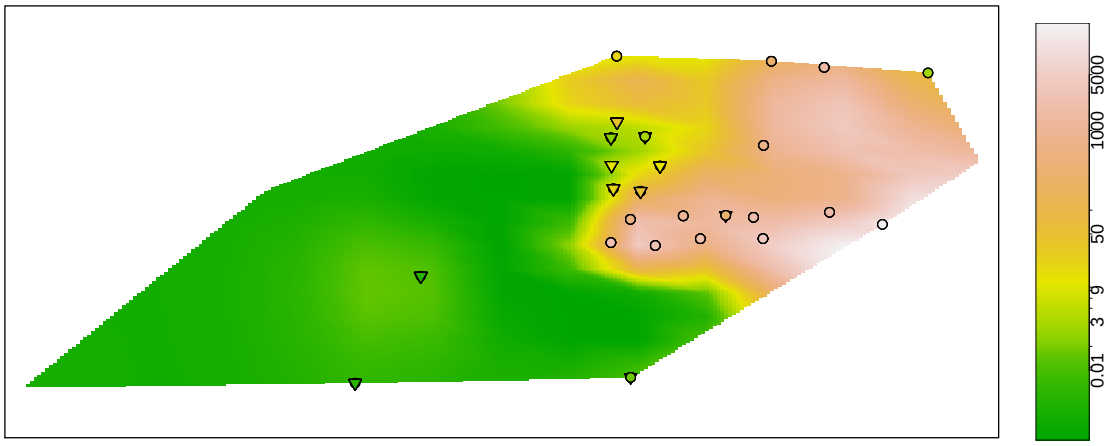


FIGURE 5.18: Predictions obtained for the real case study at time t=16.44. It corresponds to the Bayesian model averaging smoothing criterion under the relaxed assumptions (triangles represent non-detects and circles correspond to observed data)

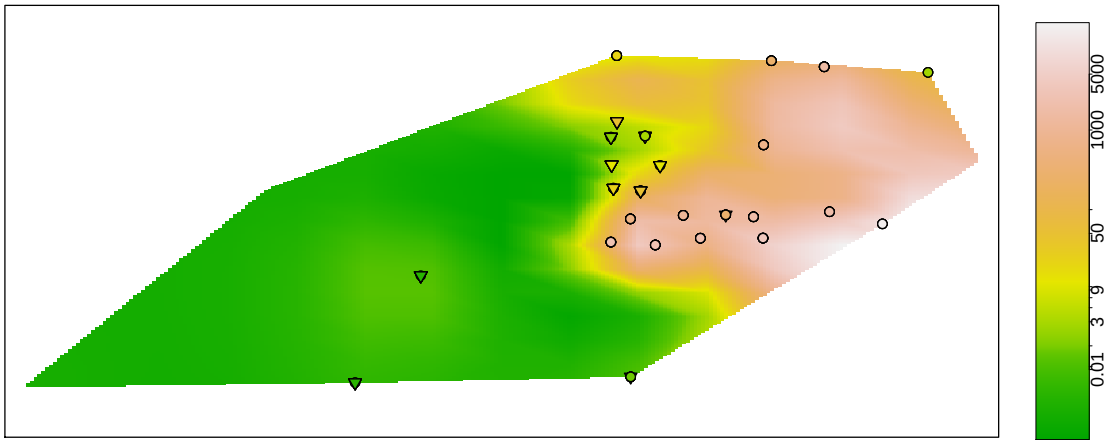


FIGURE 5.19: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=9.812e-3 was computed using the well-based CV criterion under the relaxed assumptions (triangles represent non-detects and circles correspond to observed data)

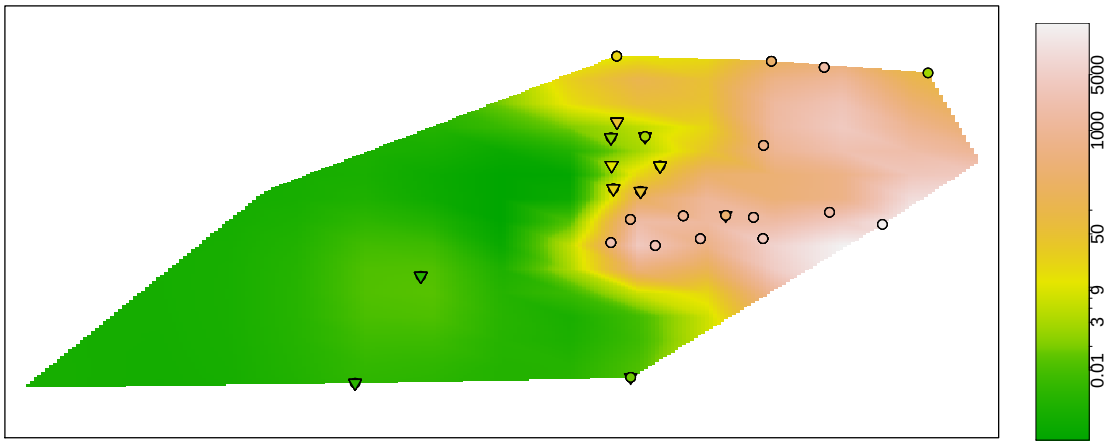


FIGURE 5.20: Predictions obtained for the real case study at time t=16.44. The penalisation parameter λ=1.455e-2 was computed using the BIC criterion under the relaxed assumptions (triangles represent non-detects and circles correspond to observed data)

5.5 Additional Example

A more extensive example of the use of these techniques is provided by retrospective analysis of a data set on a pollution event at a refinery site. MTBE (methyl tertiary butyl ether) is a petrol additive designed to reduce engine knocking and noxious emissions. MTBE is no longer in routine use at the site studied but was present in the refinery at the time of the event. On entry to groundwater, MTBE moves conservatively due to its high aqueous solubility and low retardation potential. It degrades only slowly under anaerobic conditions. Figure 5.21 shows a schematic plan of the site with colour-coded points to indicate the concentrations of MTBE measured at the monitoring wells at a date near the time of the MTBE release. Standard methods of analysis in this setting were to inspect individual well measurements over time to identify trends. Geographical information systems were available and these were helpful for individual time snapshots but these could not easily be adapted to show the evolving dynamics of the incident.

Figure 5.22 shows the estimated pollution surface using the Bayesian smoothing model described in chapter 3, using 18 basis functions for easting, 22 basis functions for northing, 14 basis functions for time and the MAP estimate of λ. The

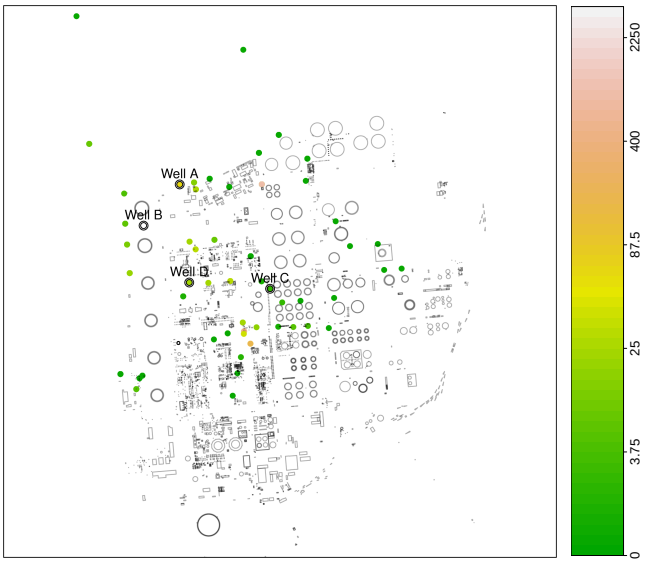
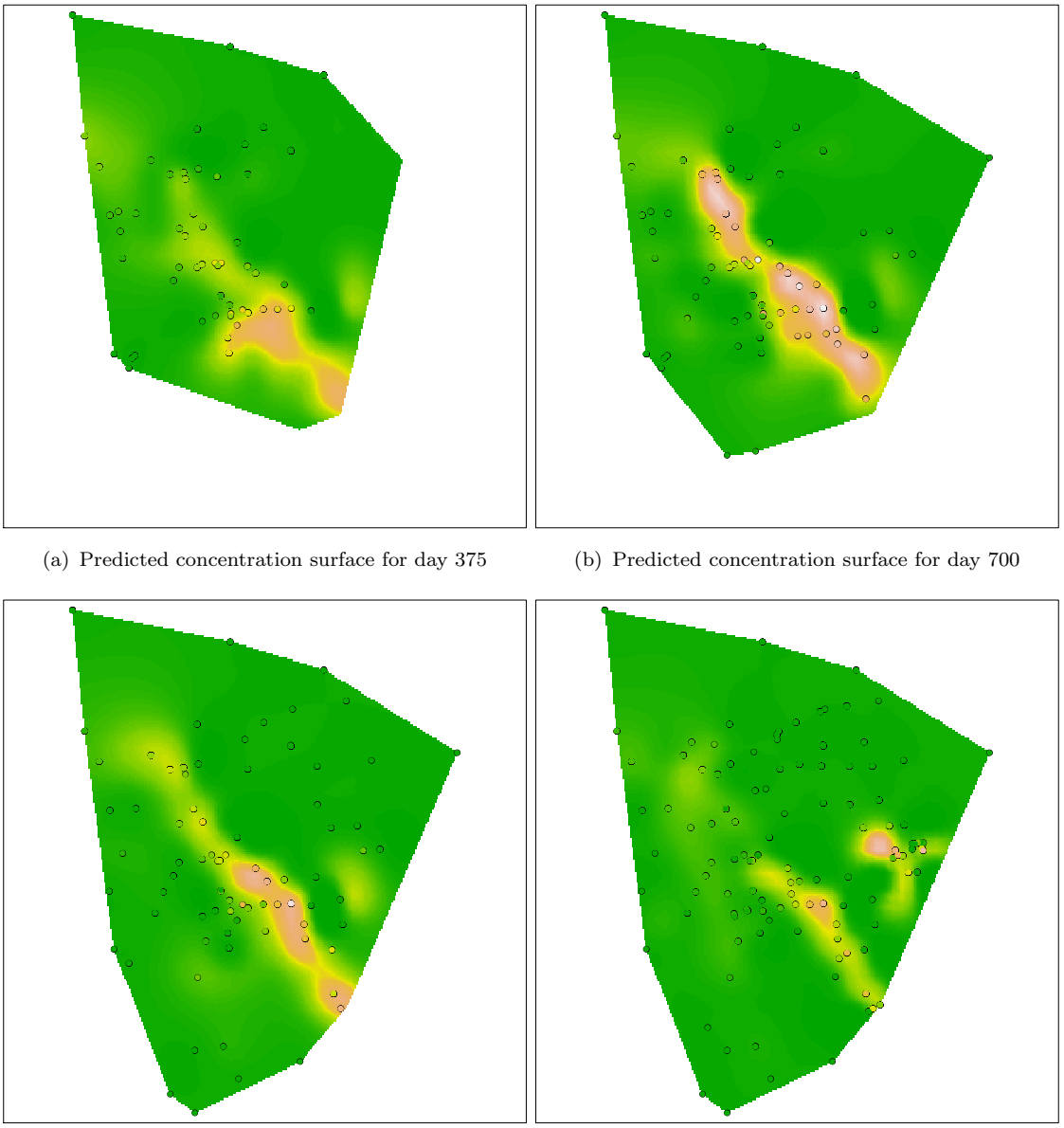


FIGURE 5.21: Plan of the refinery site and wells. The wells are colour-coded according to observed concentrations of MTBE immediately after release.

shape and direction of the plume is clear and consistent with the south-east/north-west gradient in groundwater flow. Despite the presence of protective pumping wells at the north-west boundary of the refinery site, the threat of MTBE migrating across the site boundary and potentially reaching drinking water wells required immediate action.



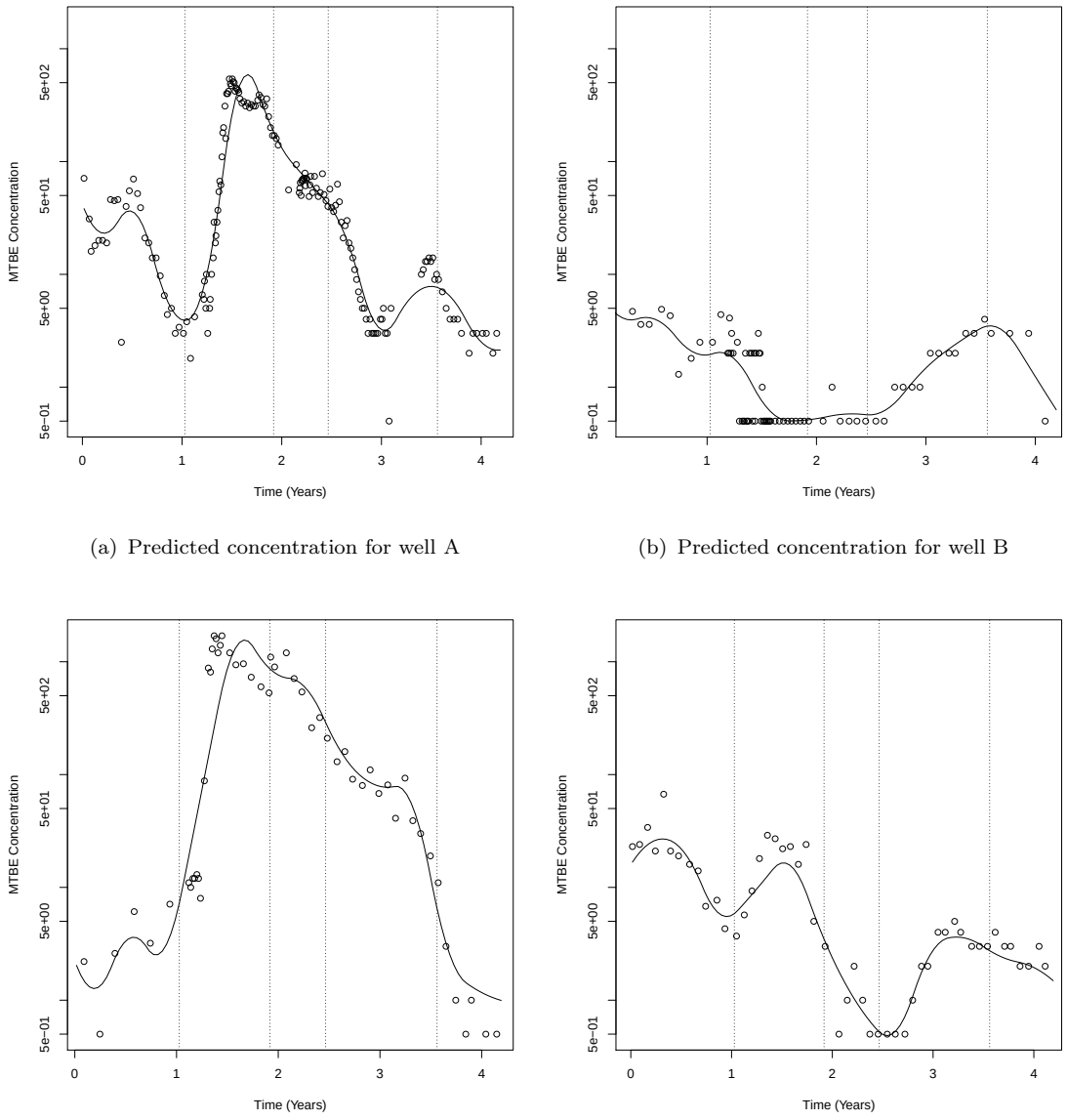
(a) Predicted concentration surface for day 375

(b) Predicted concentration surface for day 700

(c) Predicted concentration surface for day 900

(d) Predicted concentration surface for day 1300

FIGURE 5.22: Predicted levels of MTBE concentration across space obtained using the MAP estimate of the smoothing parameter for four time points. The colour scale is the same as that used in Figure 5.21



(a) Predicted concentration for well A

(b) Predicted concentration for well B

(c) Predicted concentration for well C

(d) Predicted concentration for well D

FIGURE 5.23: Predicted levels of MTBE concentration over time obtained using the MAP estimate of the smoothing parameter for four wells. The location of the wells is shown in Figure 5.21. The vertical dotted lines correspond to the time points used in Figure 5.22.

The panels of Figure 5.22 show estimates from the spatiotemporal MTBE distribution model at several further time points. The first corresponds to the upgrading of a line of wells used to form a flow barrier in the middle of the site. The effectiveness of these wells was greatly improved and the resulting curtailment of the plume to the north-west is apparent. Subsequently, the source of the MTBE release was identified near the south-east corner of the site and the model clearly tracks the dissipation and attenuation of MTBE and the end of the incident.

Figure 5.23 pictures the predicted levels of MTBE concentration over time for four wells located in the north-west area of the refinery.

5.6 Uncertainty Quantification

This section is devoted to the uncertainty quantification of the spatio-temporal predictions discussed in Sections 5.2 and 5.4. The patterns to be described generalise to the remaining spatio-temporal examples in the present work.

Figures 5.24 and 5.26 picture the 95% lower and upper confidence bands under standard and relaxed assumptions respectively, for the predictions described in the aforementioned sections. The corresponding standard errors are pictured in Figures 5.25 and 5.27 (the legends on the right correspond to the standard errors). It can be noticed that the lower is the value of the penalisation parameter λ, the larger are the standard errors. In other words, ballooning seems to affect massively the standard deviation whereas no great differences occur for large values of λ.

Conversely, larger values of the penalisation parameter yield smaller values for the variance and more biased predictions.

It is known that the eigenvector corresponding to the largest eigenvalue of the variance-covariance matrix of the estimated coefficients, namely $Var\left(\hat{\boldsymbol{\alpha}}\right)$, points to the direction of maximal variance in the parameter space, whose value is given by the eigenvalue itself.

Tables 5.3 and 5.4 reproduce Tables 5.1 and 5.2 with the addition of the largest eigenvalue of the variance-covariance matrix of $\hat{\boldsymbol{\alpha}}$ corresponding to the different penalisation parameter criteria. It can be noticed that the larger the value of λ, the smaller is the maximal variance.

In our context, the confidence intervals and standard errors were computed using the corresponding posterior predictive distribution. For a given new matrix of regressors $\tilde{\boldsymbol{B}} \in \mathbb{R}^{r\times m}$, the predicted outcomes would be described as $\tilde{\boldsymbol{Y}}|\boldsymbol{\alpha},\, \sigma^2,\, M_\lambda \sim \mathcal{N}_r\left(\tilde{\boldsymbol{B}}\boldsymbol{\alpha},\, \sigma^2\boldsymbol{I}_m\right)$. The posterior predictive distribution is given by

$$\begin{aligned} f_{\tilde{\boldsymbol{Y}}|\boldsymbol{Y},\lambda} &= \int f(\tilde{\boldsymbol{y}}|\boldsymbol{\alpha},\, \sigma^2,\, \lambda) \,\times\, f^*(\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{y},\lambda)\, d\boldsymbol{\alpha}d\sigma^2 \\ &= \int \mathcal{N}_r\left(\tilde{\boldsymbol{B}}\alpha,\, \sigma^2\boldsymbol{I}_r\right) \,\times\, \mathcal{NIG}_m\left(\boldsymbol{\mu}^*,\, \boldsymbol{V}^*(\lambda),\, a^*,\, b^*\right)\, d\boldsymbol{\alpha}d\sigma^2 \\ &= \mathcal{MVS}_{\boldsymbol{\theta},\,\boldsymbol{\Sigma},\,\nu,\,\tau}(\tilde{\boldsymbol{y}}) \end{aligned} \tag{5.1}$$

where

$$\boldsymbol{\theta} = \tilde{\boldsymbol{B}}\boldsymbol{\mu}^* \tag{5.2}$$

$$\boldsymbol{\Sigma} = \frac{b^*}{a^*}\left(\boldsymbol{I}_r + \tilde{\boldsymbol{B}}\boldsymbol{V}^*\tilde{\boldsymbol{B}}'\right) \tag{5.3}$$

$$\nu = 2a^* \tag{5.4}$$

$$\tau = r \tag{5.5}$$

and

$$\mathcal{MVS}_{\boldsymbol{\theta},\,\boldsymbol{\Sigma},\,\nu,\,\tau}(\tilde{\boldsymbol{y}}) = \frac{\Gamma\left(\frac{\nu+\tau}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\,\pi^{\frac{\tau}{2}}\,|\nu\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left[1 + \frac{(\tilde{\boldsymbol{y}}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{y}}-\boldsymbol{\theta})}{\nu}\right]^{-\frac{\nu+\tau}{2}} \tag{5.6}$$

The density defined in (5.6) is known as *multivariate-t* (see e.g. Kotz and Nadarajah, 2004). Its hyperparameters defined in equations (5.2), (5.3), (5.4) and (5.5) are functions of the hyperparameters of the Normal-Inverse Gamma posterior distribution of $\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{Y},\, M_\lambda$ defined in equations (3.15) through (3.18) and (3.21).

Prediction intervals can be obtained from equation (5.6) by computing the corresponding quantiles. The variance of the distribution in the same equation can be used to estimate the posterior uncertainty.

The technique taken in this thesis considers the simpler case of equation (5.6) with $r = 1$. In this situation we have

$$\begin{aligned} & \tilde{\boldsymbol{B}} &= \tilde{\boldsymbol{b}}' \\ \text{and} \quad & \Sigma &= \frac{b^*}{a^*}\left(1 + \tilde{\boldsymbol{b}}'\boldsymbol{V}^*\tilde{\boldsymbol{b}}\right) \\ \text{where} \quad & \boldsymbol{V}^* &= (\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1} \end{aligned}$$

The change of variables $T = \Sigma^{-\frac{1}{2}}\left(\tilde{Y} - \tilde{\boldsymbol{b}}'\boldsymbol{\mu}^*\right)$ yields $\left|\frac{d\tilde{Y}}{dT}\right| = \Sigma^{\frac{1}{2}}$ and therefore

$$f_T(t) \quad = \quad f_{\tilde{Y}}(\tilde{y})\left|\frac{d\tilde{Y}}{dT}\right| \quad = \quad \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left[1 + \frac{t^2}{\nu}\right]^{-\frac{\nu+1}{2}} \quad \sim \quad t_{2a^*} \qquad (5.7)$$

Thus a Bayes prediction interval of level α for $\tilde{Y}$ is given by

$$\tilde{\boldsymbol{b}}'\boldsymbol{\mu}^* \pm t_{2a^*,\,1-\frac{\alpha}{2}}\,\Sigma^{\frac{1}{2}} \quad = \quad \tilde{\boldsymbol{b}}'\boldsymbol{\mu}^* \pm t_{2a^*,\,1-\frac{\alpha}{2}}\sqrt{\frac{b^*}{a^*}\left(1 + \tilde{\boldsymbol{b}}'\boldsymbol{V}^*\tilde{\boldsymbol{b}}\right)} \qquad (5.8)$$

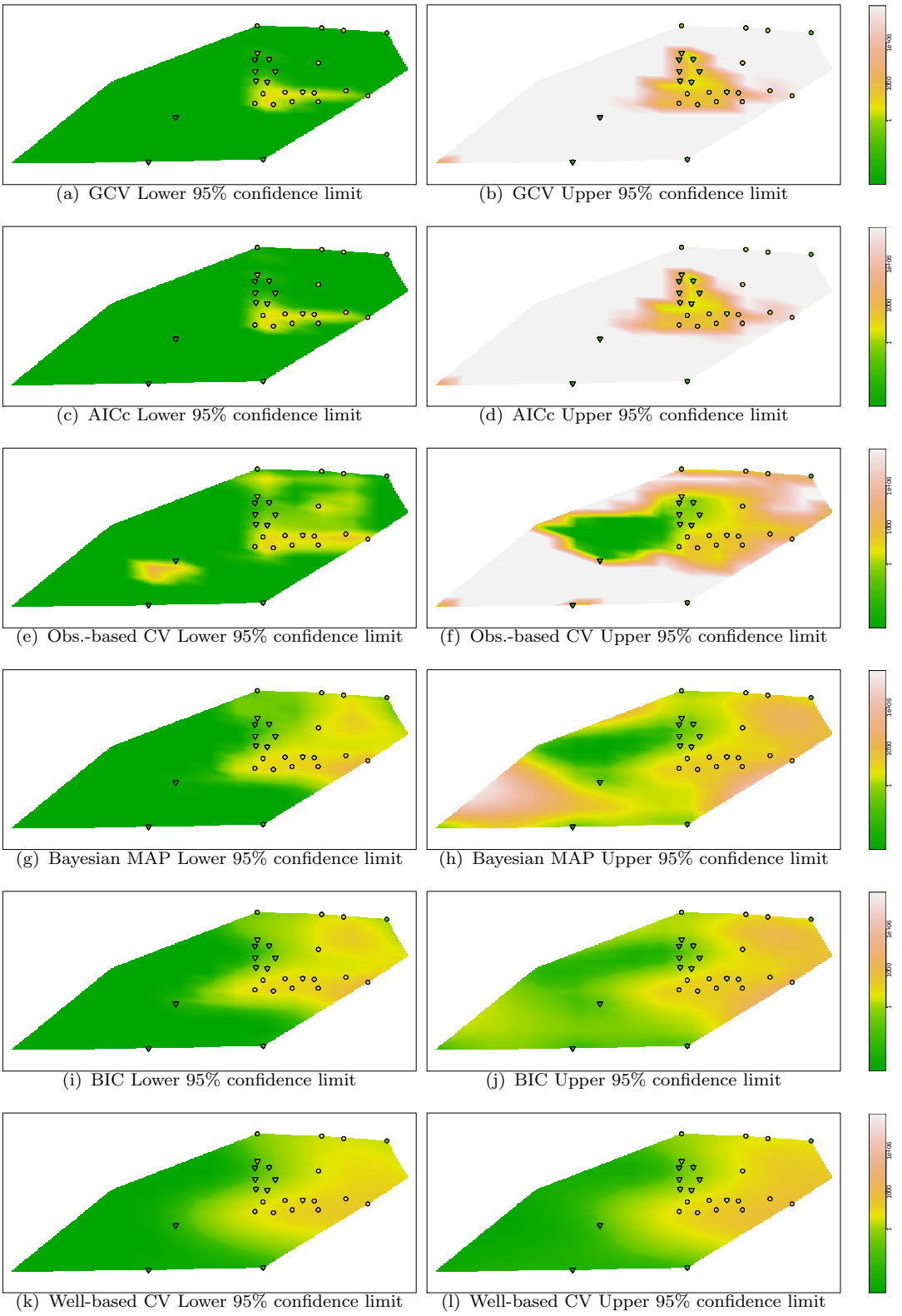
It is interesting to compare the Bayes prediction interval in (5.8) with the prediction interval we would have obtained following the frequentist approach

$$\tilde{\boldsymbol{b}}'\hat{\boldsymbol{\mu}}_{MLE} \quad \pm \quad t_{n-tr(\boldsymbol{H}),\,1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{MLE}\left(1 + \tilde{\boldsymbol{b}}'\hat{\boldsymbol{Var}}(\hat{\boldsymbol{\mu}}_{MLE})\tilde{\boldsymbol{b}}\right)} \qquad (5.9)$$

$$
\begin{array}{lll}
\text{where} & \hat{\boldsymbol{\mu}}_{MLE} & = \boldsymbol{Qy} \\
& \boldsymbol{Q} & = \left(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D}\right)^{-1} \boldsymbol{B}' \\
& \hat{\boldsymbol{Var}}(\hat{\boldsymbol{\mu}}_{MLE}) & = \boldsymbol{QQ}'\hat{\sigma}^2_{MLE} \\
& \hat{\sigma}^2_{MLE} & = \frac{\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y}}{n - tr(\boldsymbol{H})} \\
\text{and} & \boldsymbol{H} & = \boldsymbol{BQ}
\end{array}
$$

Recalling equation (3.17), we have $t_{2a^*, 1-\frac{\alpha}{2}} = t_{2a+n, 1-\frac{\alpha}{2}}$. Given the large number of observations we can replace the t-quantile in equation (5.8) by the corresponding z-quantile $z_{1-\frac{\alpha}{2}}$. The same remark applies to equation (5.9).

In alternative to (5.8), we could have used an empirical Bayesian approach to construct a prediction interval for $\tilde{Y}$, by setting $\hat{\sigma}^2_{MAP}$ to the mode of the posterior Inverse Gamma distribution, i.e. $\hat{\sigma}^2_{MAP} = \frac{b^*}{a^*+1}$ and $\hat{\tau}^2 = \frac{\hat{\sigma}^2_{MAP}}{\lambda_{MAP}}$ in equation (2.22). Recall that a^* and b^* are defined in equations (3.17) and (3.18) respectively.



(a) GCV Lower 95% confidence limit (b) GCV Upper 95% confidence limit

(c) AICc Lower 95% confidence limit (d) AICc Upper 95% confidence limit

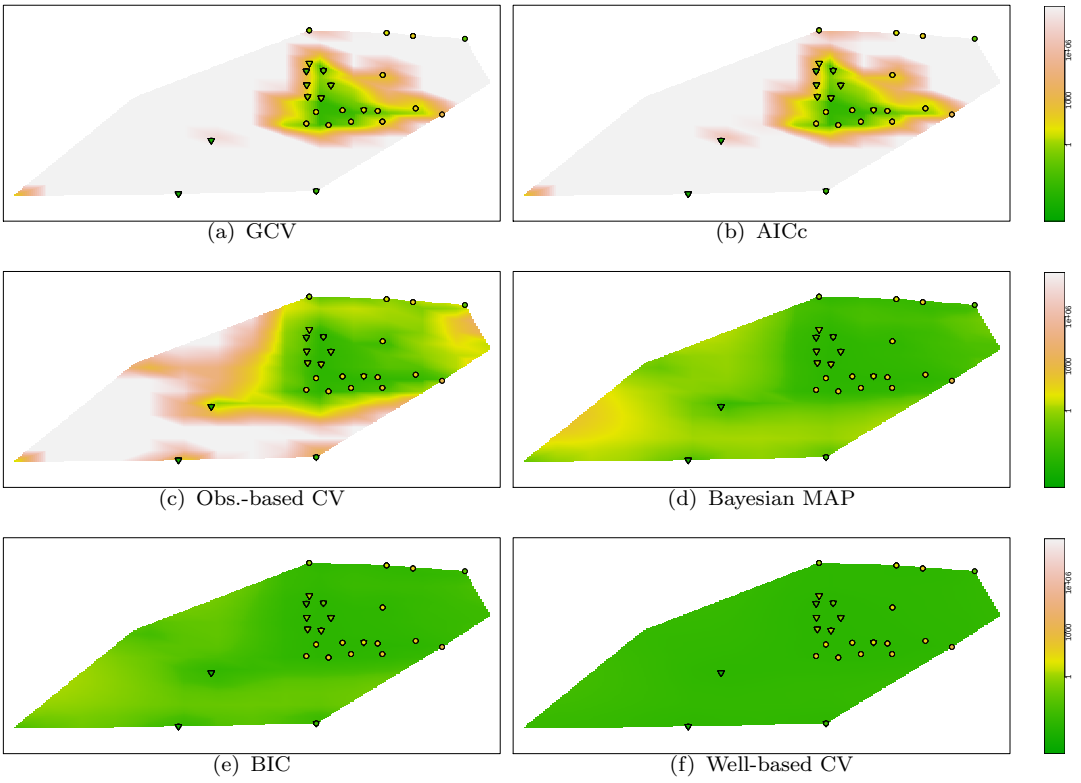
(e) Obs.-based CV Lower 95% confidence limit (f) Obs.-based CV Upper 95% confidence limit

(g) Bayesian MAP Lower 95% confidence limit (h) Bayesian MAP Upper 95% confidence limit

(i) BIC Lower 95% confidence limit (j) BIC Upper 95% confidence limit

(k) Well-based CV Lower 95% confidence limit (l) Well-based CV Upper 95% confidence limit

FIGURE 5.24: Lower and upper 95% confidence limits for the smoothing criteria used in the case study in Section 5.2 under standard assumptions (triangles represent non-detects and circles correspond to observed data)

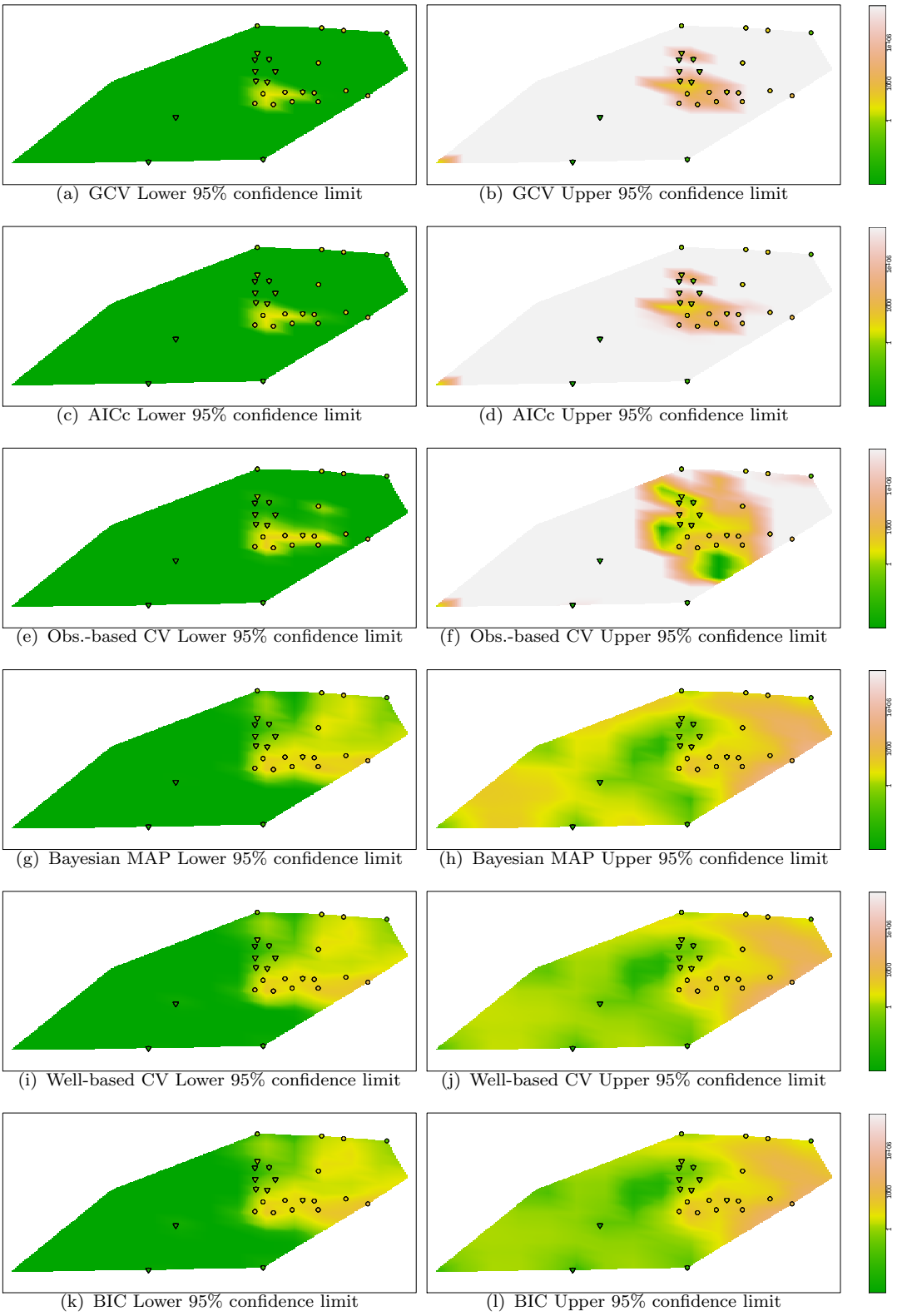


(a) GCV (b) AICc

(c) Obs.-based CV (d) Bayesian MAP

(e) BIC (f) Well-based CV

FIGURE 5.25: Standard errors for the smoothing criteria used in the case study in Section 5.2 under standard assumptions (triangles represent non-detects and circles correspond to observed data)



(a) GCV Lower 95% confidence limit (b) GCV Upper 95% confidence limit

(c) AICc Lower 95% confidence limit (d) AICc Upper 95% confidence limit

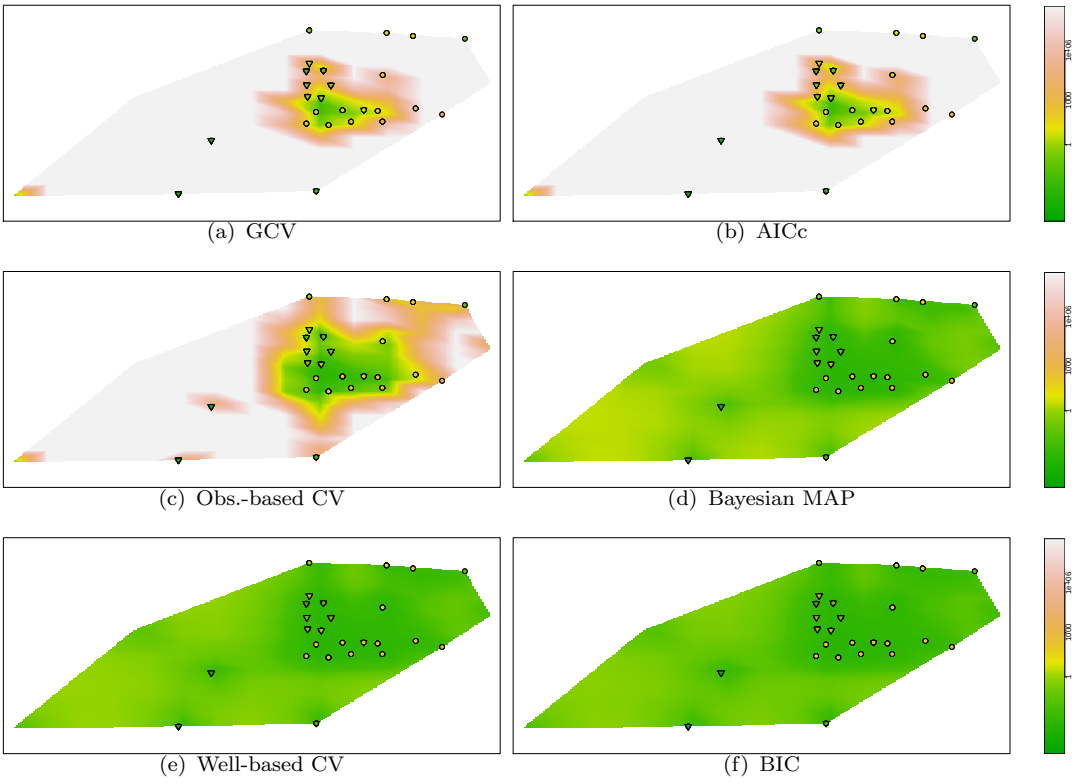
(e) Obs.-based CV Lower 95% confidence limit (f) Obs.-based CV Upper 95% confidence limit

(g) Bayesian MAP Lower 95% confidence limit (h) Bayesian MAP Upper 95% confidence limit

(i) Well-based CV Lower 95% confidence limit (j) Well-based CV Upper 95% confidence limit

(k) BIC Lower 95% confidence limit (l) BIC Upper 95% confidence limit

FIGURE 5.26: Lower and upper 95% confidence limits for the smoothing criteria used in the case study in Section 5.4 under relaxed assumptions (triangles represent non-detects and circles correspond to observed data)



(a) GCV (b) AICc

(c) Obs.-based CV (d) Bayesian MAP

(e) Well-based CV (f) BIC

FIGURE 5.27: Standard errors for the smoothing criteria used in the case study in Section 5.2 under relaxed assumptions (triangles represent non-detects and circles correspond to observed data)

Criterion used to select smoothness	Penalisation parameter λ	Maximum eigenvalue
GCV	7.961e-8	46150237.00
AICc	1.034e-7	34579190.00
Obs.-based CV	1.670e-5	318924.80
Bayesian MAP	1.220e-3	6573.86
BIC	7.394e-3	2291.86
Well-based CV	2.360e-1	425.77

TABLE 5.3: Maximum eigenvalue for the variance-covariance matrix of $\hat{\boldsymbol{\alpha}}$ for the smoothing criteria under standard assumptions in Figure 5.24

Criterion used to select smoothness	Penalisation parameter λ	Maximum eigenvalue
GCV	3.473e-7	515064.40
AICc	4.935e-7	318787.60
Obs.-based CV	2.043e-5	10167.25
Bayesian MAP	4.108e-3	112.34
Well-based CV	9.812e-3	43.28
BIC	1.455e-2	30.85

TABLE 5.4: Maximum eigenvalue for the variance-covariance matrix of $\hat{\boldsymbol{\alpha}}$ for the smoothing criteria under relaxed assumptions in Figure 5.26

Chapter 6

Approximate inference for censored data

6.1 Background

In many environmental applications data are gathered by monitors which cannot record measurements which are below (or above) a certain detection limit. For observations outside the detection range it is only known that they are below a lower detection limit or above an upper detection limit. A naïve approach, still used by many practitioners, is to replace the non-detected observations (usually referred to as *non-detects*) by some deterministic function of the detection limit.

In the case of the application presented in chapter 5 non-detects are censored observations corresponding to low recorded values of contaminants for which it is only known that they lie under a certain threshold. In addition, this threshold is not always the same: it depends on the laboratory which carried out the measurement or may vary over time. Usual practice has been to replace non-detects by one-half the detection limit. But this approach underestimates the uncertainty of the estimated mean parameters and can introduce a substantial bias as suggested in Figure 6.1.

Helsel (2006) has concluded that this kind of approach produces an invasive pattern, as the resulting estimates of correlation coefficients, regression parameters, hypothesis tests and even simple means and standard deviations are inaccurate. The corresponding values may differ substantially from the true values with the deviation being unknown.

This chapter is devoted to proposing an efficient and fast method for incorporating non-detects within the Bayesian framework used to predict the concentration of contaminants over space and time, under the constraints of time and memory described earlier in this work.

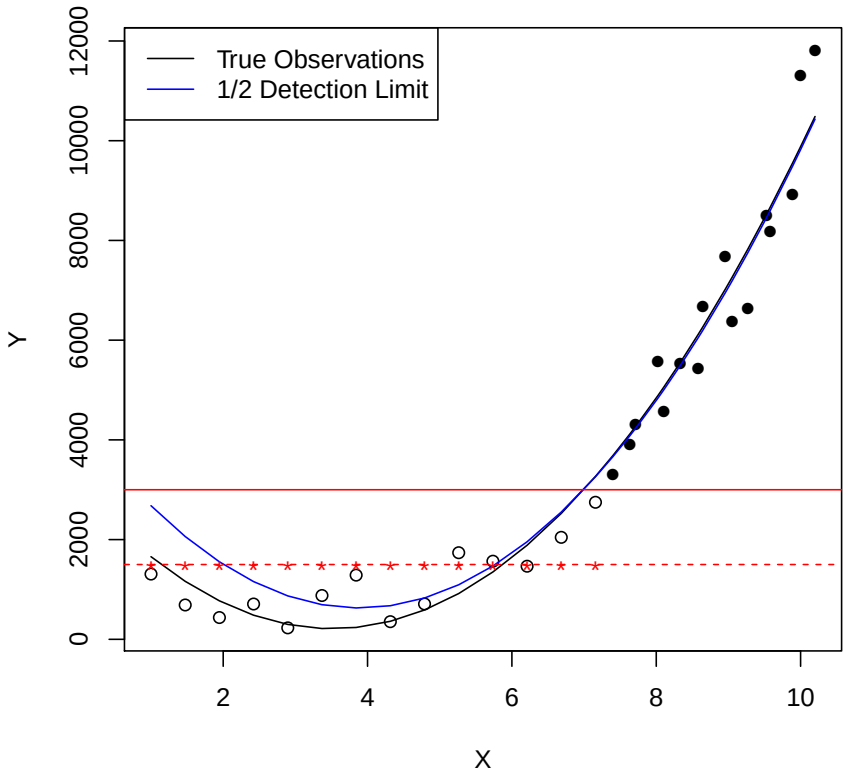


FIGURE 6.1: Fitted functions using the true observed values and by replacing non-detects by 1/2 the detection limit. The horizontal full red line indicates the detection limit while the dashed red line corresponds to 1/2 the detection limit

More formal approaches to this problem include the EM algorithm or methods derived from techniques used for time-to-event data.

6.2 The EM-algorithm

Maximum-Likelihood (ML) estimation is a widely used technique for parameter estimation, in particular within the frequentist framework. In its basic set-up, we assume to be given a random sample coming from a known density $Y_i \sim f(y|\,\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$. Estimation proceeds by finding the value of θ that maximises the joint-density function or *likelihood function* $L(\boldsymbol{Y}|\,\boldsymbol{\theta}) = \prod f(y_i|\,\boldsymbol{\theta})$ for the given random sample $\boldsymbol{Y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ which is assumed to be completely known. If $S = \{\boldsymbol{Y}|f(\boldsymbol{Y}|\,\boldsymbol{\theta}) \neq 0\}$, the support set of the density function $f(\boldsymbol{Y}|\,\boldsymbol{\theta})$, does not depend on the parameter $\boldsymbol{\theta}$ to be estimated and Θ is an open set in $\mathbb{R}^m$, this task is generally accomplished by finding the root of the *log-likelihood function* $\ell(\boldsymbol{Y}|\,\boldsymbol{\theta}) = \log L(\boldsymbol{Y}|\,\boldsymbol{\theta})$, i.e. the value of $\hat{\boldsymbol{\theta}}$ such that $\frac{\partial}{\partial \boldsymbol{\theta}}\, \ell(\boldsymbol{Y}|\,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$. Most often, the root of the log-likelihood cannot be determined analytically and hence iterative methods such as Newton-Raphson are employed to construct a sequence of values converging to $\hat{\boldsymbol{\theta}}$.

Under mild regularity conditions, it can be demonstrated that maximum-likelihood estimates (MLE) have a certain number of appealing properties such as consistency, [1] efficiency and asymptotic normal distribution [2] (see Cramér, 1946).

In our context, there are two challenges which need to be addressed: the first one is that the likelihood does not lead to a closed form estimate of $\hat{\boldsymbol{\alpha}}$, an issue that can be overcome using the EM-algorithm; the second one is that the likelihood contributions that stem from the non-censored observations are highly non-quadratic, giving the likelihood a highly skewed shape leading to a very poor asymptotic approximation by a quadratic function. In the extreme case of a setting with only non-detects, the likelihood has no maximum inside the parameter space and hence the standard asymptotic results do not hold.

[1] Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be a random sample from a distribution $f(\boldsymbol{Y}|\,\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta$ with Θ a convex open set in $\mathbb{R}^m$. Suppose that $\frac{\partial}{\partial \boldsymbol{\theta}}\, f(\boldsymbol{Y}|\,\boldsymbol{\theta})$ exists and that the support set $S = \{\boldsymbol{Y}|f(\boldsymbol{Y}|\,\boldsymbol{\theta}) \neq 0\}$ does not depend on $\boldsymbol{\theta}$ $\forall\, \boldsymbol{\theta} \in \Theta$. If $f(y|\,\boldsymbol{\theta})$ is injective on $\boldsymbol{\theta}$ and the equation $\frac{\partial}{\partial \boldsymbol{\theta}}\, \ell(\boldsymbol{Y}|\,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$ has only one solution $\hat{\boldsymbol{\theta}}$, then $\lim_{n\to\infty} \hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$ almost surely.

[2] Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be a random sample from a distribution $f(y|\,\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta$ with Θ a convex open set in $\mathbb{R}^m$. Suppose that the support set $S = \{\boldsymbol{Y}|f(\boldsymbol{Y}|\,\boldsymbol{\theta}) \neq 0\}$ does not depend on $\boldsymbol{\theta}$ $\forall\, \boldsymbol{\theta} \in \Theta$. Suppose that $\frac{\partial}{\partial \boldsymbol{\theta}} \int_{-\infty}^{\infty} f_{\boldsymbol{\theta}}\,(y|\,\boldsymbol{\theta})\, dy = \int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}}\, f_{\boldsymbol{\theta}}\,(y|\,\boldsymbol{\theta})\, dy$. Suppose that $\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \int_{-\infty}^{\infty} f_{\boldsymbol{\theta}}\,(y|\,\boldsymbol{\theta})\, dy = \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\, f_{\boldsymbol{\theta}}\,(y|\,\boldsymbol{\theta})\, dy$. Suppose that $\mathfrak{I}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\, \ell(\boldsymbol{Y}|\,\boldsymbol{\theta})\right)\right]$ exists and has rank m. Then if $\hat{\boldsymbol{\theta}}$ is an ML consistent estimator for $\boldsymbol{\theta}$, $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to^d \mathcal{N}_n\left(\mathbf{0}, \mathfrak{I}(\boldsymbol{\theta})^{-1}\right)$

Nevertheless, there are cases of incomplete-data structures in which the maximum-likelihood strategy cannot be used in a straightforward manner because the vector $\boldsymbol{Y} = (y_1, \ldots, y_n)$ is not completely known and therefore the equation $\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{Y}|\,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$ cannot be solved in closed form for $\hat{\boldsymbol{\theta}}$ as a function of the observations. Some examples correspond to grouped, censored or truncated data, multivariate data with some missing observations and data from mixtures of distributions.

The Expectation-Maximization (EM) algorithm (see Dempster et al., 1977) is a broadly applicable approach to extend the maximum-likelihood estimation technique to such cases. The underlying idea is to formulate an associated "augmented-data" problem for which it is possible to work out the MLE either analytically or computationally. In essence, it is an iterative algorithm consisting of two steps: the E-step (Expectation) and the M-step (Maximisation) converging to a local maximum under fairly general conditions.

Let us assume that $\boldsymbol{Y} = (\boldsymbol{Y}^u, \boldsymbol{Y}^c)$ where $\boldsymbol{Y}^u = (y_1^u, \ldots, y_{n_u}^u) \in \mathbb{R}^{n_u}$ is the vector associated with the "observed data" and $\boldsymbol{Y}^c = (Y_1^c, \ldots, Y_{n_c}^c) \in \mathbb{R}^{n_c}$ is the vector of random variables for the "incomplete data".

If we knew $\boldsymbol{Y}$ completely we could estimate $\hat{\boldsymbol{\theta}}$ using the appropriate maximum-likelihood technique.

Because we do not know $\boldsymbol{Y}^c$, we do not know either the vector of observations $\boldsymbol{Y} = (\boldsymbol{Y}^u, \boldsymbol{Y}^c)$ completely. In order to overcome this problem, instead of the log-likelihood function, we will consider its expectation

$$E_{\boldsymbol{Y}^c}\left\{\ell\left(\boldsymbol{Y}^u, \boldsymbol{Y}^c|\,\boldsymbol{\theta}\right)\right\} = \int \log f\left(\boldsymbol{Y}^u, \boldsymbol{Y}^c|\,\boldsymbol{\theta}\right)\, f_{\boldsymbol{Y}^c}\left(\boldsymbol{Y}^c|\,\boldsymbol{Y}^u, \boldsymbol{\theta}\right)\, d\boldsymbol{Y}^c \tag{6.1}$$

and find $\hat{\boldsymbol{\theta}}$ which maximises equation (6.1). But this equation gives rise to a new issue, because without knowing $\boldsymbol{\theta}$ we cannot compute the density $f_{\boldsymbol{Y}^c}\left(\boldsymbol{Y}^c|\,\boldsymbol{Y}^u, \boldsymbol{\theta}\right)$. We get round this problem by simply choosing and initial guess $\boldsymbol{\theta}_{old}$ for the vector of parameters and replacing the previous density by $f_{\boldsymbol{Y}^c}\left(\boldsymbol{Y}^c|\,\boldsymbol{Y}^u, \boldsymbol{\theta}_{old}\right)$. Thus, given this initial guess for $\boldsymbol{\theta}$, we then find $\boldsymbol{\theta}_{new}$ maximising

$$\mathbb{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}_{old}\right) = \int \log f\left(\boldsymbol{Y}^u,\, \boldsymbol{Y}^c|\, \boldsymbol{\theta}_{new}\right)\, f_{\boldsymbol{Y}^c}\left(\boldsymbol{Y}^c|\, \boldsymbol{Y}^u,\, \boldsymbol{\theta}_{old}\right)\, d\boldsymbol{Y}^c \tag{6.2}$$

The value of $\hat{\boldsymbol{\theta}}$ found to maximise 6.2 can be used as a new guess and start again.

The full iterative EM-algorithm could be described as follows:

1. Choose an initial value for the vector of parameters $\boldsymbol{\theta} = \boldsymbol{\theta}_{old}$.
2. (E-step) Compute $\mathbb{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}_{old}\right)$.
3. (M-step) Choose $\boldsymbol{\theta}_{new}$ to be the value of $\boldsymbol{\theta}$ which maximises $\mathbb{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}_{old}\right)$.
4. Replace $\boldsymbol{\theta}_{old}$ by $\boldsymbol{\theta}_{new}$ in the E-step and repeat until convergence.

It can be proved that $\ell\left(\boldsymbol{Y}|\, \boldsymbol{\theta}_{new}\right) \geq \ell\left(\boldsymbol{Y}|\, \boldsymbol{\theta}_{old}\right)$ with equality achieved at a maximum (although not necessarily a global maximum) of $\ell\left(\boldsymbol{Y}|\, \boldsymbol{\theta}\right)$. If $\hat{\boldsymbol{\theta}}$ corresponds to the limiting value at convergence of the sequence of parameters $\boldsymbol{\theta}$ constructed with the algorithm described above, clearly the value to be imputed to the vector of missing observations $\boldsymbol{Y}^c$ is $\tilde{\boldsymbol{Y}}^c = \mathbb{E}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{Y}^c)$.

Some of the advantages of the EM-algorithm are (see McLachlan and Krishnan, 2008):

- It is a numerically stable, with each iteration leading to a non-decreasing sequence of the likelihood function.
- It is generally easy to implement.
- It can be used to impute estimated values to "incomplete data".

whereas some of its disadvantages are:

- It may converge slowly.
- It does not provide a procedure to estimate the covariance matrix of the parameters estimates. It is however possible to construct bounds of the variance of the parameters. This can be obtained by computing the second

derivative of the log-likelihood (see Husmeier, 2000) and exploiting the information inequality. This procedure only yields a lower bound rather than an estimate of the variance and is for these reasons not shown in Figures 6.20 and 6.21.

- It does not guarantee convergence to a global maximum when there are multiple maxima.

For the sake of comparison with the Laplace-type method to be proposed, we will use the EM-algorithm in section 6.7 as an alternative to impute values to the censored observations.

6.3 Other existing approaches

Helsel (2005), (see also Helsel, 2012) proposes applying methods used for time-to-event data, based on standard procedures in medical and industrial studies, to handle censored data in the environmental sciences.

Some of these techniques are based on ML estimation, which assumes that the data follow a particular known distribution. The corresponding parameters are fitted matching both the values for observed data and the proportion of those that fall below a known detection limit, which efficiently captures the information contained in non-detects.

Let us assume that $Y \sim f(y, \boldsymbol{\theta})$ with $F(y, \boldsymbol{\theta})$ representing the corresponding cumulative distribution function. In addition, let us consider that the vector of available data is $\boldsymbol{Y} = (y_1, \ldots, y_{n_u}, y_{n_u+1}, \ldots, y_{n_u+n_c})$ where the first n_u figures represent the observed or uncensored values and the last n_c figures correspond to the detection limits of the censored observations. The likelihood function corresponds to the joint probability of the vector $\boldsymbol{Y}$, which under the assumption of independence, takes the form

$$L(\boldsymbol{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n_u} f(y_i, \boldsymbol{\theta}) \prod_{i=n_u+1}^{n_u+n_c} F(y_i, \boldsymbol{\theta})$$

As mentioned in the previous section, point estimation of the vector of parameters $\boldsymbol{\theta}$ is carried out by finding the value $\hat{\boldsymbol{\theta}} \in \Theta$ which maximises $L(\boldsymbol{Y}, \boldsymbol{\theta})$ using suitable numerical methods to accomplish this task, if necessary.

In addition, the usual likelihood-ratio tests and Wald tests can be used to perform hypothesis testing on the parameters. Generally likelihood-ratio tests are preferred to Wald tests which are much more conservative, though differences in p-values are often small.

ML methods generally fail to work properly for data sets with fewer than $30 - 50$ detected values, where one or two outliers may mislead the estimation or there is not enough evidence supporting the assumed probability distribution model.

In such cases nonparametric or "distribution-free" methods which do not assume a shape of data or a specific distribution (and hence do not involve estimating parameters), would be preferred. These methods use the relative position or ranks of data and are specially useful for censored data because they efficiently use the available information. The data's percentiles are reflected in the ranks attached to them. In the case of non-detects, because they are known to be lower than the values above their reporting limit, they are assigned a lower rank.

Nonparametric methods are more powerful than their parametric counterparts when dealing with skewed distributions and outliers; Helsel and Hirsch (2002) have demonstrated their usefulness in the framework of environmental studies.

Most of these nonparametric methods use standard statistical software for "survival analysis" or "reliability analysis" that deal with right-censored data rather than left-censored observations, which are typical in environmental data. Helsel (2012) proposes a "flipping" transformation to adapt the use of these nonparametric methods with environmental data.

6.4 Motivation

As mentioned earlier, the goal of this chapter is to investigate whether the information regarding non-detects can be included in the model in an accurate fashion using the Bayesian framework.

Our hope is that if we can be fully Bayesian on how to deal with non-detect data, we can also extend the Bayesian approach for selecting the penalistation parameter λ.

As we have seen in chapter 2, if we have no censored data at all, the Bayesian approach leads to a closed Normal-Inverse Gamma posterior distribution of the parameters. On the other extreme, if only censored data are available, the model to be applied should be a Bayesian probit model (see Denison et al., 2002). A drawback in this case, is that there is no closed form for the posterior and sampling techniques such as MCMC are required for exact inference. But in addition, σ^2 is not identifiable in the Bayesian probit model.

We propose to approximate the posterior distribution of the parameters by means of a Laplace-type approximation which essentially resembles a Normal-Inverse Gamma density on the parameters. The likelihood function $L(\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma^2)$ will include now the factors corresponding to the censored data and as usual, the prior on the parameters will be a proper Normal-Inverse Gamma.

6.5 The Model

We will consider a Bayesian linear model corresponding to a P-spline regression case. Thus, for the likelihood, $\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma^2 \sim \mathcal{N}_n\left(\boldsymbol{B}\boldsymbol{\alpha}, \sigma^2\boldsymbol{I}_n\right)$ where the $\boldsymbol{B}$ matrix is made up of a set of basis functions.

We will assume a Normal-Inverse Gamma prior distribution for the parameters, $\boldsymbol{\alpha}, \sigma^2|\lambda \sim \mathcal{NIG}_m(\boldsymbol{\mu}, \boldsymbol{V}(\lambda), a, b)$. As suggested earlier, we choose $\boldsymbol{\mu} = \mathbf{0}$ to acknowledge our prior uncertainty about the sign of the regression coefficients; similarly we set $\boldsymbol{V} = (\lambda \boldsymbol{D'D})^{-1}$ (with λ standing for the penalisation parameter and $\boldsymbol{D}$ for a conveniently chosen difference matrix) to mimic the smoothing term in the objective function to be minimised in the P-spline context. Here, a and b are the hyperparameters of the marginal Inverse Gamma distribution for σ^2. These hyperparameters are typically set to a very small value such as 0.001. As mentioned in section 3.4, the rationale behind this choice is that these values yield a limiting approximation to the corresponding uninformative Jeffreys' prior for σ^2. To complete the model we assume an improper uniform prior on λ.

Let us call $\boldsymbol{y^u} \in \mathbb{R}^{n_u}$ the vector of uncensored or continuous observed values and $\boldsymbol{y^c} \in \mathbb{R}^{n_c}$ the one corresponding to non-detects or censored observations, i.e. the observations which could not be directly observed. In agreement with our initial set-up and by splitting $\boldsymbol{B}$ into two submatrices $\boldsymbol{B}^u$ and $\boldsymbol{B}^c$, $\boldsymbol{Y^u}|\boldsymbol{\alpha}, \sigma^2 \sim \mathcal{N}_{n_u}(\boldsymbol{B^u\alpha}, \sigma^2\boldsymbol{I_{n_u}})$ and $\boldsymbol{Y^c}|\boldsymbol{\alpha}, \sigma^2 \sim \mathcal{N}_{n_c}(\boldsymbol{B^c\alpha}, \sigma^2\boldsymbol{I_{n_c}})$ with $\boldsymbol{B^u} \in \mathbb{R}^{n_u \times m}$, $\boldsymbol{B^c} \in \mathbb{R}^{n_c \times m}$, $n = n_u + n_c$ and $\boldsymbol{\alpha} \in \mathbb{R}^m$. In addition, for the censored data, it is only known that $y_i^c \leq d_i, \quad i = 1, 2, \ldots, n_c \quad$ where d_i represents the *i-th* detection limit. In econometrics, this model is known as *Tobit regression model* (see Johnston and diNardo, 1997).

The likelihood takes the form

$$\begin{aligned} L(\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma^2) &\propto \prod_{i=1}^{n_u} \frac{1}{\sigma}\varphi\Big(\frac{y_i^u - \boldsymbol{B}_i^{u\prime}\boldsymbol{\alpha}}{\sigma}\Big) \prod_{i=1}^{n_c} \Phi\Big(\frac{d_i - \boldsymbol{B}_i^{c\prime}\boldsymbol{\alpha}}{\sigma}\Big) \\ &= \Big[\sigma^{-n_u} \exp\Big\{-\frac{1}{2\sigma^2}\|\boldsymbol{y}^u - \boldsymbol{B}^u\boldsymbol{\alpha}\|^2\Big\}\Big] \prod_{i=1}^{n_c} \Phi\Big(\frac{d_i - \boldsymbol{B}_i^{c\prime}\boldsymbol{\alpha}}{\sigma}\Big) \end{aligned} \tag{6.3}$$

As mentioned in section 6.2, the classical frequentist approach to compute the MLE for the parameters from 6.3, would not lead to a closed form solution due to the censored observations. Additionally, in the presence of a large number of censored observations, the shape of the log-likelihood tends to be highly skewed and hence it cannot be well approximated by a quadratic function, a condition

that needs to be met to obtain an asymptotic normal distribution. In the extreme case in which almost all the observations are censored, the log-likelihood is so far from quadratic that it can lead to misleading results. See Hauck and Donner (1977) for a discussion of this phenomenon in the case of logistic regression, which is closely linked to the case of censored observations. In the case in which all the observations are censored, the log-likelihood is monotonic in some parameters, leading to an estimate of $\pm\infty$ for these parameters.

The use of a (Gaussian) prior distribution on the parameters in our Bayesian model, will partially overcome this issue by approximating the true posterior with a quadratic form. This will ensure that the approximate posterior is not monotonic in any of the parameters. However if the proportion of censored observations is very large, the shape of this approximate posterior might still be far from quadratic. The prior is

$$\begin{aligned} f(\boldsymbol{\alpha}, \sigma^2, \lambda) &\propto (\sigma^2)^{-(a+1+\frac{m}{2})} \exp\left\{ -\frac{1}{2\sigma^2}\Big[(\boldsymbol{\alpha}-\boldsymbol{\mu})'\boldsymbol{V}(\lambda)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu}) + 2b\Big]\right\} \\ &= (\sigma^2)^{-(a+1+\frac{m}{2})} \exp\left\{ -\frac{1}{2\sigma^2}\Big[\boldsymbol{\alpha}'\boldsymbol{V}(\lambda)^{-1}\boldsymbol{\alpha} + 2b\Big]\right\} \end{aligned}$$

Up to a normalising constant, the true posterior distribution of the parameters is therefore given by

$$f^*(\boldsymbol{\alpha}, \sigma^2 | \boldsymbol{Y}, \lambda) \propto L(\boldsymbol{Y}|\boldsymbol{\alpha}, \sigma^2) \times f(\boldsymbol{\alpha}, \sigma^2, \lambda) \tag{6.4}$$

6.6 Approximation to the log-likelihood

Let us consider for the moment σ^2 as a nuisance parameter; we will aim at replacing the posterior distribution (6.4) by its Laplace approximation. Let $\hat{\boldsymbol{\alpha}}$ be the point at which the posterior density attains its maximum. Because we are only concerned about $\hat{\boldsymbol{\alpha}}$, for the sake of notation, we will only consider f^* as a

function of the coefficients of our regression model. By expanding the logarithm of f^* around $\hat{\boldsymbol{\alpha}}$ using Taylor's series up to the second order term, we have that

$$\begin{aligned}\ell^*(\boldsymbol{\alpha}) = \log f^*(\boldsymbol{\alpha}) &\approx \ell^*(\hat{\boldsymbol{\alpha}}) + \frac{\partial \ell^*(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}) \\ &\quad + \frac{1}{2}(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})' \frac{\partial^2 \ell^*(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^2}\Big|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}) \\ &= \ell^*(\hat{\boldsymbol{\alpha}}) - \frac{1}{2}(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})' \boldsymbol{Q^{-1}} (\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})\end{aligned}$$

with $\boldsymbol{Q^{-1}} = \boldsymbol{P} = -\frac{\partial^2 \ell^*(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}^2}\Big|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}$. Taking into consideration that the term of first order vanishes at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$, the Laplace approximation to the posterior is

$$f^*(\boldsymbol{\alpha}) \approx f^*(\hat{\boldsymbol{\alpha}}) \exp\Big\{ -\frac{1}{2}(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})' \boldsymbol{Q^{-1}} (\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})\Big\}$$

where $\boldsymbol{Q}$ and $\boldsymbol{P}$ can be regarded as the covariance and the precision matrices respectively.

The steps for the computation of $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{Q^{-1}}$ can be described more easily by initially considering separately the prior, the uncensored and the censored observations. We can write $\ell^* = \ell_{n_u} + \ell_{n_c} + \ell_f$ where

$$\ell_f = k - (a+1+\frac{m}{2})\log(\sigma^2) - \frac{1}{2\sigma^2}\Big[\boldsymbol{\alpha}'\boldsymbol{V}^{-1}\boldsymbol{\alpha} + 2b\Big]$$

$$\begin{aligned}\ell_{n_u} &= \tilde{k} - n_u \log(\sigma) - \frac{1}{2\sigma^2}\|\boldsymbol{y}^u - \boldsymbol{B}^u\boldsymbol{\alpha}\|^2 \\ &= \tilde{k} - n_u \log(\sigma) - \frac{1}{2\sigma^2}\big[\boldsymbol{\alpha'}\boldsymbol{B}^{u\,\prime}\boldsymbol{B}^u\boldsymbol{\alpha} - 2\boldsymbol{\alpha}'\boldsymbol{B}^{u\,\prime}\boldsymbol{y}^u + \boldsymbol{y}^{u\,\prime}\boldsymbol{y}^u\big]\end{aligned}$$

and

$$\ell_{n_c} = \sum_{i=1}^{n_c} \log \Phi\Big(\underbrace{\frac{d_i - \boldsymbol{B}_i^{c\,'}\boldsymbol{\alpha}}{\sigma}}_{=t_i} \Big) = \sum_{i=1}^{n_c} \log \Phi(t_i) \tag{6.5}$$

Recalling that $\frac{\partial}{\partial x}(u'x) = \frac{\partial}{\partial x}(x'u) = u'$, $\frac{\partial}{\partial x}(x'Ax) = x'(A + A')$ and $\frac{\partial^2}{\partial x^2}(x'Ax) = A + A'$, we have

$$\frac{\partial \ell_f}{\partial \boldsymbol{\alpha}} = -\frac{1}{\sigma^2}\boldsymbol{\alpha}'\boldsymbol{V}^{-1}$$

$$\frac{\partial^2 \ell_f}{\partial \boldsymbol{\alpha}^2} = -\frac{1}{\sigma^2}\boldsymbol{V}^{-1}$$

$$\frac{\partial \ell_{n_u}}{\partial \boldsymbol{\alpha}} = \frac{\boldsymbol{y}^{u\,'}\boldsymbol{B}^u - \boldsymbol{\alpha}'\boldsymbol{B}^{u\,'}\boldsymbol{B}^u}{\sigma^2}$$

$$\frac{\partial^2 \ell_{n_u}}{\partial \boldsymbol{\alpha}^2} = -\frac{\boldsymbol{B}^{u\,'}\boldsymbol{B}^u}{\sigma^2}$$

and

$$\frac{\partial \ell_{n_c}}{\partial \alpha_j} = \sum_{i=1}^{n_c} \left(-\frac{\boldsymbol{B}_{ij}^c}{\sigma}\right) \frac{\varphi}{\Phi}\Big|_{t=t_i} = -\frac{1}{\sigma}\sum_{i=1}^{n_c} \frac{\varphi}{\Phi}\Big|_{t=t_i} \boldsymbol{B}_{ij}^c$$

$$\frac{\partial^2 \ell_{n_c}}{\partial \alpha_j \partial \alpha_k} = -\frac{1}{\sigma}\sum_{i=1}^{n_c} \left(-\frac{\boldsymbol{B}_{ik}^c}{\sigma}\right) \frac{\varphi'\Phi - \varphi^2}{\Phi^2}\Big|_{t=t_i} \boldsymbol{B}_{ij}^c = -\frac{1}{\sigma^2}\sum_{i=1}^{n_c} \boldsymbol{B}_{ki}^{c\,'} \frac{\varphi^2 - \varphi'\Phi}{\Phi^2}\Big|_{t=t_i} \boldsymbol{B}_{ij}^c$$

The last two relationships can be written in a more compact fashion as

$$\frac{\partial \ell_{n_c}}{\partial \boldsymbol{\alpha}} = -\frac{1}{\sigma}\boldsymbol{v}^{c\,'}\boldsymbol{B}^c \qquad \text{with} \qquad v_i^c = \frac{\varphi}{\Phi}\Big|_{t=t_i} \tag{6.6}$$

$$\frac{\partial^2 \ell_{n_c}}{\partial \boldsymbol{\alpha}^2} = -\frac{1}{\sigma^2}\boldsymbol{B}^{c\prime}\boldsymbol{W}^c\boldsymbol{B}^c \qquad \text{with} \qquad W_{ij}^c = \frac{\varphi^2 - \varphi'\Phi}{\Phi^2}\Big|_{t=t_i}\delta_{ij} \tag{6.7}$$

where δ_{ij} represents the usual *Kronecker's delta* function. We can now derive the full expressions of the first and second partial derivatives of the log-posterior, yielding

$$\begin{aligned}
\Big(\frac{\partial \ell^*}{\partial \boldsymbol{\alpha}}\Big)' &= \Big(\frac{\partial \ell_{n_u}}{\partial \boldsymbol{\alpha}} + \frac{\partial \ell_{n_c}}{\partial \boldsymbol{\alpha}} + \frac{\partial \ell_f}{\partial \boldsymbol{\alpha}}\Big)' \\
&= \left(\frac{\boldsymbol{y}^{u\prime}\boldsymbol{B}^u - \boldsymbol{\alpha}'\boldsymbol{B}^{u\prime}\boldsymbol{B}^u - \sigma\boldsymbol{v}^{c\prime}\boldsymbol{B}^c - \boldsymbol{\alpha}'\boldsymbol{V}^{-1}}{\sigma^2}\right)' \\
&= \frac{\boldsymbol{B}^{u\prime}\boldsymbol{y}^u - \boldsymbol{B}^{u\prime}\boldsymbol{B}^u\boldsymbol{\alpha} - \sigma\boldsymbol{B}^{c\prime}\boldsymbol{v}^c - \boldsymbol{V}^{-1}\boldsymbol{\alpha}}{\sigma^2}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 \ell^*}{\partial \boldsymbol{\alpha}^2} &= \frac{\partial^2 \ell_{n_u}}{\partial \boldsymbol{\alpha}^2} + \frac{\partial^2 \ell_{n_c}}{\partial \boldsymbol{\alpha}^2} + \frac{\partial^2 \ell_f}{\partial \boldsymbol{\alpha}^2} \\
&= -\frac{\boldsymbol{B}^{u\prime}\boldsymbol{B}^u + \boldsymbol{B}^{c\prime}\boldsymbol{W}^c\boldsymbol{B}^c + \boldsymbol{V}^{-1}}{\sigma^2} \qquad = \quad -\boldsymbol{Q}^{-1}
\end{aligned}$$

The optimal value of the parameters $\hat{\boldsymbol{\alpha}}$ maximising the likelihood $L(\boldsymbol{\alpha})$ is determined by solving the equation $\dfrac{\partial \ell^*}{\partial \boldsymbol{\alpha}} = 0$. In practice, this task is accomplished numerically using the iterative Newton-Raphson algorithm

$$\boldsymbol{\alpha}_{k+1} = \boldsymbol{\alpha}_k - \left[\Big(\frac{\partial^2 \ell^*}{\partial \boldsymbol{\alpha}^2}\Big)^{-1}\Big(\frac{\partial \ell^*}{\partial \boldsymbol{\alpha}}\Big)'\right]_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_k}$$

The iterative algorithm for computing $\boldsymbol{\alpha}$ implies evaluating the inverse of the $m \times m$ matrix $\dfrac{\partial^2 \ell^*}{\partial \boldsymbol{\alpha}^2}$ at each step. Taking into account that

$$\frac{\partial^2 \ell^*}{\partial \boldsymbol{\alpha}^2} \propto \underbrace{\boldsymbol{B}^{u\prime}\boldsymbol{B}^u + \boldsymbol{V}^{-1}}_{\boldsymbol{A}} + \underbrace{\boldsymbol{B}^{c\prime}\boldsymbol{W}^c\boldsymbol{B}^c}_{\text{of rank } n_c} = \boldsymbol{A} + \boldsymbol{B}^{c\prime}\boldsymbol{W}^c\boldsymbol{B}^c$$

efficiency can be achieved if $n_c << m$ using the Sherman-Morrison-Woodbury formula (see e.g. Gentle, 2007), according to which

$$\begin{aligned}\left(\frac{\partial^2 \ell^*}{\partial \boldsymbol{\alpha}^2}\right)^{-1} &= \left(\boldsymbol{A} + \boldsymbol{B}^{c\prime}\boldsymbol{W}^c\boldsymbol{B}^c\right)^{-1} \\ &= \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}^{c\prime}\left[\boldsymbol{B}^c\boldsymbol{A}^{-1}\boldsymbol{B}^{c\prime} + (\boldsymbol{W}^c)^{-1}\right]^{-1}\boldsymbol{B}^c\boldsymbol{A}^{-1}\end{aligned}$$

In these terms, the computational complexity can be decreased as it would imply only the evaluation of the inverses of the diagonal matrix $\boldsymbol{W}^c$ and that of the $n_c \times n_c$ matrix $\boldsymbol{B}^c\boldsymbol{A}^{-1}\boldsymbol{B}^{c\prime} + (\boldsymbol{W}^c)^{-1}$.

The quadratic approximation to the posterior distribution of the parameter $\boldsymbol{\alpha}$ is presented in the illustrative example in Figure 6.2.

As for the nuisance parameter $S = \sigma^2$, if it is not known (as it is generally the case), we can estimate it by applying again the Newton-Raphson algorithm until convergence, at each iteration in the computation of $\hat{\boldsymbol{\alpha}}$. By calling

$$\begin{aligned}G &= \frac{1}{2}\left[(\boldsymbol{\alpha} - \boldsymbol{\mu})'\boldsymbol{V}^{-1}(\boldsymbol{\alpha} - \boldsymbol{\mu}) + 2b + \|\boldsymbol{y}^u - \boldsymbol{B}^u\boldsymbol{\alpha}\|^2\right] \\ H &= -\left(a + 1 + \frac{m + n_u}{2}\right)\end{aligned}$$

and remembering that

$$t_i = \frac{d_i - \boldsymbol{B}_i^{c\prime}\boldsymbol{\alpha}}{\sigma} = (d_i - \boldsymbol{B}_i^{c\prime}\boldsymbol{\alpha})S^{-\frac{1}{2}}$$

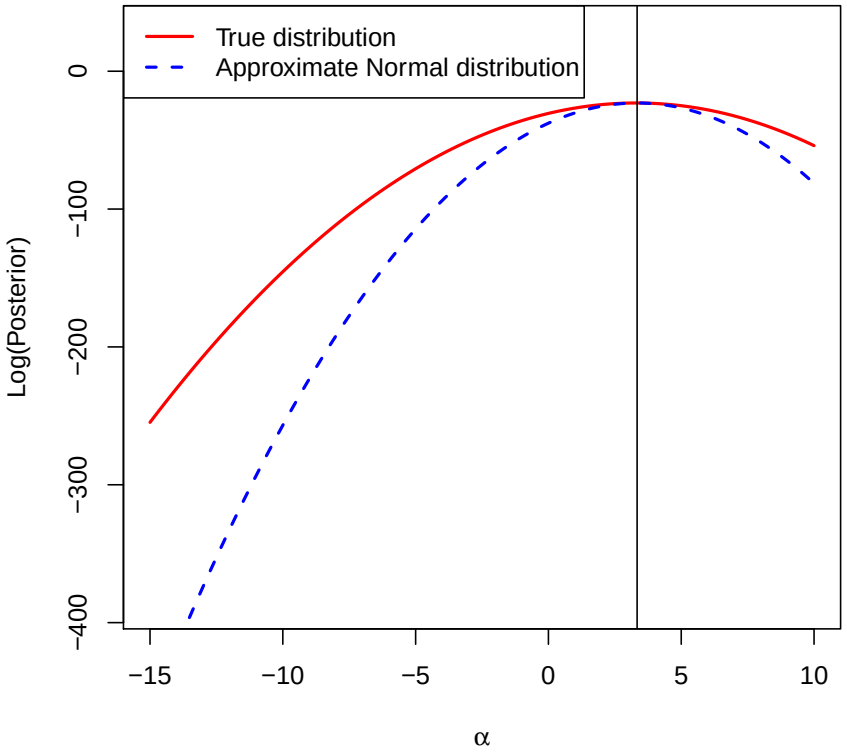


FIGURE 6.2: Illustrative example corresponding to the Laplace approximation for the posterior distribution of $\boldsymbol{\alpha}$

we obtain

$$\begin{aligned}
\frac{\partial \ell^*}{\partial S} &= H\,S^{-1} + G\,S^{-2} - \frac{1}{2}S^{-1}\sum_{i=1}^{n_c} t_i \frac{\varphi(t_i)}{\Phi(t_i)} \\
\frac{\partial^2 \ell^*}{\partial S^2} &= -H\,S^{-2} - 2G\,S^{-3} + \frac{1}{4}S^{-2}\sum_{i=1}^{n_c} t_i \frac{\varphi(t_i)}{\Phi(t_i)}\left(3 - t_i^2 - t_i\frac{\varphi(t_i)}{\Phi(t_i)}\right) \qquad (6.8)
\end{aligned}$$

yielding the recurrence formula

$$\sigma^2_{h+1} = S_{h+1} = \left\{ S - \left(\frac{\partial^2 \ell^*}{\partial S^2} \right)^{-1} \left(\frac{\partial \ell^*}{\partial S} \right) \right\}_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{k+1},\, S=S_h}$$

$$= \left\{ \frac{12G + S\left[8H - \sum_{i=1}^{n_c} t_i \frac{\varphi(t_i)}{\Phi(t_i)} \left(5 - t_i^2 - t_i \frac{\varphi(t_i)}{\Phi(t_i)}\right)\right]}{8G + S\left[4H - \sum_{i=1}^{n_c} t_i \frac{\varphi(t_i)}{\Phi(t_i)} \left(3 - t_i^2 - t_i \frac{\varphi(t_i)}{\Phi(t_i)}\right)\right]} S \right\}_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{k+1},\, S=S_h}$$

6.7 Interpretation in terms of the imputed values

Without loss of generality, we can think of the $\boldsymbol{B}$ matrix as having its rows ordered in such a way that the first n_u of them correspond to the actual observed values while the last n_c are those of the censored observations, i.e. $\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}^u \\ \boldsymbol{B}^c \end{pmatrix}$.

We can write

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{I}_{n_u} & 0 \\ 0 & \boldsymbol{W}^c \end{pmatrix} \qquad \boldsymbol{r} = \begin{pmatrix} \boldsymbol{y}^u - \boldsymbol{B}^u \boldsymbol{\alpha} \\ -\sigma \boldsymbol{v}^c \end{pmatrix} \tag{6.9}$$

With this notation

$$\left(\frac{\partial \ell^*}{\partial \boldsymbol{\alpha}}\right)' = \frac{\boldsymbol{B}'\boldsymbol{r} - \boldsymbol{V}^{-1}\boldsymbol{\alpha}}{\sigma^2}$$

$$\frac{\partial^2 \ell^*}{\partial \boldsymbol{\alpha}^2} = -\frac{\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B} + \boldsymbol{V}^{-1}}{\sigma^2} = -\boldsymbol{Q}^{-1} \tag{6.10}$$

and thus, the Newton-Raphson algorithm can be rewritten as

$$\begin{aligned}
\boldsymbol{\alpha}_{k+1} &= \boldsymbol{\alpha}_k - \left[\left(\frac{\partial^2 \ell^*}{\partial \boldsymbol{\alpha}^2}\right)^{-1}\left(\frac{\partial \ell^*}{\partial \boldsymbol{\alpha}}\right)'\right]_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_k} \\
&= \boldsymbol{\alpha}_k - \left[\Big(-\frac{\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B}+\boldsymbol{V}^{-1}}{\sigma^2}\Big)^{-1}\Big(\frac{\boldsymbol{B}'\boldsymbol{r}-\boldsymbol{V}^{-1}\boldsymbol{\alpha}}{\sigma^2}\Big)\right]_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_k} \\
&= \boldsymbol{\alpha}_k + \Big[(\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B}+\boldsymbol{V}^{-1})^{-1}(\boldsymbol{B}'\boldsymbol{r}-\boldsymbol{V}^{-1}\boldsymbol{\alpha})\Big]_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_k} \qquad (6.11) \\
&= \left\{(\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B}+\boldsymbol{V}^{-1})^{-1}\Big[(\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B}+\boldsymbol{V}^{-1})\boldsymbol{\alpha}+\boldsymbol{B}'\boldsymbol{r}-\boldsymbol{V}^{-1}\boldsymbol{\alpha}\Big]\right\}_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_k} \\
&= \Big[(\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B}+\boldsymbol{V}^{-1})^{-1}\boldsymbol{B}'\boldsymbol{W}(\boldsymbol{B}\boldsymbol{\alpha}+\boldsymbol{W}^{-1}\boldsymbol{r})\Big]_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_k}
\end{aligned}$$

Taking limits on both sides, the previous expression yields

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B}+\boldsymbol{V}^{-1})^{-1}\boldsymbol{B}'\boldsymbol{W}(\boldsymbol{B}\hat{\boldsymbol{\alpha}}+\boldsymbol{W}^{-1}\boldsymbol{r})$$

where $\hat{\boldsymbol{\alpha}} = \lim_{k\to\infty}\boldsymbol{\alpha_k}$. Therefore

$$\boldsymbol{B}\hat{\boldsymbol{\alpha}}+\boldsymbol{W}^{-1}\boldsymbol{r} = \begin{pmatrix} \boldsymbol{y}^u \\ \boldsymbol{B}^c\hat{\boldsymbol{\alpha}} - \sigma(\boldsymbol{W}^c)^{-1}\boldsymbol{v}^c \end{pmatrix} \qquad (6.12)$$

Thus, the iterative method can now be viewed as a weighted regression model where the response is made up of the true observed responses and imputed values.

We can also demonstrate that these fake observations are always smaller than the corresponding fitted values, i.e. $\left(\boldsymbol{B}\hat{\boldsymbol{\alpha}}+\boldsymbol{W}^{-1}\boldsymbol{r}\right)_i \leq \left(\boldsymbol{B}\hat{\boldsymbol{\alpha}}\right)_i$ or equivalently $\left(\boldsymbol{W}^{-1}\boldsymbol{r}\right)_i \leq 0$ for $i = n_u+1,\ldots,n_u+n_c$.

For the imputed values, it is $\boldsymbol{W}^{-1}\boldsymbol{r} = -\sigma(\boldsymbol{W}^c)^{-1}\boldsymbol{v}^c$; because by definition it is $\boldsymbol{v}_i^c \geq 0$, it suffices to show that $\boldsymbol{W}_i^c \geq 0$.

Recalling the definition of $\boldsymbol{W}_i^c$ and that

$$\varphi(x)' = -x\varphi(x)\,\forall x, \tag{6.13}$$

our claim holds if $0 < \varphi(x)^2 - \varphi(x)'\Phi(x) = \varphi(x)\big(\varphi(x) + x\Phi(x)\big)$ or simply if $\varphi(x) + x\Phi(x) \geq 0\,\forall x$.

Let us consider $\mathbb{E}(X|X \leq x) = \displaystyle\int_{-\infty}^{+\infty} s\, f_{X|X\leq x}(s|x)\, ds$. Then

$$\begin{aligned}
F_{X|X\leq x}(s|x) &= P(X \leq s|X \leq x) \\
&= 1 - P(X \geq s|X \leq x) \\
&= 1 - \frac{P(s \leq X \leq x)}{P(X \leq x)} \\
&= 1 - \frac{F_X(x) - F_X(s)}{F_X(x)}\, I_{(-\infty,x)}(s)
\end{aligned}$$

yielding

$$f_{X|X\leq x}(s|x) = \frac{f_X(s)}{F_X(x)}\, I_{(-\infty,x)}(s)$$

and thus

$$\mathbb{E}(X|X \leq x) = \int_{-\infty}^{+\infty} s\, \frac{f_X(s)}{F_X(x)}\, I_{(-\infty,x)}(s)\, ds = \int_{-\infty}^{x} s\, \frac{f_X(s)}{F_X(x)}\, ds$$

$$\leq \int_{-\infty}^{x} x \frac{f_X(s)}{F_X(x)}\, ds \qquad = x \int_{-\infty}^{+\infty} f_{X|X\leq x}(s|x)\, ds$$

$$= x$$

In particular, for the Normal distribution, we have that

$$\mathbb{E}(X|X \leq x) = \int_{-\infty}^{x} s \frac{\varphi(s)}{\Phi(x)}\, ds = \frac{1}{\Phi(x)} \int_{-\infty}^{x} -\varphi(s)'\, ds$$

$$= \frac{1}{\Phi(x)} \int_{x}^{-\infty} d\varphi(s) = -\frac{\varphi(x)}{\Phi(x)}$$

Putting all together, it is $-\frac{\varphi(x)}{\Phi(x)} = \mathbb{E}(X|X \leq x) \leq x$ leading to $0 \leq \varphi(x) + x\Phi(x)$.

The same result can be achieved by noticing that for the function $H(x) = \varphi(x) + x\Phi(x)$ it is $H'(x) = \Phi(x) > 0\, \forall x$ recalling equation (6.13). Therefore $H(x)$ is a strictly increasing function.

Because $\mathbb{E}(X)$ exists, it holds that $\lim_{x\to-\infty} x\Phi(x) = 0$ (see Rényi, 2007); thus $\lim_{x\to-\infty} H(x) = 0$ and therefore $0 = \lim_{x\to-\infty} \varphi(x) + x\Phi(x) < \varphi(x) + x\Phi(x)\, \forall x$ due to the monotonicity of $H(x)$.

Figure 6.3 represents the weight function defined as $w(t) = \frac{\varphi(t)^2 - \varphi(t)'\Phi(t)}{\Phi(t)^2}$. At convergence, these weights are computed at $t_i = \frac{d_i - \boldsymbol{B}_i^{c'}\hat{\boldsymbol{\alpha}}}{\sigma}$, i.e., as a function of the distance between the detection limit and the fitted value corresponding to the *i-th* censored value. Thus we notice that the longer the distance the smaller is the weight given to the corresponding fake observation.

The *ad-hoc* example in Figure 6.4 shows a very good agreement between the Laplace-type approximation and the fitted function that results from using all the data. Instead, replacing non-detects by the detection-limit introduces a remarkable bias. The red points correspond to the imputed values computed according

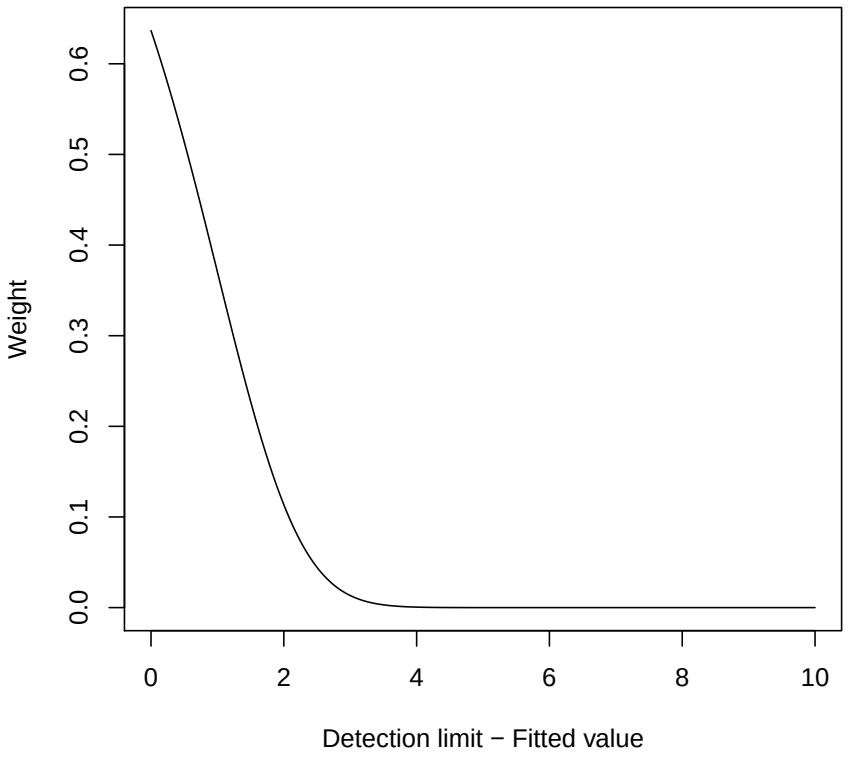


FIGURE 6.3: Weight function for the imputed observations

to equation (6.12). The higher the weight assigned to these points, the darker they are depicted. Appendix C provides the details of this example.

Alternatively, we can use the EM-algorithm to impute values for the non-detect data. According to section 6.2, we need to compute $\mathbb{E}(Y|Y<d)$ under the assumption that it is $Y|\mu,\sigma^2 \sim \mathcal{N}(\mu,\sigma^2)$. We have that

$$\begin{aligned} F_{Y|Y<d}(y) &= P(Y<y|Y<d) \\ &= 1-P(Y\geq y|Y<d) \\ &= 1-\frac{P(y\leq Y<d)}{P(Y<d)} \end{aligned}$$

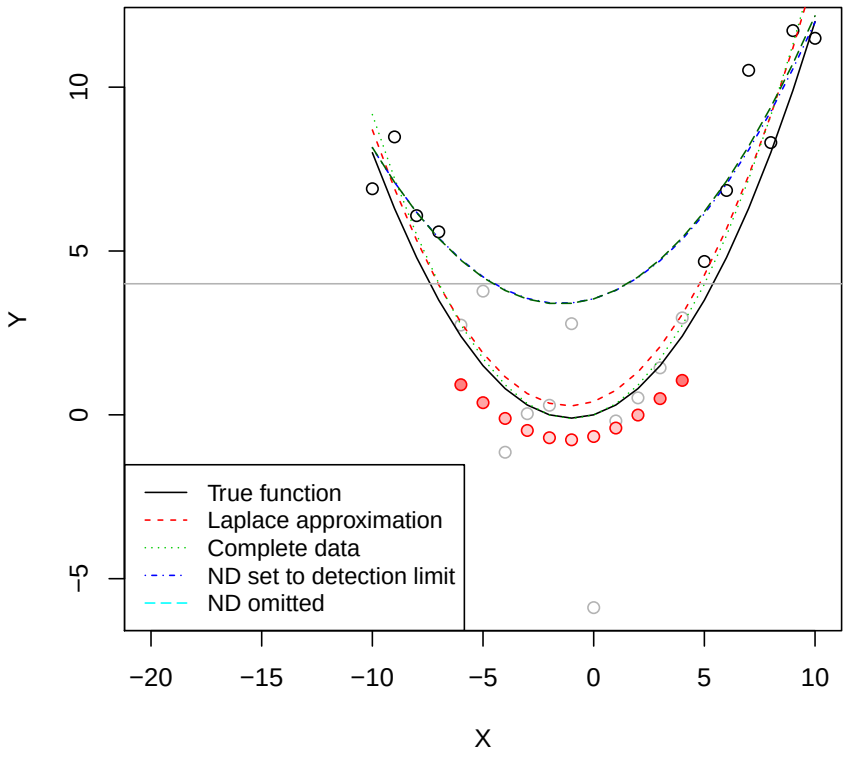


FIGURE 6.4: Comparison between the Laplace-type approximation and the standard approach. The red points correspond to the imputed values. The higher the intensity, the higher the weight assigned to these points

If $y > d$, $F_{Y|Y<d}(y) = 1$ and therefore $f_{Y|Y<d}(y) = \frac{d}{dy} F_{Y|Y<d}(y) = 0$. Otherwise it holds that

$$
\begin{aligned}
F_{Y|Y<d}(y) &= 1 - \frac{P(\frac{y-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{d-\mu}{\sigma})}{P(\frac{Y-\mu}{\sigma} < \frac{d-\mu}{\sigma})} \\
&= 1 - \frac{\Phi(\frac{d-\mu}{\sigma}) - \Phi(\frac{y-\mu}{\sigma})}{\Phi(\frac{d-\mu}{\sigma})} \\
&= \frac{\Phi(\frac{y-\mu}{\sigma})}{\Phi(\frac{d-\mu}{\sigma})}
\end{aligned}
$$

Hence for $y < d$, $f_{Y|Y<d}(y) = \frac{d}{dy} F_{Y|Y<d}(y) = \frac{1}{\sigma} \frac{\varphi(\frac{y-\mu}{\sigma})}{\Phi(\frac{d-\mu}{\sigma})}$. By denoting $\beta = \frac{d-\mu}{\sigma}$ and using the indicator function, the conditional density $f_{Y|Y<d}(y)$ can be written in a more compact form as

$$f_{Y|Y<d}(y) = \frac{1}{\sigma} \frac{\varphi(\frac{y-\mu}{\sigma})}{\Phi(\beta)} I(y < d)$$

Now we can compute the conditional expectation $\mathbb{E}(Y|Y < d)$ as

$$\begin{aligned}
\mathbb{E}(Y|Y < d) &= \int_{-\infty}^{+\infty} y\, f_{Y|Y<d}(y)\, dy \\
&= \int_{-\infty}^{+\infty} y\, \frac{1}{\sigma} \frac{\varphi(\frac{y-\mu}{\sigma})}{\Phi(\beta)} I(y < d)\, dy \\
&= \frac{1}{\sigma\, \Phi(\beta)} \int_{-\infty}^{d} y\, \varphi(\frac{y-\mu}{\sigma})\, dy
\end{aligned}$$

With the change of variables $y = \sigma\, t + \mu$ we obtain $dy = \sigma\, dt$. Considering also that for the normal density $\varphi(t)$ it is $d\varphi(t) = -t\, \varphi(t)\, dt$, we have that

$$\begin{aligned}
\mathbb{E}(Y|Y < d) &= \frac{1}{\sigma\, \Phi(\beta)} \int_{-\infty}^{d} y\, \varphi(\frac{y-\mu}{\sigma})\, dy \\
&= \frac{1}{\Phi(\beta)} \int_{-\infty}^{\beta=\frac{d-\mu}{\sigma}} (\sigma\, t + \mu)\, \varphi(t)\, dt \\
&= \frac{1}{\Phi(\beta)} \Big[\sigma \int_{-\infty}^{\beta} t\, \varphi(t)\, dt + \mu \int_{-\infty}^{\beta} \varphi(t)\, dt\Big] \\
&= \frac{1}{\Phi(\beta)} \Big[- \sigma \int_{-\infty}^{\beta} d\varphi(t) + \mu\, \Phi(\beta)\Big]
\end{aligned}$$

$$= \quad -\sigma \frac{\varphi(\beta)}{\Phi(\beta)} + \mu$$

$$= \quad \mu - \sigma \frac{\varphi(\beta)}{\Phi(\beta)}$$

Therefore, recalling equations (2.15) and (2.16), it is

$$\tilde{y}_i^{EM} = \begin{cases} y_i^u & i = 1, \ldots, n_u \\ y_i^c = \mathbb{E}(Y_i^c | \boldsymbol{\alpha}, \sigma, y_i^c < d_i) = \boldsymbol{B}_i^{c'} \boldsymbol{\alpha}_{old} - \sigma_{old} \frac{\varphi(\beta_i)}{\Phi(\beta_i)} & \text{with } \beta_i = \dfrac{d_i - \boldsymbol{B}_i^{c'} \boldsymbol{\alpha}_{old}}{\sigma_{old}} \\ & i = 1, \ldots, n_c \end{cases}$$

where

$$\boldsymbol{B}\boldsymbol{\alpha}_{old} = \boldsymbol{B} \left(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D} \right)^{-1} \boldsymbol{B}' \tilde{\boldsymbol{y}}_{old}^{EM} = \boldsymbol{H}\, \tilde{\boldsymbol{y}}_{old}^{EM} \qquad \text{and}$$

$$\sigma_{old}^2 = \frac{\| \tilde{\boldsymbol{y}}_{old}^{EM} - \boldsymbol{B}\boldsymbol{\alpha}_{old} \|^2}{n - tr(\boldsymbol{H})}$$

yielding an expression similar to equation (6.12) with $\boldsymbol{W}_{EM}^c = \boldsymbol{I}_{n_c}$. Thus, the EM-algorithm gives the same weight to all the imputed values. As mentioned earlier this method needs many more iterations than the proposed Laplace-type approximation and constructing confidence bands is not straightforward.

6.8 Approximation to the posterior distribution of σ^2

When dealing only with fully observed data, without non-detects, an Inverse Gamma distribution of parameters a^* and b^* is typically used to describe the

distribution of σ^2. We will try to estimate these parameters by approximating the actual posterior distribution of $\boldsymbol{\alpha}$ and σ^2 with a Normal-Inverse Gamma. By assuming that $(\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{Y}, M_\lambda) \approx \mathcal{NIG}_m(\boldsymbol{\mu}^*,\, \boldsymbol{V}^*,\, a^*,\, b^*)$, we will obtain an approximation for the parameters defining the posterior distribution on σ^2 by matching the first and second derivatives (with respect to σ^2) between the actual and postulated posterior on the full parameters. This approximation gives exact results if only uncensored data are to be dealt with.

Let us recall that for the posterior $\mathcal{NIG}_m^*$ density is

$$f^*(\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{Y}, M_\lambda) \propto [\sigma^2]^{-\frac{2a^*+m+2}{2}} \exp\left\{ -\frac{1}{2\sigma^2}\Big[2b^* + (\boldsymbol{\alpha}-\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu}^*)\Big] \right\}$$

and hence

$$\begin{aligned}
\ell^* &= \log f^*(\boldsymbol{\alpha},\, \sigma^2|\boldsymbol{Y}, M_\lambda) \\
&= k - \frac{1}{2}(2a^*+m+2)\log\sigma^2 - \frac{1}{2\sigma^2}\Big[2b^* + (\boldsymbol{\alpha}-\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu}^*)\Big] \\
&= k - \frac{1}{2}(2a^*+m+2)\log\, S - \frac{1}{2}S^{-1}\Big[2b^* + (\boldsymbol{\alpha}-\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}(\boldsymbol{\alpha}-\boldsymbol{\mu}^*)\Big]
\end{aligned}$$

Because the maximum of the log-posterior is attained at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ and $\sigma^2 = \hat{S}$ it must hold that

$$0 = \left.\frac{\partial \ell^*}{\partial \boldsymbol{\alpha}}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}},\, S=\hat{S}} = -\frac{1}{2}\hat{S}^{-1}\Big[2\,(\hat{\boldsymbol{\alpha}}-\boldsymbol{\mu}^*)'\,(\boldsymbol{V}^*)^{-1}\Big] \tag{6.14}$$

$$\begin{aligned}
0 = \left.\frac{\partial \ell^*}{\partial\, S}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}},\, S=\hat{S}} &= -\frac{1}{2}(2a^*+m+2)\hat{S}^{-1} \\
&\quad + \frac{1}{2}\hat{S}^{-2}\Big[2b^* + (\hat{\boldsymbol{\alpha}}-\boldsymbol{\mu}^*)'(\boldsymbol{V}^*)^{-1}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\mu}^*)\Big]
\end{aligned}$$

yielding

$$\boldsymbol{\mu}^* = \hat{\boldsymbol{\alpha}}$$

$$\frac{2b^*}{2a^* + m + 2} = \hat{S} = \hat{\sigma}^2 \tag{6.15}$$

Taking into account the relationship (6.15), for the second derivative with respect to S,

$$\begin{aligned} \left.\frac{\partial^2 \ell^*}{\partial\, S^2}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}},\, S=\hat{S}} &= \frac{1}{2}(2a^* + m + 2)\hat{S}^{-2} - 2b^*\,\hat{S}^{-3} \\ &= \frac{(2a^* + m + 2)^3}{8(b^*)^2} - \frac{(2a^* + m + 2)^3}{4(b^*)^2} \\ &= -\frac{(2a^* + m + 2)^3}{8(b^*)^2} \end{aligned} \tag{6.16}$$

By solving for a^* and b^* from the equations (6.15) and (6.16) we get

$$a^* = -\hat{S}^2 \left.\frac{\partial^2 \ell^*}{\partial S^2}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}},\, S=\hat{S}} - \frac{m+2}{2} \tag{6.17}$$

$$b^* = -\hat{S}^3 \left.\frac{\partial^2 \ell^*}{\partial S^2}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}},\, S=\hat{S}} \tag{6.18}$$

where the value of $\left.\dfrac{\partial^2 \ell^*}{\partial S^2}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}},\, S=\hat{S}}$ should be approximated by evaluating equation (6.8) at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ and $S = \hat{S}$.

Finally, from equation (6.14) we obtain $\dfrac{\partial^2 \ell^*}{\partial \boldsymbol{\alpha}^2} = -(\boldsymbol{V}^*)^{-1}\hat{S}^{-1}$. Matching this equation with (6.10), it turns out that

$$\boldsymbol{V}^* = \left(\boldsymbol{B}'\boldsymbol{W}\boldsymbol{B} + \boldsymbol{V}^{-1}\right)^{-1} \tag{6.19}$$

Figure 6.5 is an illustrative example depicting the approximated posterior Inverse Gamma distribution for σ^2.

Chib (1992) proposed proper Laplace approximations including a Gaussian approximation to the posterior distribution of σ^2. Note that, in contrast to the approach set out in this chapter, this does not yield an exact answer even if there are no censored observations.

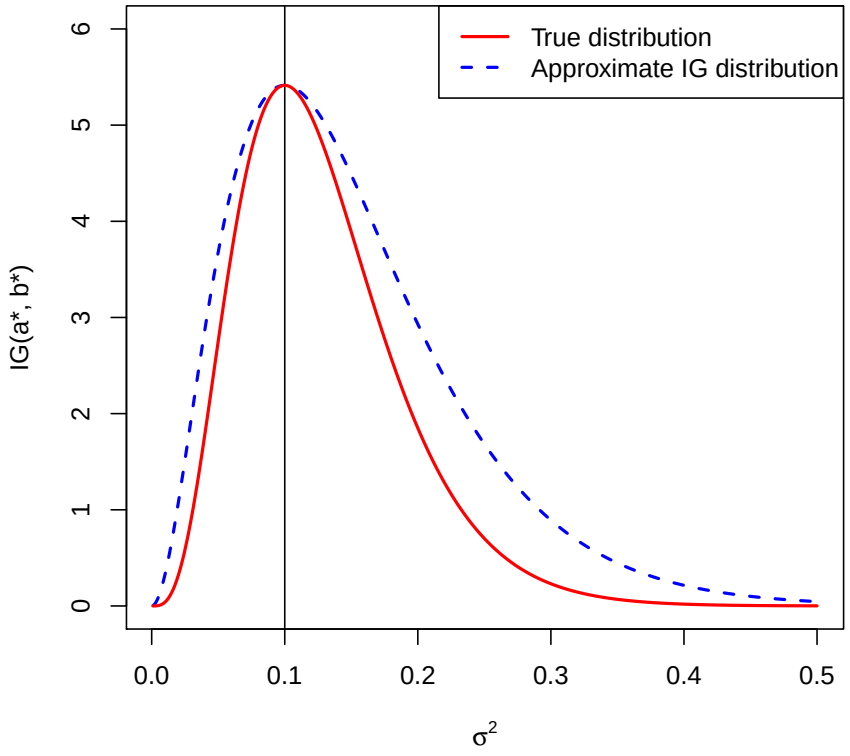


FIGURE 6.5: Illustrative example corresponding to the Laplace-type approximation for the posterior distribution of σ^2

6.9 Other possible approximations

If we denote by $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \sigma^2)$ our full vector of parameters to be estimated, let us recall that our objective is to approximate the posterior $f^*(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda)$ given by equation (6.4), by another function $q(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda)$ which is more tractable. The Laplace-type quadratic approximation proposed is far from being optimal when the number of non-detects increases. In this section we shall consider briefly two other alternatives.

6.9.1 Variational Bayes approximation

One can show (see i.e. Bishop, 2006) that for any approximation $q(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda)$ to the log-marginal posterior $f^*(\boldsymbol{Y}, \lambda)$ it holds that

$$\log f^*(\boldsymbol{Y}, \lambda) = \mathcal{L}(q) + KL(q\|f^*) \tag{6.20}$$

$$\text{where} \quad \mathcal{L}(q) = \int q(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda) \log \frac{f^*(\boldsymbol{Y}, \lambda, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda)} d\boldsymbol{\theta} \tag{6.21}$$

$$\text{and} \quad KL(q\|f^*) = \int q(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda) \log \frac{q(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda)}{f^*(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda)} d\boldsymbol{\theta} \tag{6.22}$$

The expression $KL(q\|f^*)$ is known as the *Kullback-Leibler divergence* and is always non-negative (see Bishop, 2006). Consequently it holds that $\log f^*(\boldsymbol{Y}, \lambda) \geq \mathcal{L}(q)$ and the lower bound is attained when $q(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda) = f^*(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda)$.

In our context of censored data, the variational strategy can closely follow the approaches employed for probit regression and classification (see Girolami and Rogers, 2006; Seeger, 2000).

6.9.2 Expectation Propagation

This technique is also based on minimising the Kullback-Leibler divergence but in the reverse form, leading to an approximation with rather different properties.

We assume that our true posterior can be written as

$$f^*\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right) \quad = \quad \frac{1}{K}\, f_0(\boldsymbol{\theta}) \prod f_i\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right) \tag{6.23}$$

where $\prod f_i\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right)$ represents a factorisation of the likelihood function, $f_0(\boldsymbol{\theta})$ is the prior over the vector of parameters $\boldsymbol{\theta}$ and K is the normalising constant. We will look for an approximation to $f^*\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right)$ of the form

$$q\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right) \quad = \quad \frac{1}{Z}\, \tilde{f}_0\left(\boldsymbol{\theta}\right) \prod \tilde{f}_i\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right) \tag{6.24}$$

where each $\tilde{f}_i$ approximates the corresponding f_i and Z is the normalising constant. Ideally, we would aim at minimising $KL\left(f^*\|q\right)$ with respect to $q\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right)$ considering $f^*\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right)$ fixed but this task would involve averaging over the true distribution. Expectation propagation proceeds by minimising the divergences between f_i and $\tilde{f}_i$. Each factor $\tilde{f}_i$ is constrained to come from an exponential family and hence the whole approximation $q\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right)$ will belong to this family too. The idea of this method is to proceed iteratively to revise each factor $\tilde{f}_j$ in turn, in the context of all the remaining factors.

If we call q^{new} the approximation to $q\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right)$ at each new step, expectation propagation seeks to minimise $KL\left(f_j \prod_{i\neq j} \tilde{f}_i,\, q^{new}\right)$. Bishop (2006) shows that because $q\left(\boldsymbol{\theta}|\boldsymbol{Y},\,\lambda\right)$ is assumed to belong to an exponential family, the minimisation is achieved by matching the moments between q^{new} and $f_j \prod_{i\neq j} \tilde{f}_i$, provided that this task is tractable. Thus we can make $\tilde{f}_j \propto \frac{q^{new}}{\prod_{i\neq j} \tilde{f}_i}$. The iteration process is repeated till convergence. The normalisation constant Z is finally chosen as $\frac{1}{\int \prod \tilde{f}_i\, d\boldsymbol{\theta}}$. Bishop (2006) provides the whole algorithm in detail.

Additionally to tractability in all the operations involved, another disadvantage of expectation propagations is that there is no guarantee of convergence. Besides, if the true distribution $f^*(\boldsymbol{\theta}|\boldsymbol{Y}, \lambda)$ were multimodal (which is not our case), minimising $KL(f^*\|q)$ might yield a poor approximation.

Given that the censored data component of our model is equivalent to probit regression, one can use the same strategies for deriving expectation propagation approximations as used for probit regression (see e.g. Rasmussen and Williams, 2006). It should be noticed that no approximations are needed for the non-censored part of the data.

6.10 The choice of the penalisation parameter

Recalling equation (3.25), the posterior distribution of the penalisation parameter is

$$f_{M_\lambda|\boldsymbol{Y}} \propto \lambda^{\frac{\text{rank}(D'D)}{2}} \times \frac{\Gamma(a^*)\left|V^*\right|^{\frac{1}{2}}}{(b^*)^{a^*}}$$

Therefore, the idea is to identify the value of λ that maximises $f_{M_\lambda|\boldsymbol{Y}}$ using the estimators of a^*, b^* and V^* obtained in equations (6.17), (6.18) and (6.19) respectively. It should be noticed that these estimators of the posterior parameters are functions of λ and hence they have to be recomputed for every potential value of the penalisation parameter to determine its optimal value.

6.11 Importance Sampling

One way to assess the goodness of our approximation and correct it, is by means of the technique of *importance sampling* (see Gentle, 2002; Tanner, 1996).

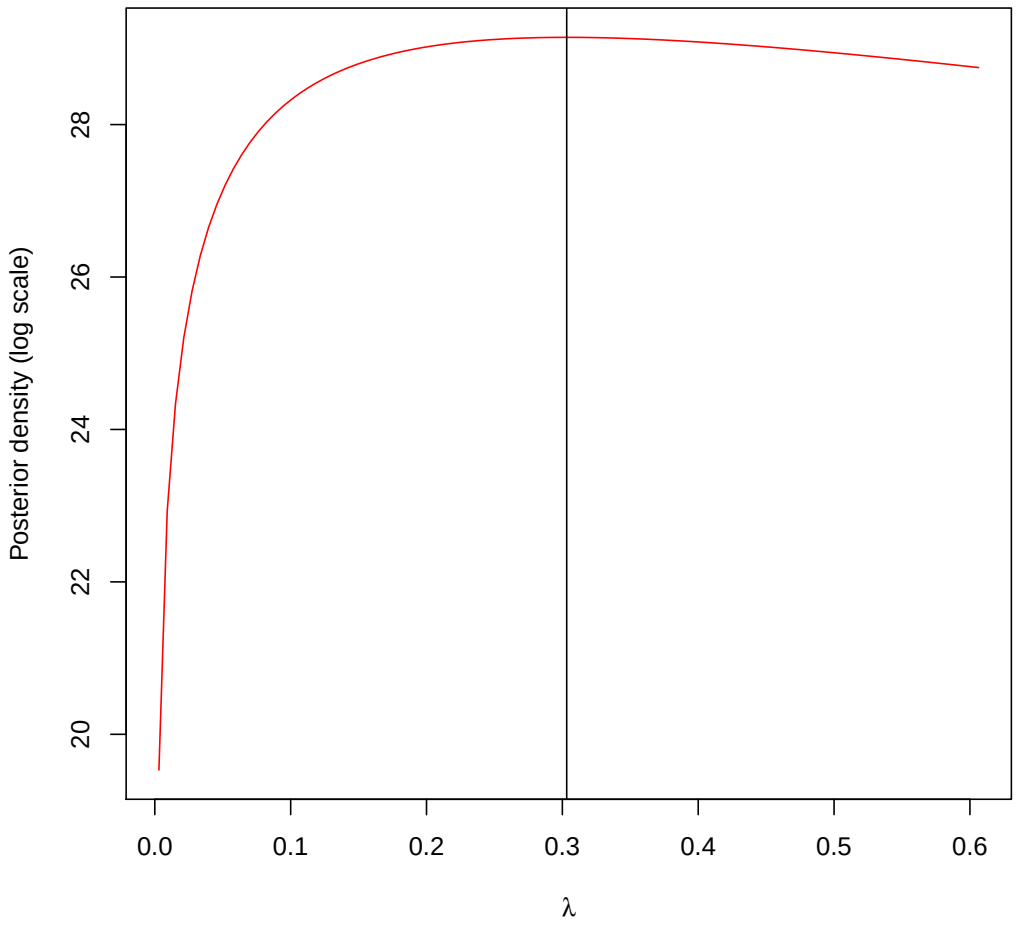


FIGURE 6.6: Optimal value of λ based on the approximated $f_{M_\lambda|\boldsymbol{Y}}$

For the sake of notation, let us call $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \sigma^2)$. Up to a normalising constant K, we can compute the true posterior distribution of $\boldsymbol{\eta}$ using equation (6.4). Let us denote by $f^*(\boldsymbol{\eta}|\boldsymbol{Y}, \lambda)$ our $\mathcal{NIG}_m^*(\boldsymbol{\mu}^*, \boldsymbol{V}(\lambda)^*, a^*, b^*)$ Laplace-type approximation to the **true** posterior distribution $K\, f(\boldsymbol{\eta}|\boldsymbol{Y}, \lambda)$.

Because our approximation is fully known, we can draw a sample $\boldsymbol{\eta}_i$ from it with $i = 1, \ldots, N$. Using this sample, we can construct the sequence of *weights* $w_i = \dfrac{f(\boldsymbol{\eta}_i|\boldsymbol{Y}, \lambda)}{f^*(\boldsymbol{\eta}_i|\boldsymbol{Y}, \lambda)}$. Assuming that the expectation of the random vector $w\boldsymbol{\eta}$ with respect to $K\, f(\boldsymbol{\eta}|\boldsymbol{Y}, \lambda)$ exists, it must be

$$\lim_{N\to\infty} \frac{1}{N} \sum w_i \boldsymbol{\eta}_i \quad = \quad \mathbb{E}\Big[w\,\boldsymbol{\eta}|\boldsymbol{Y}, \lambda\Big]$$

$$
\begin{aligned}
&= \int w\boldsymbol{\eta}\, f^*(\boldsymbol{\eta}|\boldsymbol{Y},\lambda)\, d\boldsymbol{\eta} \\
&= \frac{1}{K}\int K\frac{f(\boldsymbol{\eta}|\boldsymbol{Y},\lambda)}{f^*(\boldsymbol{\eta}|\boldsymbol{Y},\lambda)}\,\boldsymbol{\eta}\, f^*(\boldsymbol{\eta}|\boldsymbol{Y},\lambda)\, d\boldsymbol{\eta} \\
&= \frac{1}{K}\int \boldsymbol{\eta}\, K f(\boldsymbol{\eta}|\boldsymbol{Y},\lambda)\, d\boldsymbol{\eta} \\
&= \frac{1}{K}\mathbb{E}(\boldsymbol{\eta}|\boldsymbol{Y},\lambda) \qquad (6.25)
\end{aligned}
$$

In particular, for $\boldsymbol{\eta} = (1, \ldots, 1)'$, equation (6.25) yields

$$\lim_{N\to\infty} \bar{w} = \lim_{N\to\infty} \frac{1}{N}\sum w_i = \frac{1}{K} \qquad (6.26)$$

From equations (6.25) and (6.26) we can obtain an approximation for the **true** posterior expectation of the vector of parameters $\boldsymbol{\eta}$

$$\mathbb{E}(\boldsymbol{\eta}|\boldsymbol{Y},\lambda) \approx \sum \tilde{w}_i\boldsymbol{\eta}_i = \hat{\mathbb{E}}(\boldsymbol{\eta}|\boldsymbol{Y},\lambda) \qquad \text{with} \qquad \tilde{w}_i = \frac{w_i}{N\,\bar{w}} \qquad (6.27)$$

Although $\hat{\mathbb{E}}(\boldsymbol{\eta}|\boldsymbol{Y},\lambda)$ is not an unbiased estimator of $\mathbb{E}(\boldsymbol{\eta}|\boldsymbol{Y},\lambda)$, it can be proved that under weak assumptions, $\lim_{N\to\infty}\hat{\mathbb{E}}(\boldsymbol{\eta}|\boldsymbol{Y},\lambda) = \mathbb{E}(\boldsymbol{\eta}|\boldsymbol{Y},\lambda)$ almost surely (see Tanner, 1996).

The relationship (6.27) can be used to estimate any linear function of the parameters; in particular, the mean of the regression function for a given value of the covariates.

The proposed Laplace-type approximation produces always symmetric confidence intervals; hence, it works fairly well when the percentage of censored values is small. Equation (6.27) can also be used to correct for this effect as it allows

the computation of quantiles. Using Quasi-Monte Carlo strategies (see Lemieux, 2009), convergence can be achieved for importance sampling with a smaller number of points. In spite of this, the algorithm is not very efficient in practice as it is also very time-consuming.

6.12 Illustrative Example

As an illustrative example, we have simulated $N = 100$ points from the linear model

$$Y = -10 + 2X + \varepsilon \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0,\, \sigma^2 = 5^2) \tag{6.28}$$

A simple linear model was fitted to the data, after censoring had been applied (a two-parameters linear model was used to be able to illustrate the results). We have plotted (on a log scale) the weights $\tilde{w}_i$ defined in equation (6.27) for the distribution of the intercept vs. the slope using normal distance and the Mahalanobis distance. In the first case the values were normalised and centered in the second.

Figure 6.7 corresponds to the case of fully uncensored observations. It can be noticed, that except for error representation, the values of the weights are practically 1 everywhere.

The plots related to the case with a 30% of censored observations are displayed in Figure 6.8. In this case we note that the weights vary between $0.6 \approx e^{-0.5}$ and $1.5 \approx e^{0.4}$. The differences occur in the tails of the distributions where our proposed posterior approximation $f^*(\boldsymbol{\eta}|\boldsymbol{Y}, \lambda)$ differs from the true posterior density $f(\boldsymbol{\eta}|\boldsymbol{Y}, \lambda)$. However these differences do not seem to be remarkable.

Figure 6.9 pictures the more extreme case with a 50% of censored observations. Here the weights range between $0.5 \approx e^{-0.8}$ and $1.8 \approx e^{0.6}$. Again, we see that differences in weights occur in the tails although they are not really important.

The corresponding values of the smoothing parameter are $\lambda_0 = 0.60$, $\lambda_{30} = 0.53$ and $\lambda_{50} = 0.84$.

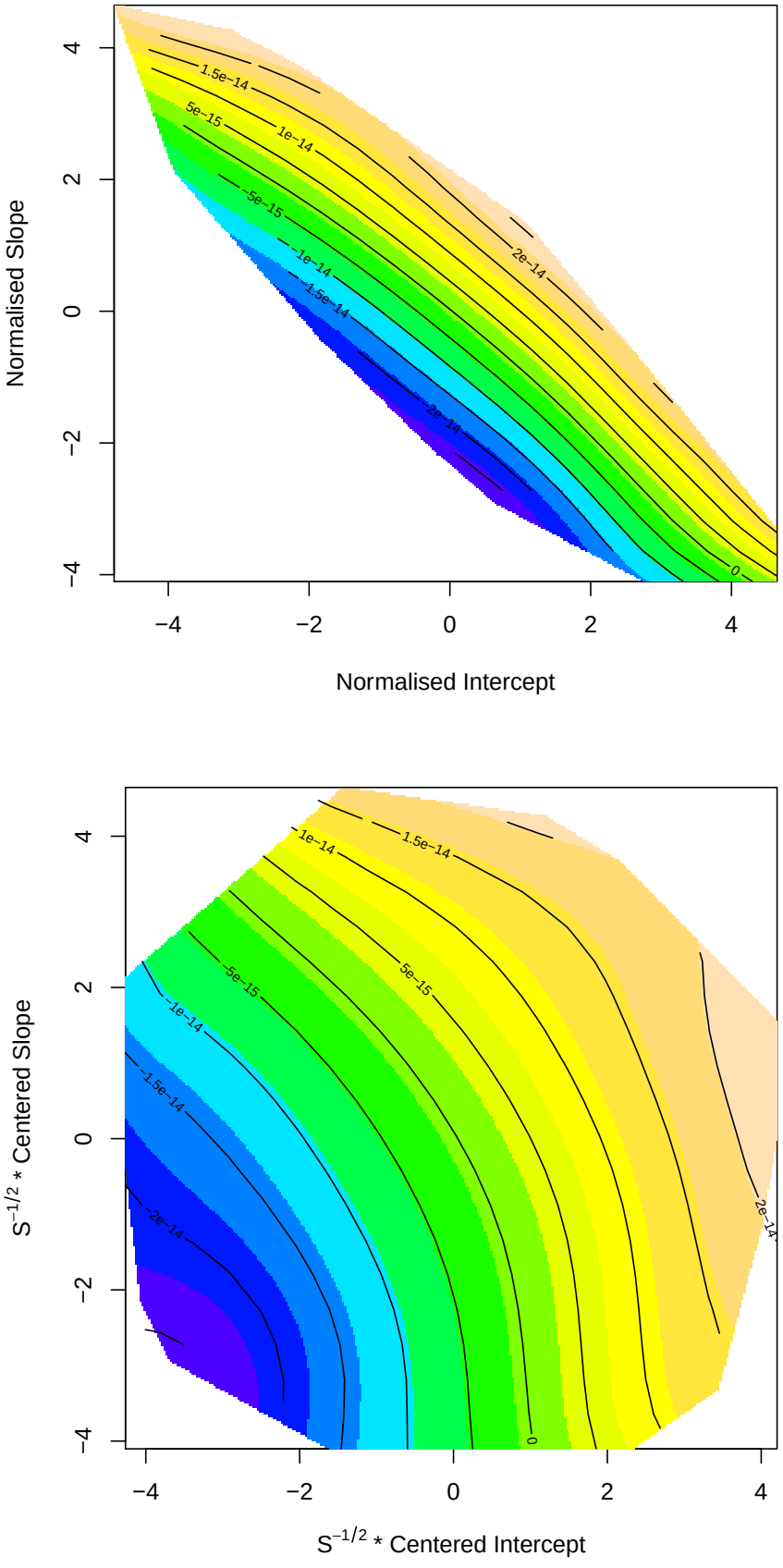


FIGURE 6.7: Weights (in log scale) of normalised parameters using normal distance (top) and Mahalanobis distance (bottom) - Without non-detects

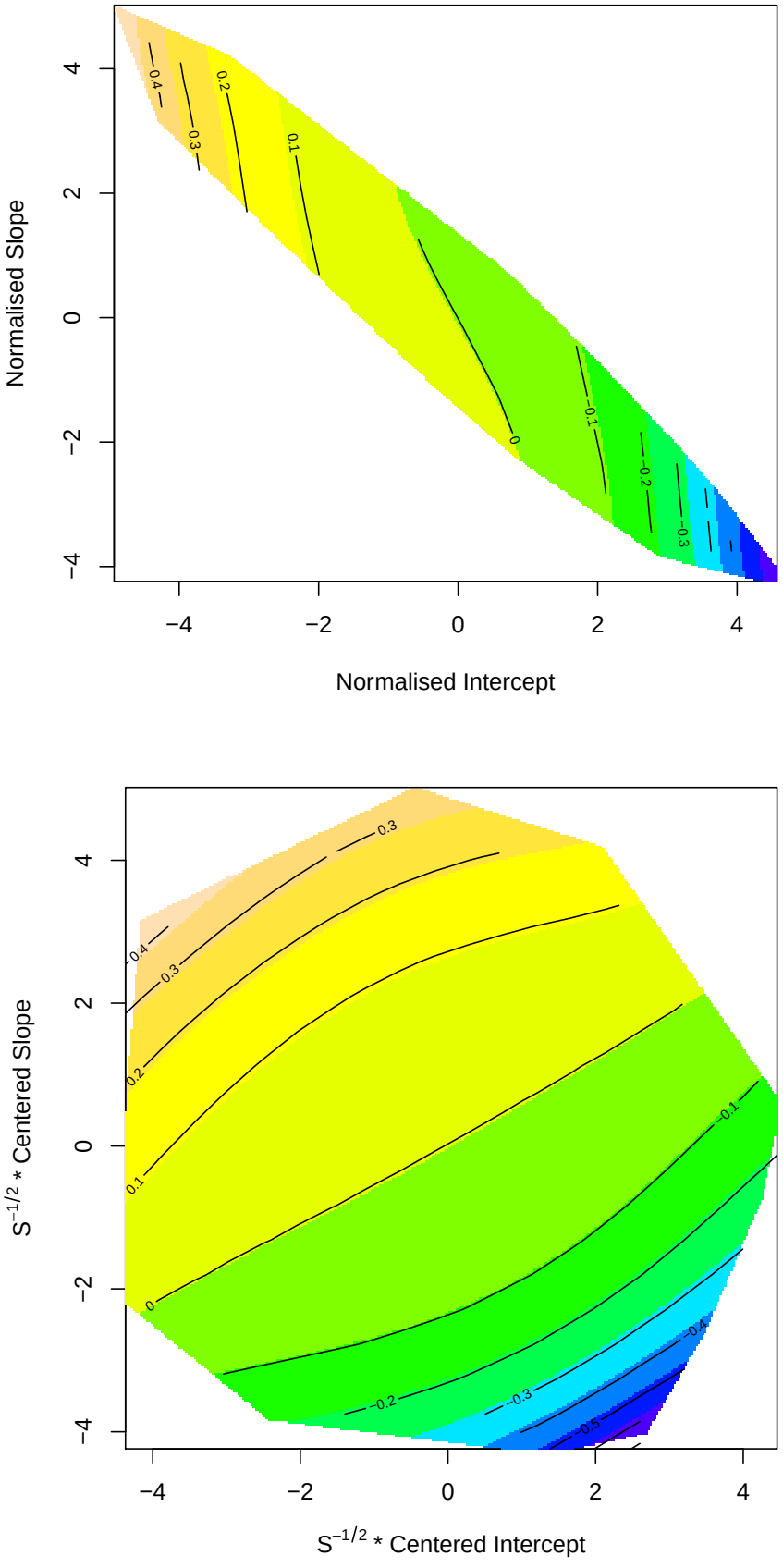


FIGURE 6.8: Weights (in log scale) of normalised parameters using normal distance (top) and Mahalanobis distance (bottom) - Percentage of non-detects: 30%

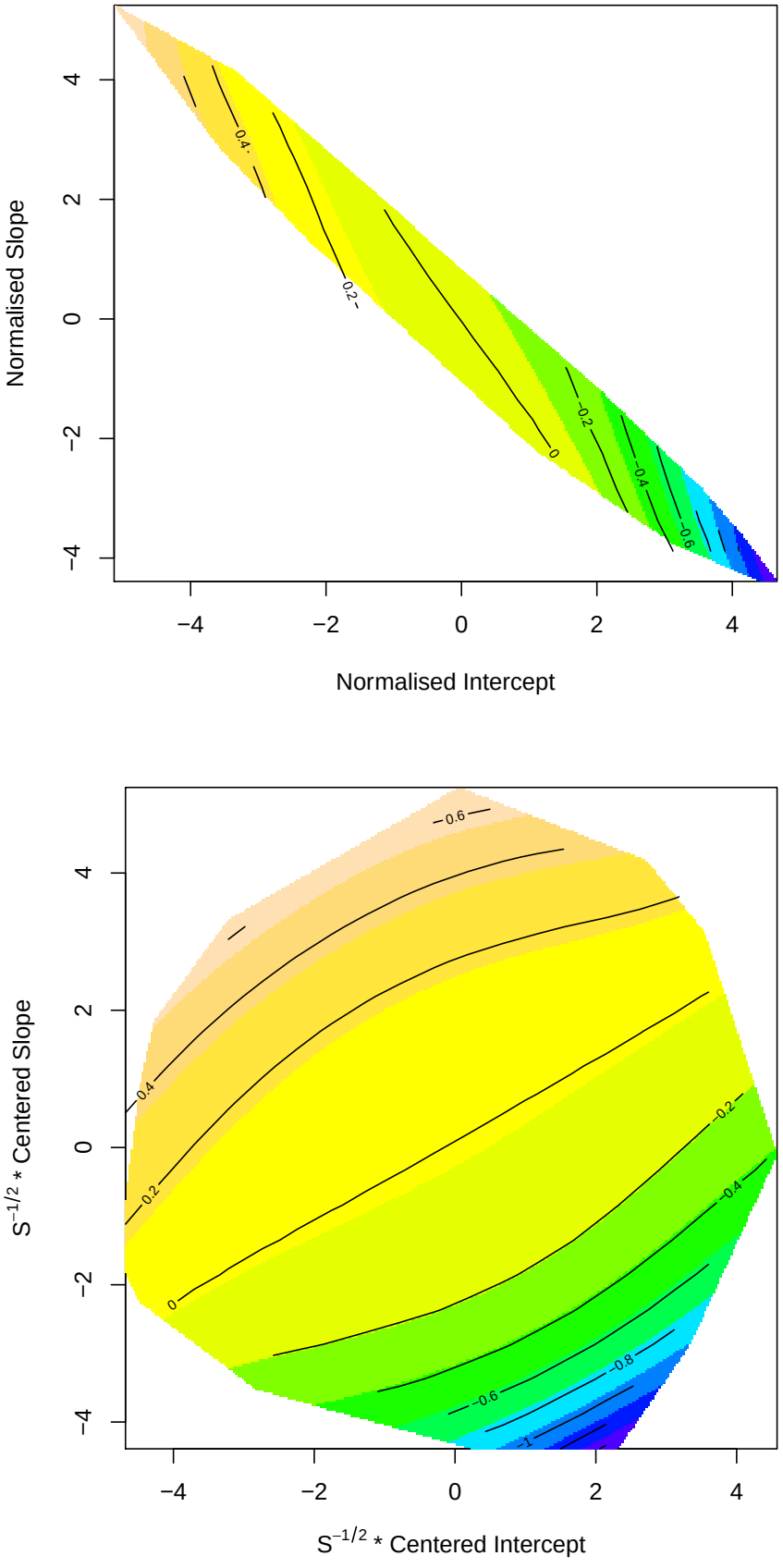


FIGURE 6.9: Weights (in log scale) of normalised parameters using normal distance (top) and Mahalanobis distance (bottom) - Percentage of non-detects: 50%

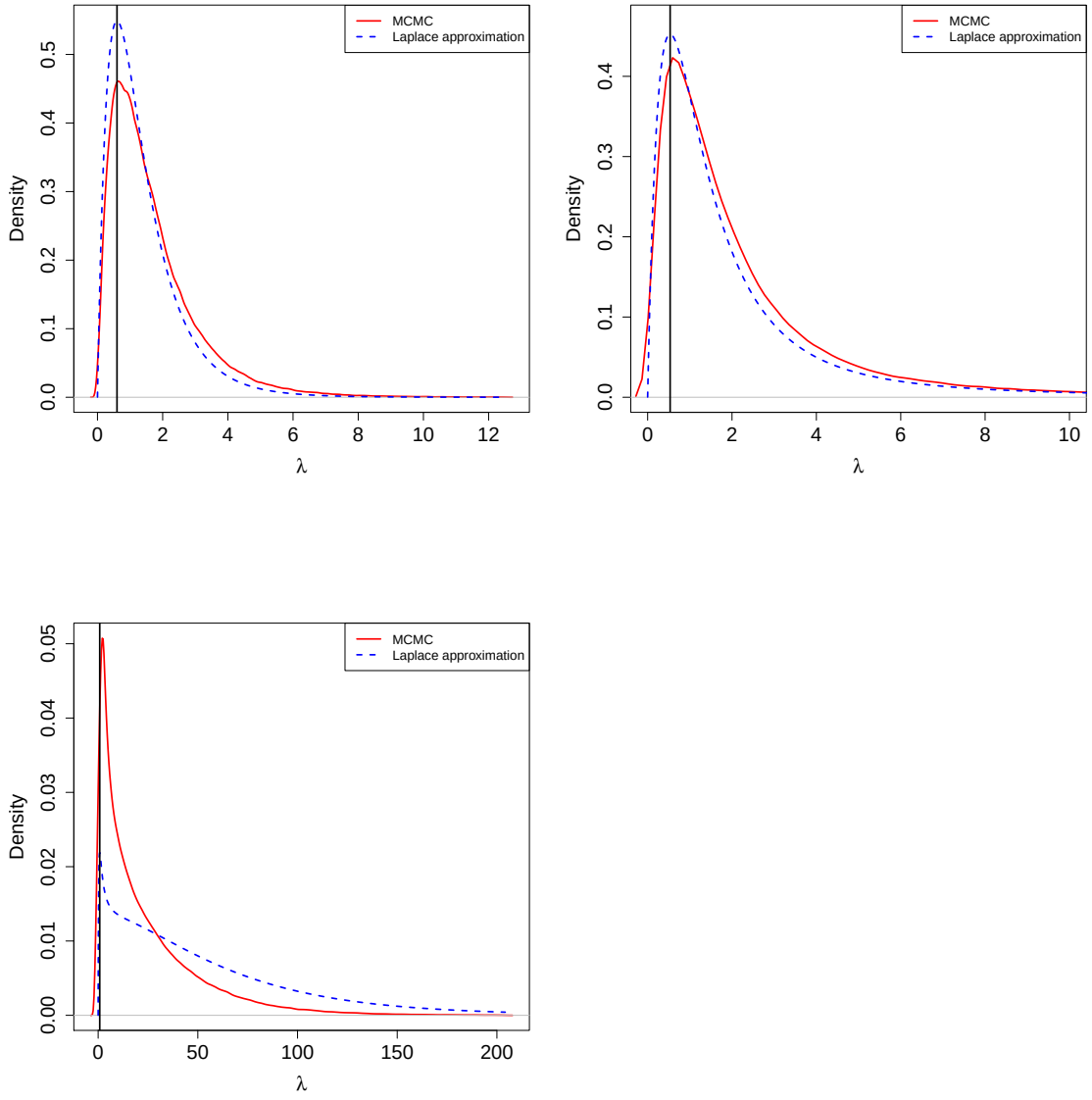


FIGURE 6.10: Comparison of the distribution of the penalisation parameter λ evaluated using MCMC and the approximate $f_{M_\lambda|\boldsymbol{Y}}$ for different percentages of non-detects: 0% (top left), 30% (top right) and 50% (bottom)

Figure 6.10 shows the distribution of the penalisation parameter λ evaluated using MCMC and the approximate $f_{M_\lambda|\boldsymbol{Y}}$ (equation (3.25)) for different percentages of contamination. It can be noticed that although the mode is the same in all cases, the shape differs noticeably if high levels of non-detects are considered.

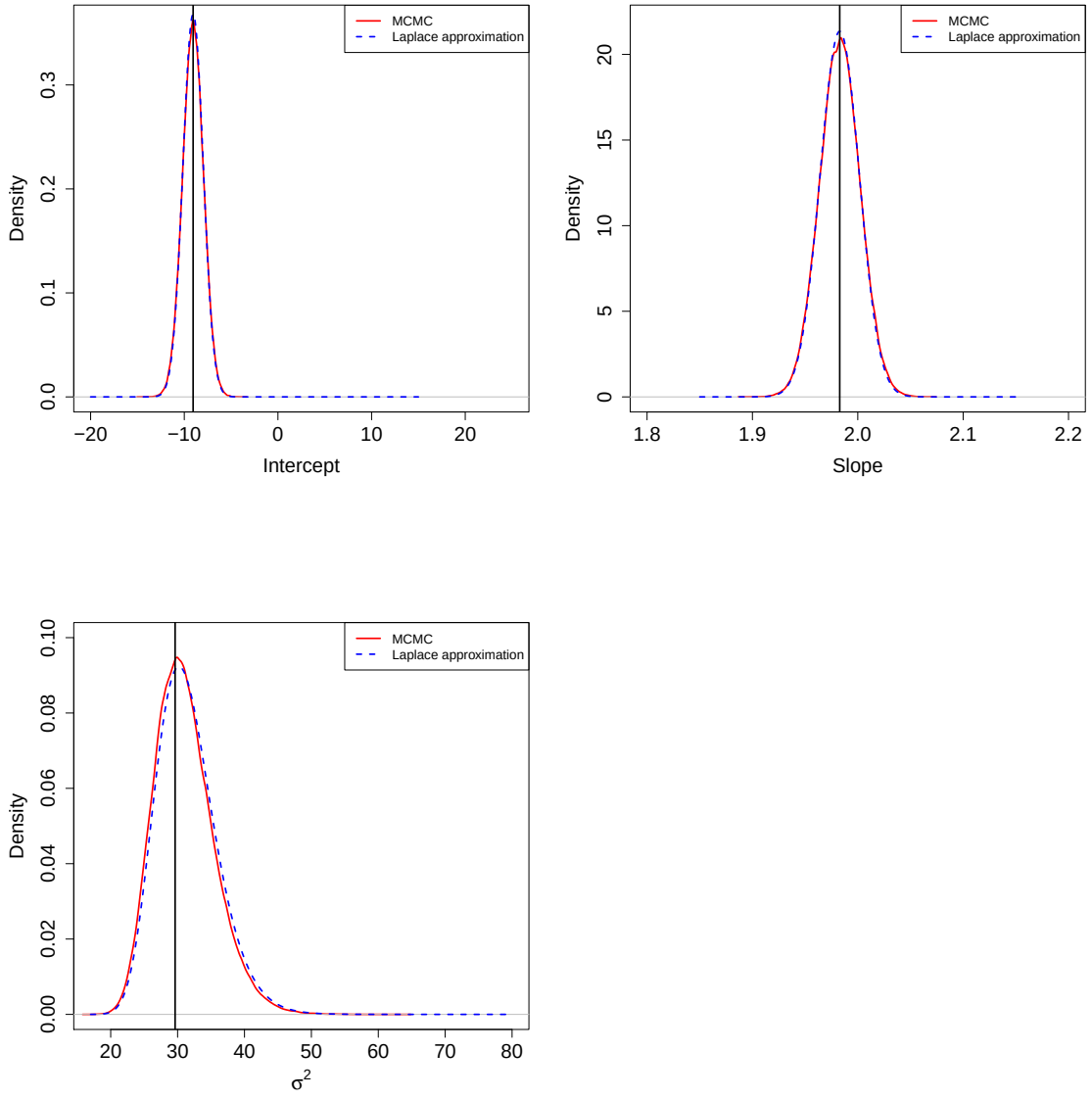


FIGURE 6.11: Comparison of the distribution of the Intercept, Slope and σ^2 using MCMC and the Laplace-type approximation without non-detects. The penalisation parameter was computed using $f_{M_\lambda|\boldsymbol{Y}}$ and is the same in both cases

Figure 6.11 shows the distribution of the parameters of the model without non-detects, computed using MCMC and the Laplace-type approximation. For both methods, the value of the penalisation parameter is fixed and corresponds to the optimal value according to $f_{M_\lambda|\boldsymbol{Y}}$, which is exact in this case. As expected, both methods yield the same distribution; but remarkably, this is even the case when we have non-detects, as it can be noticed in Figures 6.12 and 6.13.

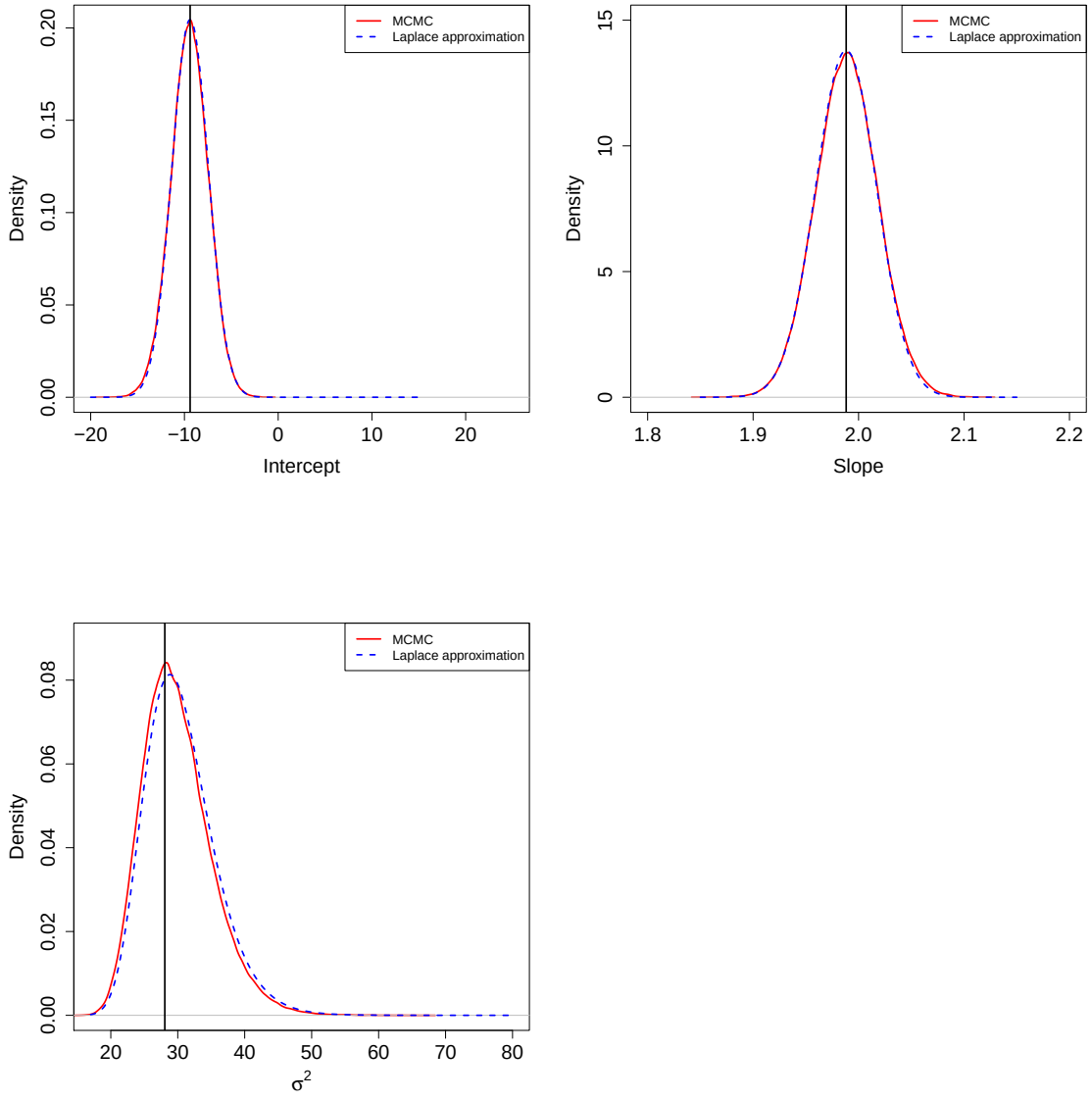


FIGURE 6.12: Comparison of the distribution of the Intercept, Slope and σ^2 using MCMC and the Laplace-type approximation for 30% of non-detects. The penalisation parameter was computed using the approximate $f_{M_\lambda|\boldsymbol{Y}}$ and is the same in both cases

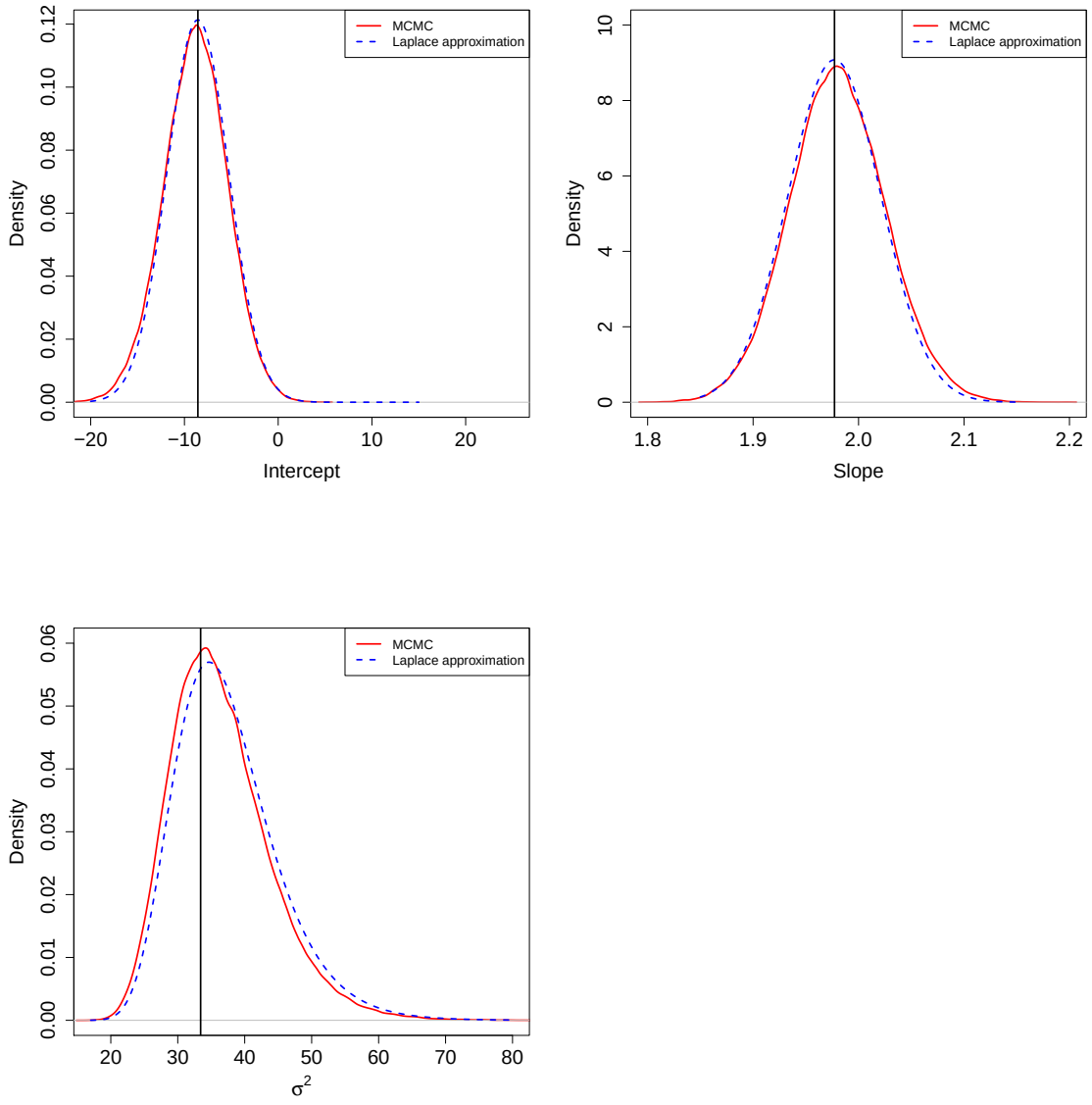


FIGURE 6.13: Comparison of the distribution of the Intercept, Slope and σ^2 using MCMC and the Laplace-type approximation for 50% of non-detects. The penalisation parameter was computed using the approximate $f_{M_\lambda|\boldsymbol{Y}}$ and is the same in both cases

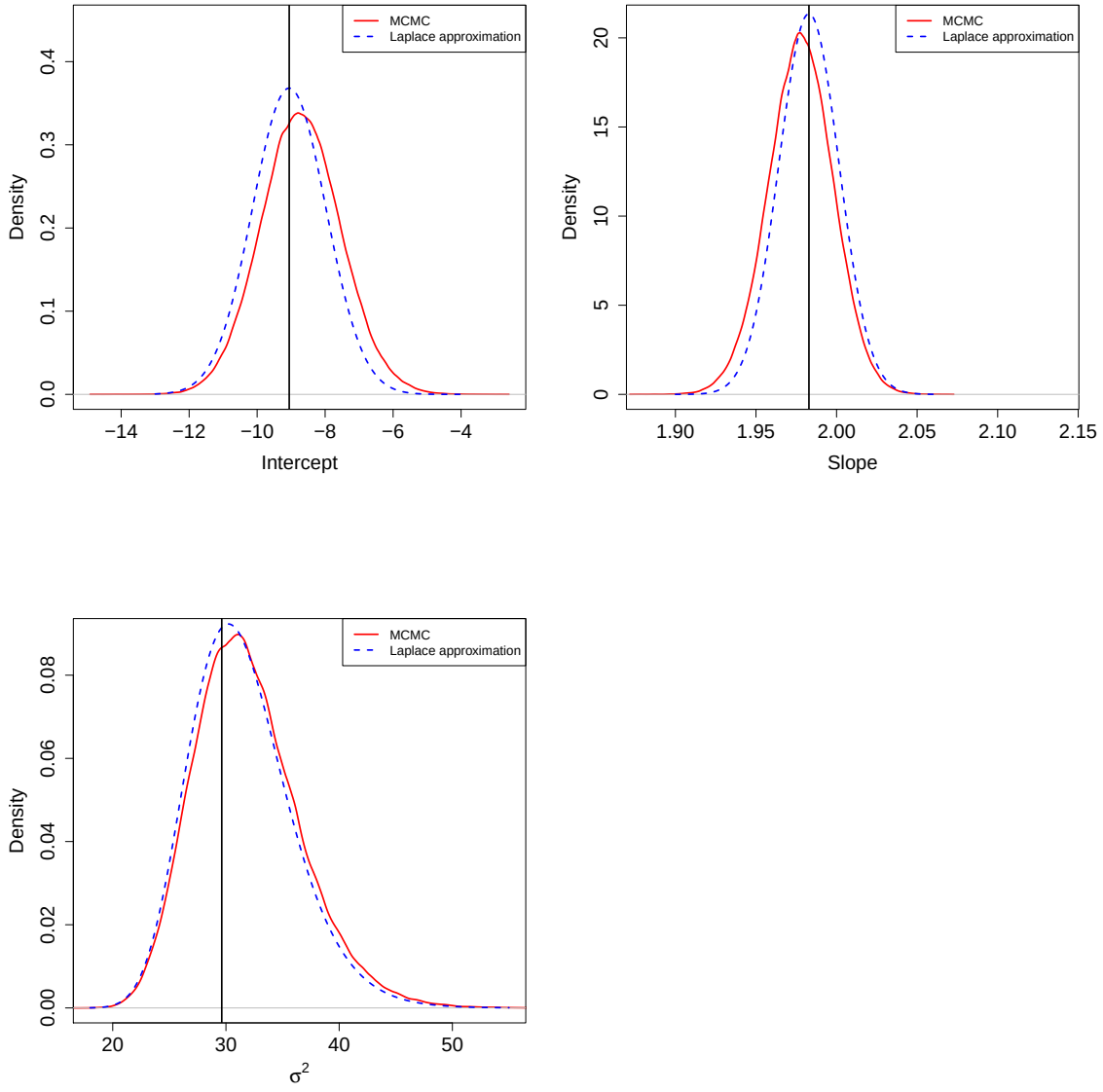


FIGURE 6.14: Comparison of the distribution of the Intercept, Slope and σ^2 using MCMC and the Laplace-type approximation without non-detects. The penalisation parameter was computed using MCMC in the first case and the approximate $f_{M_\lambda|\boldsymbol{Y}}$ in the second

Figures 6.14, 6.15 and 6.16 reproduce the same situation, except that the actual posterior distribution of the penalisation parameter is considered for MCMC. We notice that in this case, when there are censored data, the distribution of the regression parameters is very skewed and very different from the Laplace-type approximation. The reason for this is that the approximation of the posterior distribution of the penalisation parameter λ is rather poor.

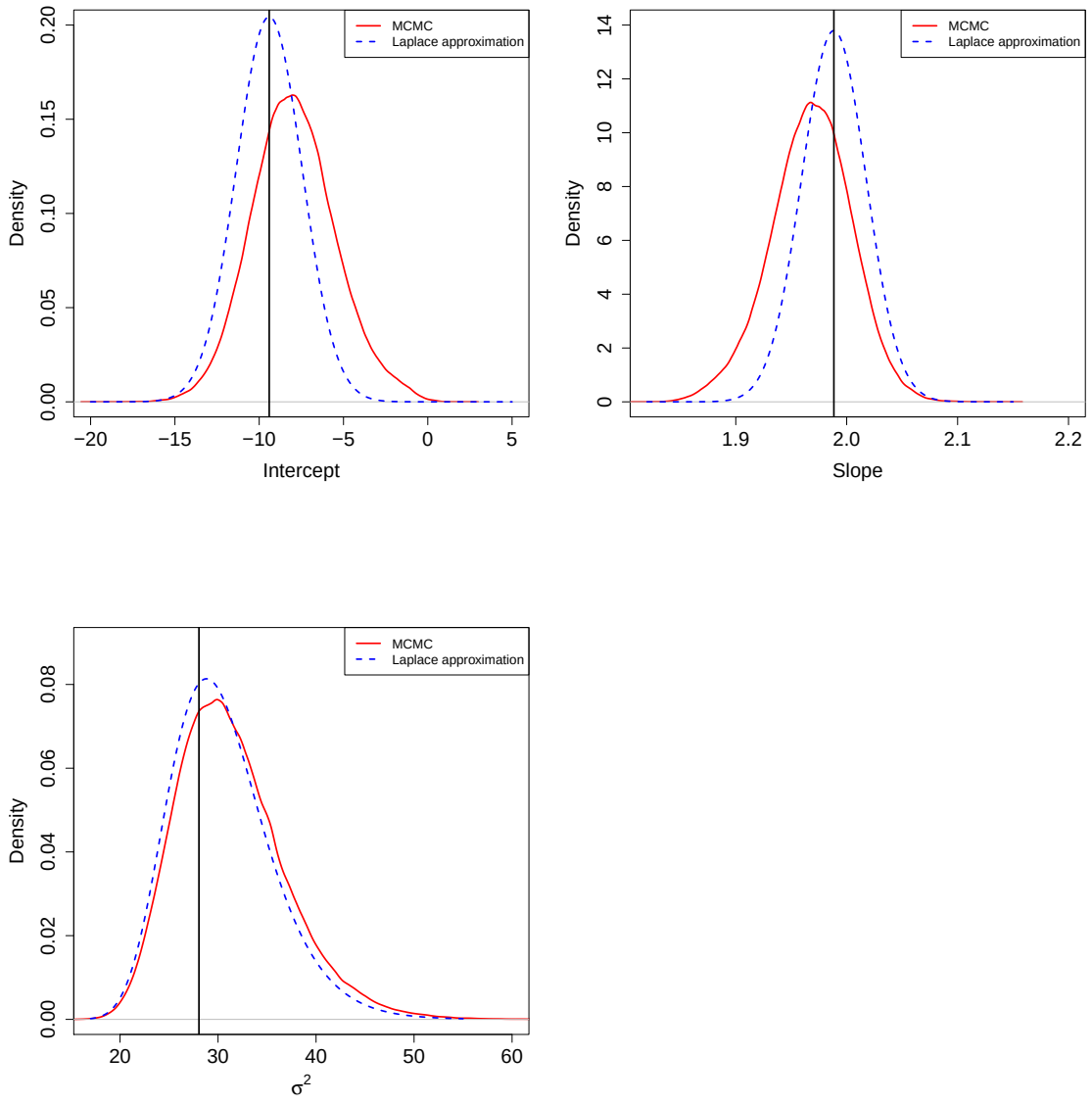


FIGURE 6.15: Comparison of the distribution of the Intercept, Slope and σ^2 using MCMC and the Laplace-type approximation for 30% of non-detects. The penalisation parameter was computed using MCMC in the first case and the approximate $f_{M_\lambda|\boldsymbol{Y}}$ in the second

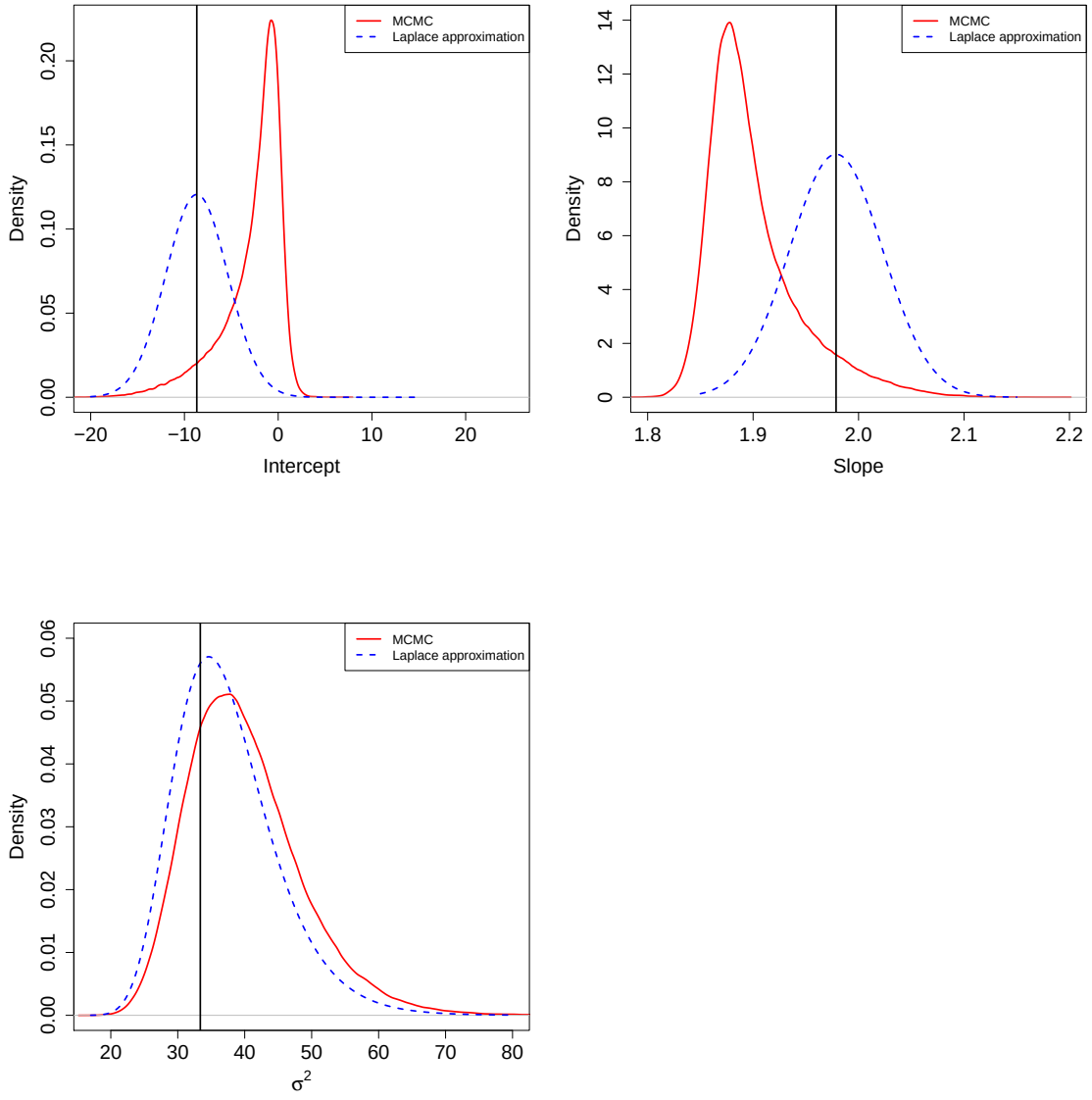


FIGURE 6.16: Comparison of the distribution of the Intercept, Slope and σ^2 using MCMC and the Laplace-type approximation for 50% of non-detects. The penalisation parameter was computed using MCMC in the first case and the approximate $f_{M_\lambda|\boldsymbol{Y}}$ in the second

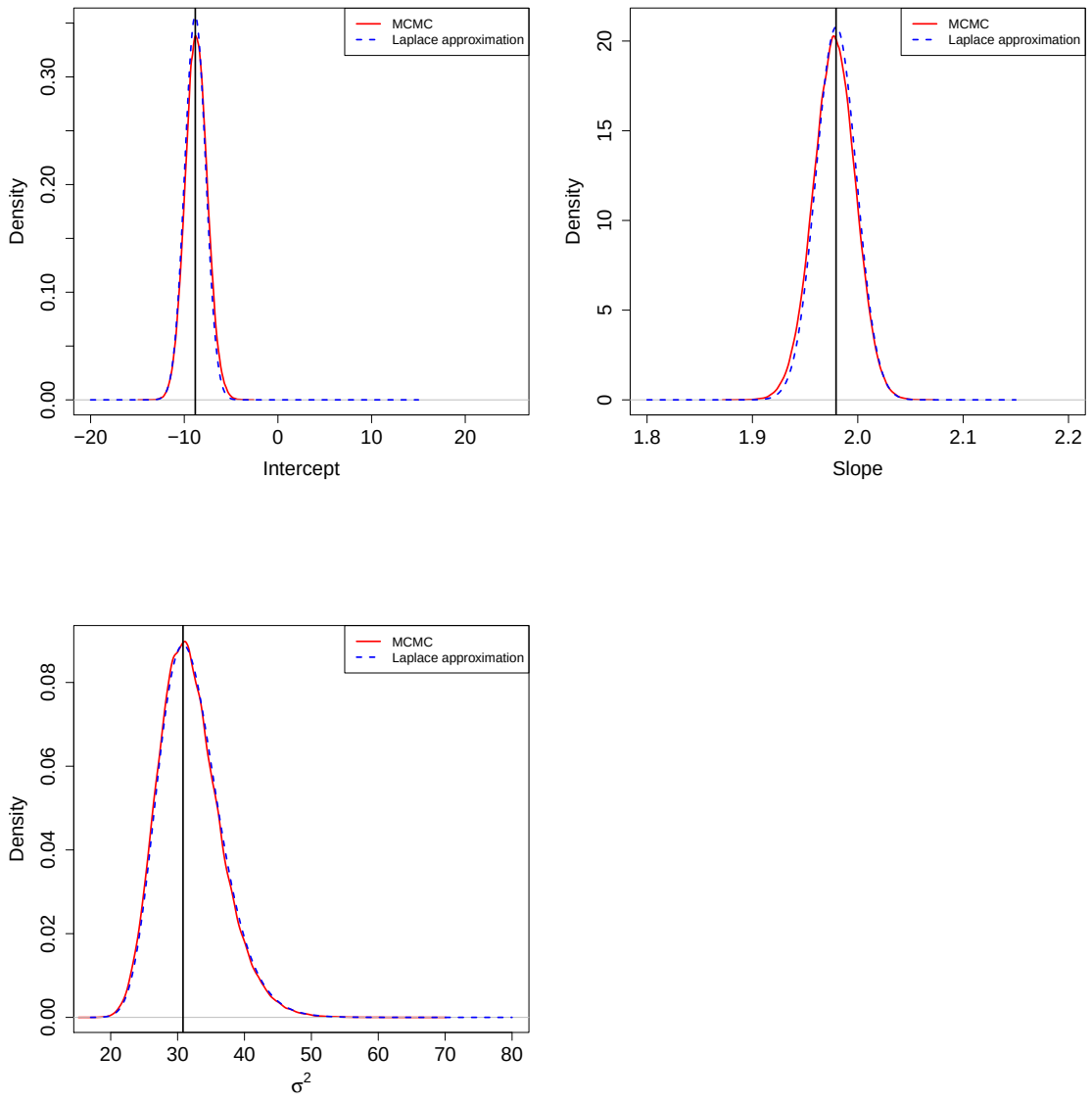


FIGURE 6.17: Comparison of the distribution of the Intercept, Slope and σ^2 using MCMC and the Laplace-type approximation with model averaging without non-detects

In Figures 6.17, 6.18 and 6.19 we have used model averaging rather than the MAP criterion to compute the approximate posterior distribution of the parameters for the Laplace-type approximation. We can see that by this means, the effect of skewness can be softened although not quite avoided.

As mentioned in section 3.6, not all the values of λ with posterior positive density are used in the averaging process. For computation efficiency, we select a subset

from the overall possible values and consider those such that $w_\lambda \geq \frac{1}{K}\max\{w_\lambda\}$. We have set $K = 20$, a typical value suggested in the literature (see Raftery et al., 1997).

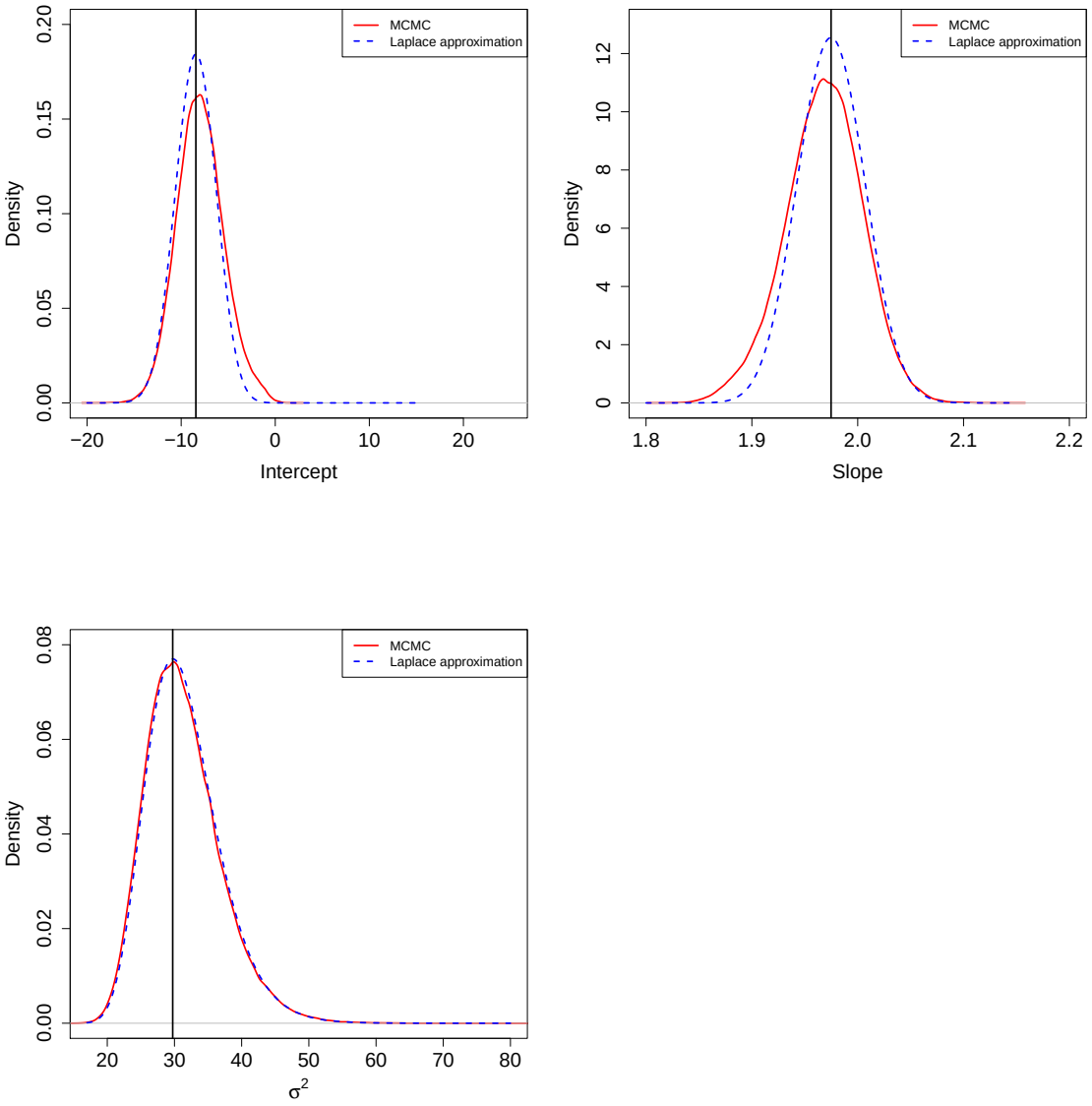


FIGURE 6.18: Comparison of the distribution of the Intercept, Slope and σ^2 using MCMC and the Laplace-type approximation with model averaging for 30% of non-detects

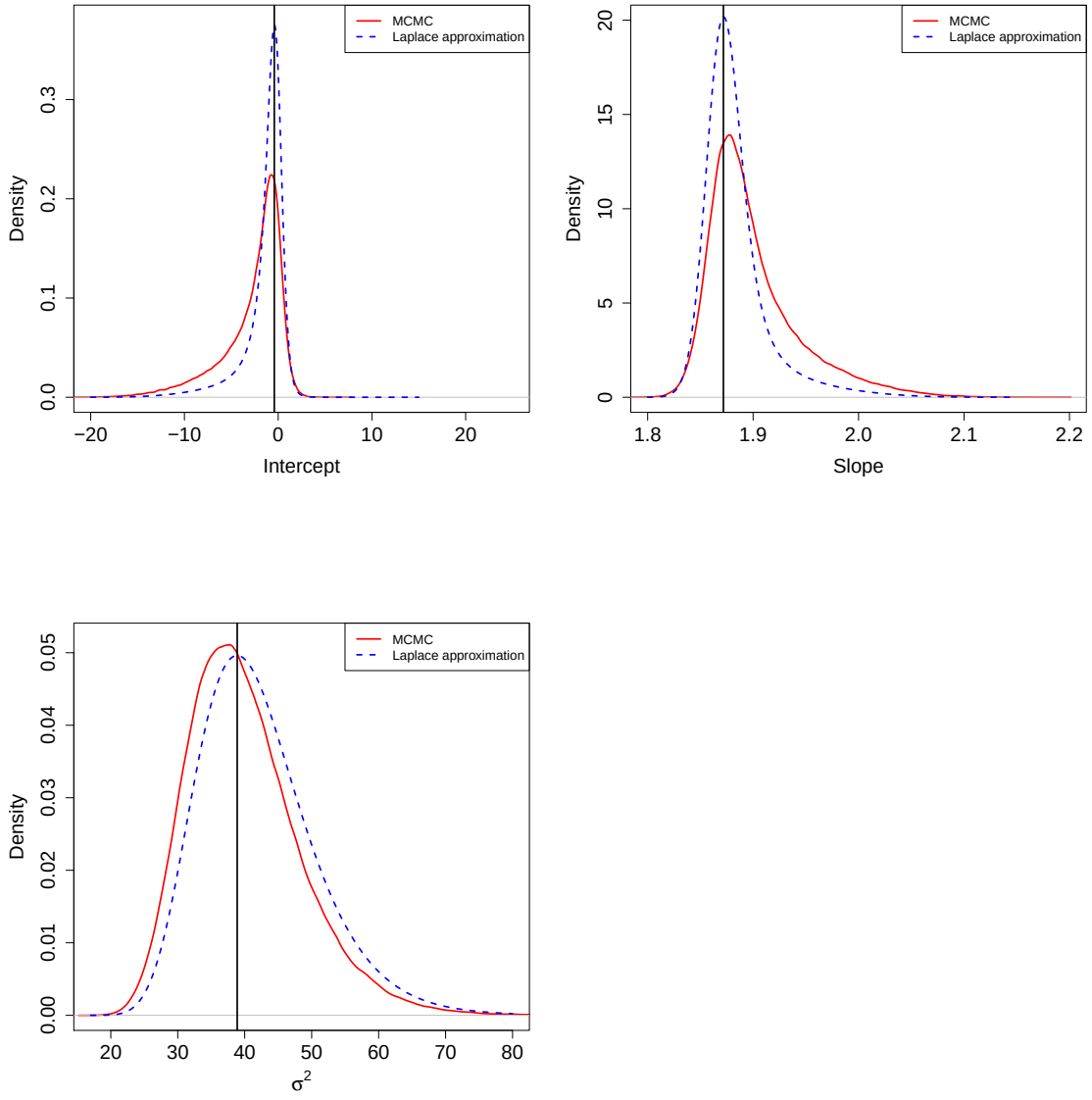


FIGURE 6.19: Comparison of the distribution of the Intercept, Slope and σ^2 using MCMC and the Laplace-type approximation with model averaging for 50% of non-detects

We conclude that the Laplace-type approximation proposed gives very accurate results, conditional on the smoothing parameter. Otherwise, the approximation works fairly well only for a low proportion of non-detects.

But under uncertainty on the penalisation parameter and as the proportion of non-detects increases, these posterior distributions become more skewed. Although the Laplace-type approximation generally manages to identify the mode of the

posterior distribution of the parameters, it fails to capture their shape properly. This undesired effect can be lessened by not quite avoided by using Bayesian model averaging rather than the Bayesian MAP approach for model selection.

6.13 Univariate example using Shell data

We will consider a time series of concentrations of a groundwater contaminant recorded over 1379 days at a well at an industrial site. 75 observations have been recorded, 49 (65%) of which are below the detection threshold. A B-spline basis with 25 basis functions is used for the design matrix.

The data, together with the predicted mean function and 95% prediction intervals, are shown in Figure 6.20 employing

- Laplace-type approximation with model averaging,
- Laplace-type approximation with MAP determination for the penalisation parameter,
- EM algorithm,
- Replacement of the censored values with one-half the detection limit

In the last two cases, the estimate of the penalty parameter maximising the approximation to the corresponding posterior distribution of λ is used. The values are $\lambda_{MAP} = 8.21$, $\lambda_{EM} = 1.87$ and $\lambda_{1/2DL} = 4.36$.

For the sake of comparison, Figure 6.21 reproduces Figure 6.20 using the same value of $\lambda = \lambda_{MAP} = 8.21$ in all cases. It can be noticed that whereas in the case of replacing the censored values by one-half the detection limit there is no sensible difference, in the EM case the fitted curve is smoother due to the higher order magnitude in the penalisation parameter.

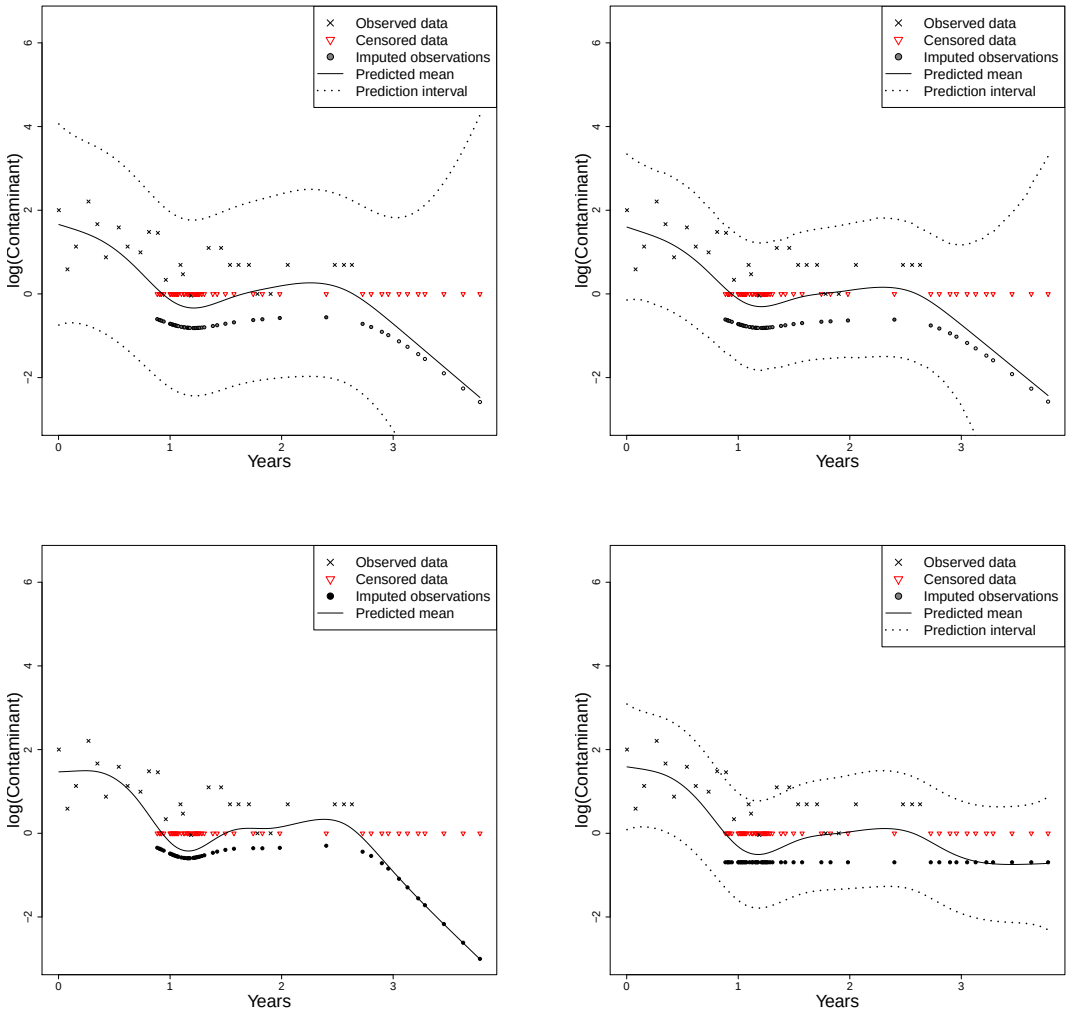


FIGURE 6.20: Predicted mean function and 95% prediction intervals for the contamination data obtained using the Laplace-type approximation with MAP (top left) and model averaging (top right), by replacing non-detects by 1/2 the detection-limit (bottom right) and predicted mean function using EM-algorithm (bottom left)

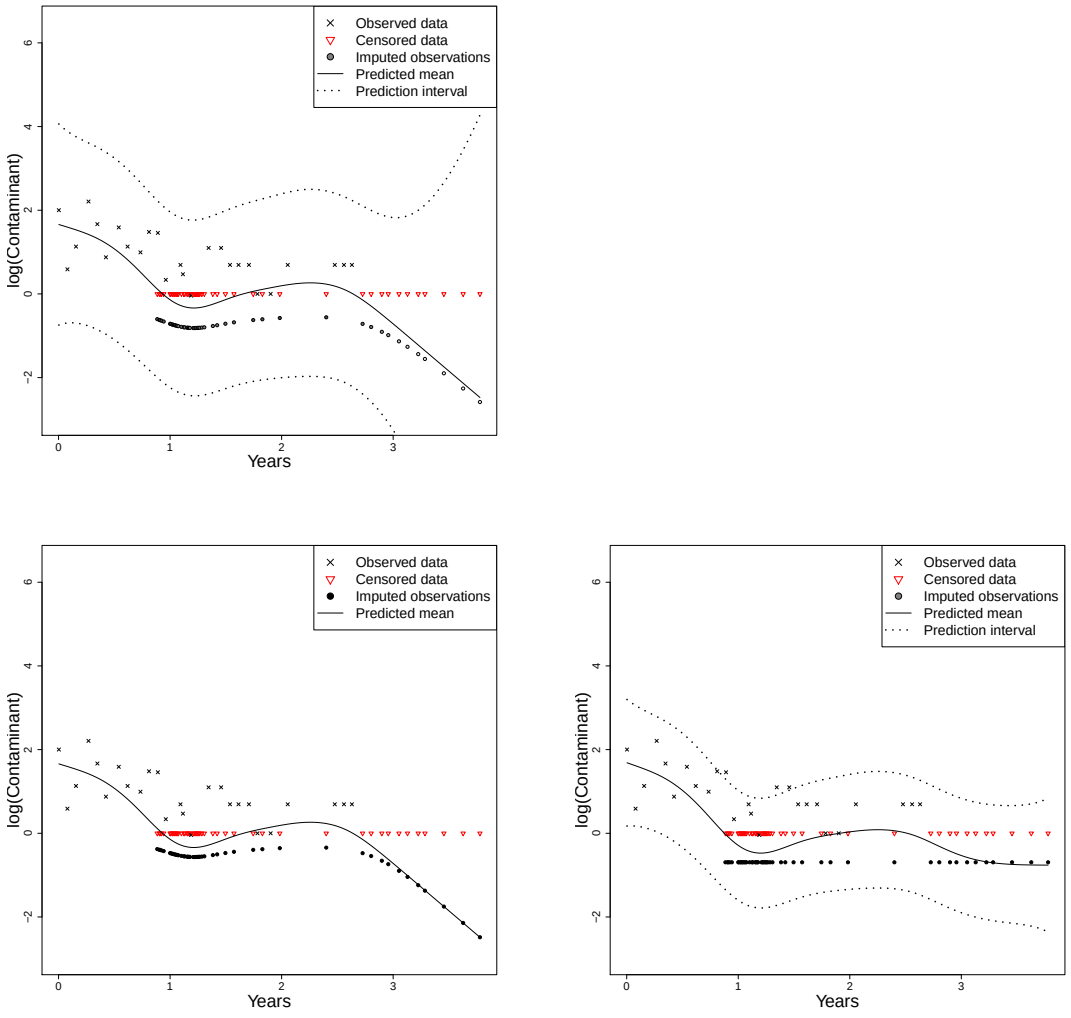


FIGURE 6.21: Predicted mean function and 95% prediction intervals for the contamination data obtained using the Laplace-type approximation (top left), by replacing non-detects by 1/2 the detection-limit (bottom right) and predicted mean function using EM-algorithm (bottom left). In all cases, the penalisation parameter λ corresponds to the MAP

Figure 6.22 shows the MCMC solution for the same fixed value of the penalisation parameter computed for the Laplace-type approximation, i.e. $\lambda = \lambda_{MAP} = 8.21$ (see Figure 6.20, top-right). It can be seen that, except at the very end of the series, the Laplace-type approximation is close to the fixed Bayesian solution. Using half the detection limit underestimates the uncertainty and yields too narrow prediction bands.

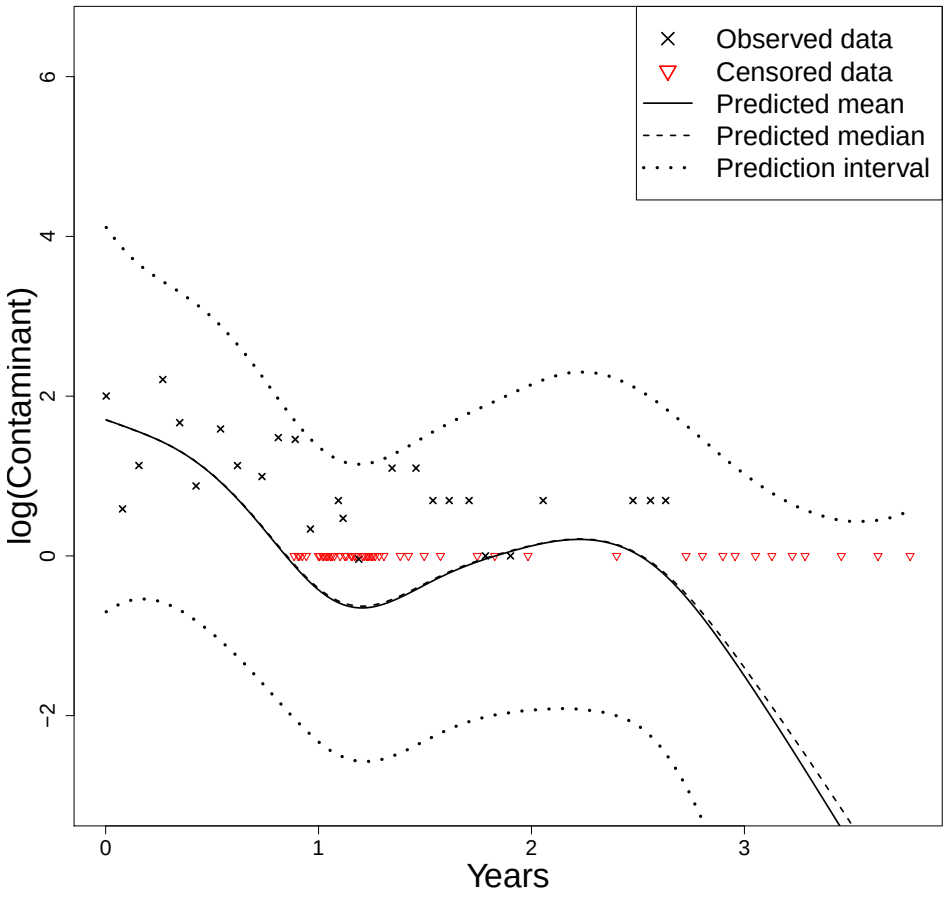


FIGURE 6.22: Predicted mean function and 95% prediction intervals for the contamination data obtained using MCMC with fixed value of λ

We see that the Laplace-type approximation yields symmetric (and hence unrealistic) confidence intervals. Furthermore, these intervals become very large on the ending extreme because here is where most of the non-detects are typically located.

An additional pitfall is that most of the uncertainty should be located on the upper bound side. Besides, the use of a log scale on the observations yields extremely high upper bounds in the natural scale making them implausible.

6.14 Case Study Revisited

We analyse again our case study from section 5.2 by dealing with non-detects using the Laplace-type approximation described in this chapter.

Figure 6.23 uses the standard assumptions whereas Figure 6.27 corresponds to the relaxed ones. Figures 6.24, 6.25, 6.26 and 6.28, 6.29, 6.30 picture the 95% lower and upper confidence intervals and standard errors for the predictions in both cases. In practice, the estimation of the coefficients $\hat{\boldsymbol{\alpha}}$ was carried out using equation (6.11), which in turn implied the computation of $\boldsymbol{v}^c$ (equation (6.6)) and $\boldsymbol{W}^c$ (equation (6.7)) required for evaluating $\boldsymbol{r}$ and $\boldsymbol{W}$ (recall equations (6.9)).

In the case of the standard assumptions, when the argument t_i (equation (6.5)) approached $-\infty$, the evaluation of $\boldsymbol{v}^c$ and $\boldsymbol{W}^c$ gave rise to indeterminate forms $\frac{0}{0}$, even if the computations were carried out using logarithms. Finally, this problem was overcome by repeated application of the rule of L'Hôpital when the argument approached extremely negative values.

Another pitfall that arose several times in the case of the standard assumptions, was that the inverse of $\boldsymbol{P} = \boldsymbol{B}'\boldsymbol{W}\boldsymbol{B} + \boldsymbol{V}^{-1}$ to be used in equation (6.11), could not be evaluated straightforward because the matrix $\boldsymbol{P}$ was ill-conditioned. Fortunately this issue could be also surmounted using a QR decomposition.

Using the relaxed assumptions, the algorithm converged for the optimal Bayesian MAP value for the smoothing parameter (λ=3.792e-3) without any computational issues.

The comparison of Figure 6.23 with Figure 6.27 reflects again that the undesired effect of ballooning when the standard assumptions are used, can be overcome with the use of relaxed ones. This is also confirmed by noticing that the eigenvalues yielding the maximal variance are 18294798.00 and 121.94 respectively. As expected, in the case of relaxed assumptions the upper confidence bound is low where non-detects are situated.

In addition, if we compare Figures 5.17 and 6.27, we see that the latter tends to predict lower levels of concentrations where observations with non-detects fall outside the area dominated by observed high values (in all cases, non-detects are represented by triangles). This is a sensible improvement due to our Laplace-type approximation using the relaxed assumptions framework as they manage to deal properly with observations for which it is only known to be below the value depicted in the wells.

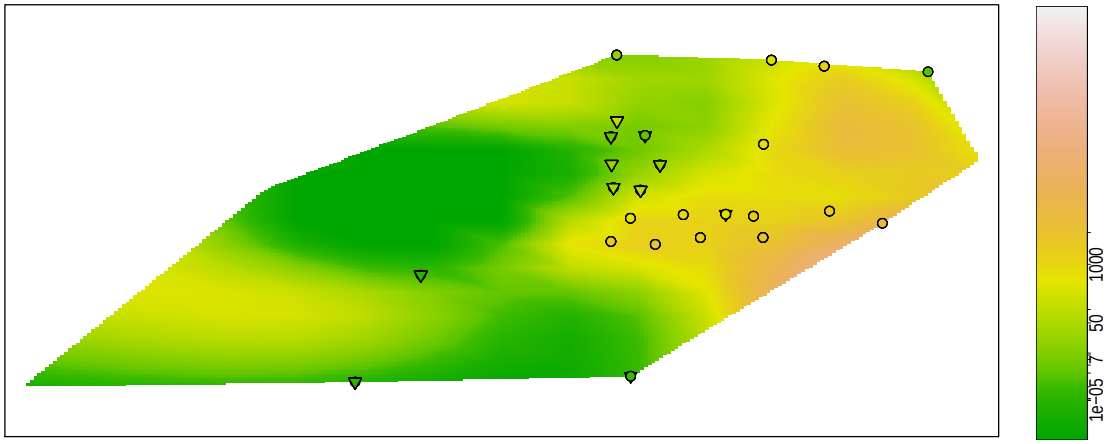


FIGURE 6.23: Predictions obtained for the real case study at time t=16.44 using the Laplace-type approximation under the standard assumptions. The penalisation parameter λ=9.123e-4 was computed using the Bayesian MAP criterion (triangles represent non-detects and circles correspond to observed data)

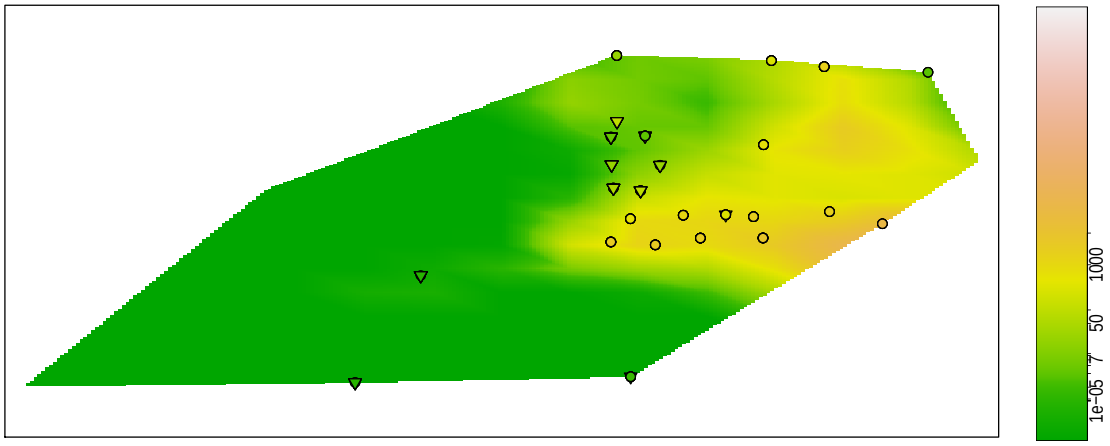


FIGURE 6.24: Lower 95% confidence limit for the predictions in Figure 6.23

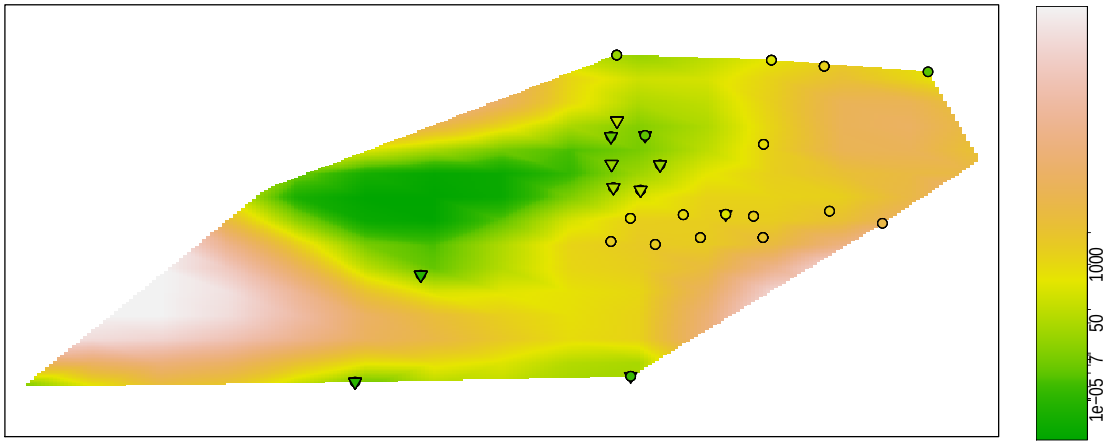


FIGURE 6.25: Upper 95% confidence limit for the predictions in Figure 6.23

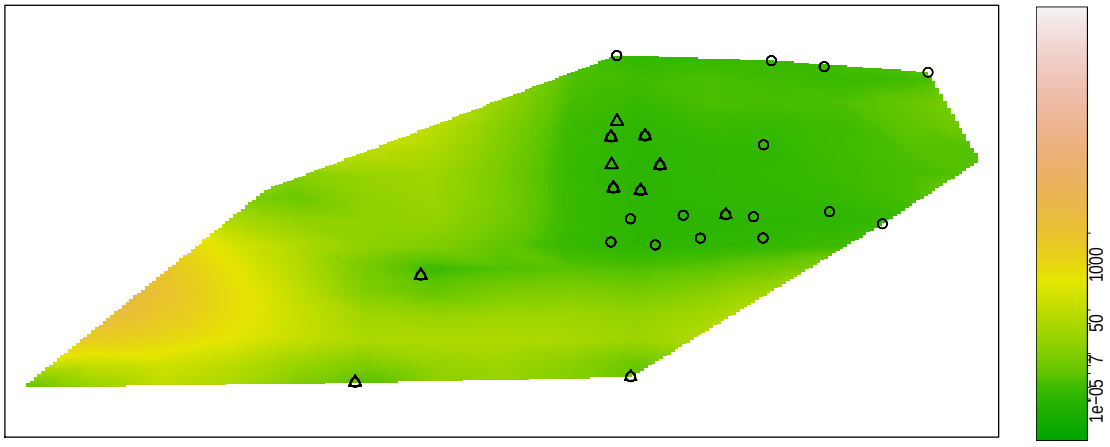


FIGURE 6.26: Standard errors for the predictions in Figure 6.23

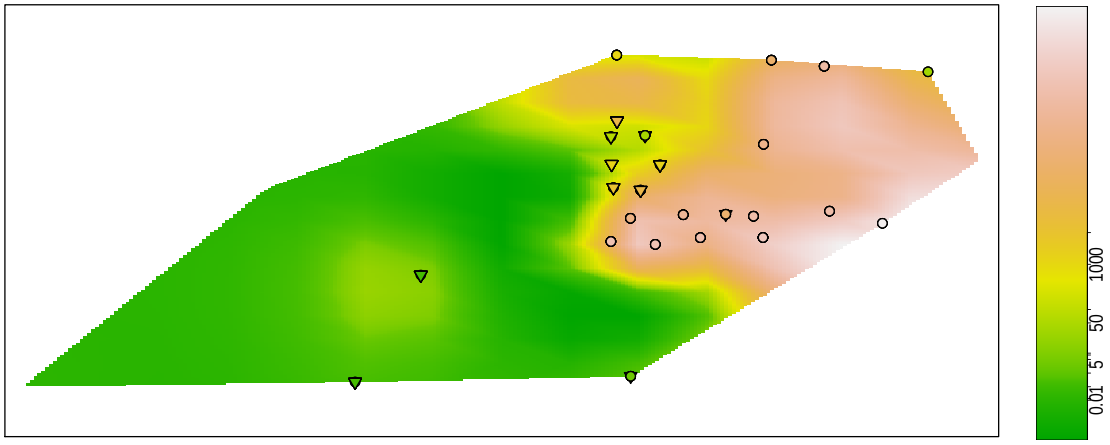


FIGURE 6.27: Predictions obtained for the real case study at time t=16.44 using the Laplace-type approximation under the relaxed assumptions. The penalisation parameter λ=3.792e-3 was computed using the Bayesian MAP criterion (triangles represent non-detects and circles correspond to observed data)

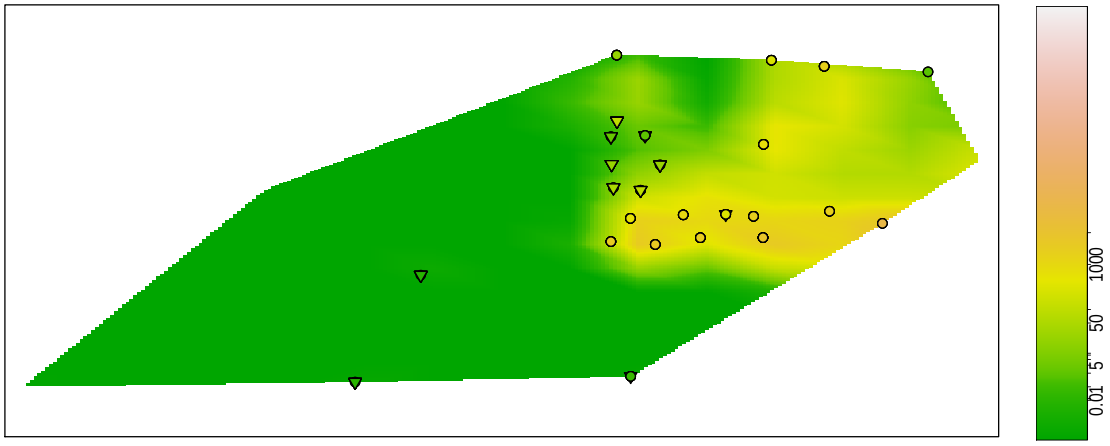


FIGURE 6.28: Lower 95% confidence limit for the predictions in Figure 6.27

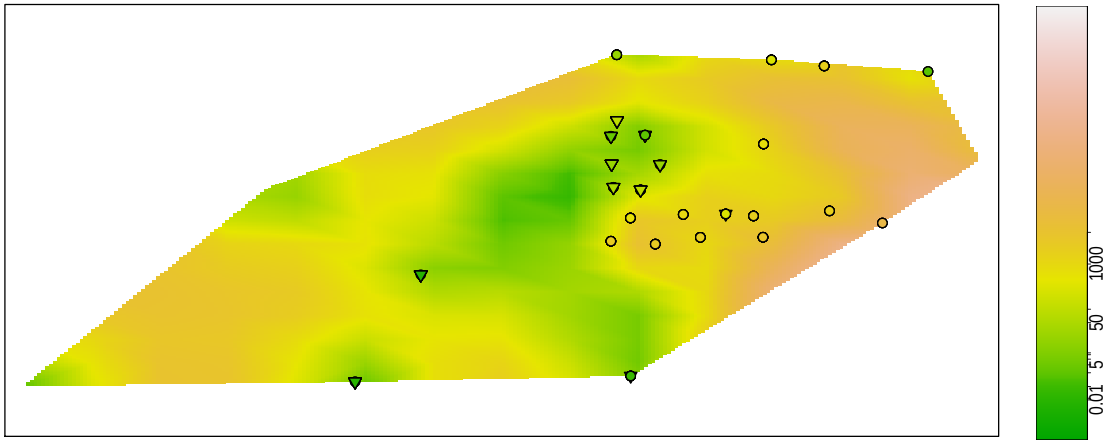


FIGURE 6.29: Upper 95% confidence limit for the predictions in Figure 6.27

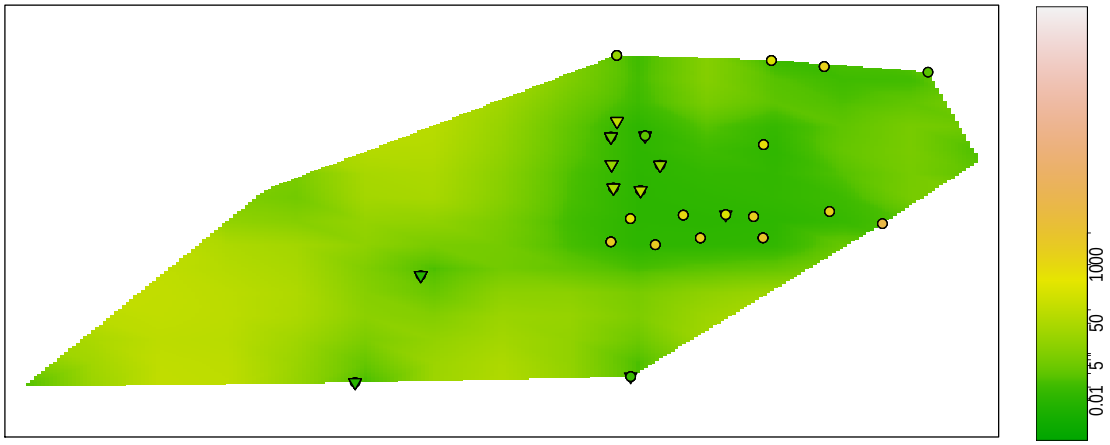


FIGURE 6.30: Standard errors for the predictions in Figure 6.27

Chapter 7

Discussion

7.1 Discussion

This thesis has focussed on the modelling of spatiotemporal data under two major conditions, namely that the model has to be fitted in an unsupervised setting and that it has to be fitted very quickly. The first of these conditions requires strong stability, in particular of the choice of penalty parameter which controls the degree of smoothing applied to the data. The second condition requires careful attention to the computational aspects of the model to avoid lengthy delay in the production of the results.

A standard spatial approach, also available to some extent in the spatiotemporal setting, would be to use kriging. However, this has significant disadvantages. One is in the assumption of a particular structure of separability usually made on the covariance matrix, while another is the requirement for the inversion of an $N \times N$ matrix (where N denotes the number of observations in the data set). Due to the fact that some of the actual data sets Shell has to deal with are very large, a P-splines model addresses this problem by providing a 'low rank' solution where the number of parameters in the model can be controlled. The tensor-product basis matrices required for three-dimensional (space-time) covariates remain large but lie within the scope of modern matrix inversion methods.

The assessment of the model proposed, carried out in chapters 3, 4 and 5, shows that the conditions of stability and speed have both been met successfully. Indeed, Shell have now rolled out software which provide these methods for use by consultants around the world, in evaluating sampled spatiotemporal groundwater data.

There are several new directions which could be taken for further research. There are clear potential advantages in allowing a more flexible model with at least two smoothing parameters: one for space and another for time. This would allow different degrees of smoothness for the evolution of pollution patterns over space and time. This issue has been addressed in the thesis through adjustment of the number of basis functions over space and time, but the availability of further adjustment of the penalty parameters would be welcome. This could be achieved through the use of computationally intensive methods such as MCMC but the time involved in this approach makes it infeasible for the present context.

On the specific issue of ballooning, an interesting avenue to explore is the use of random effects to describe the particular characteristics of the wells. The repeated measurements over time at each well would allow a model of this type to be considered and it may have the advantage of ameliorating the tendency for local differences to generate unjustifiably high predictions in sparsely sampled regions.

In the non-detects issue, the Laplace-type approximation proposed gives very accurate results, conditional on the smoothing parameter. The approximation works well for a modest proportion of non-detects. However, under uncertainty on the penalisation parameter, and as the proportion of non-detects increases, these posterior distributions become more skewed. Although the Laplace-type approximation is successful in identifying the mode of the posterior distribution of the parameters, it fails to capture their shape properly. This undesirable effect can be lessened, but not quite avoided, by using Bayesian model averaging rather than the Bayesian MAP approach for model selection.

The underlying issue is that the Laplace-type approximation yields symmetric confidence intervals while the actual distribution is asymmetric. Furthermore, these intervals can become very large at the edges of the sample space, where

most of the non-detects are typically located. Another issue is that principal attention should be focussed on the upper (rather than lower) bound. The use of a log scale on the observations yields extremely high upper bounds in the natural scale making them implausible.

One approach to model fitting might be to use a generalised linear model, for example with a gamma distribution for the response, although the iterative nature of the fitting process will generally cause computational problems, in settings where flexible modelling of the covariates is required. So these issues identify a very interesting line of future research, namely to construct asymmetric confidence intervals for a non-negative regression function in the presence of censored observations.

In terms of the application context, Shell may be interested in extending these models to measurements on multiple substances, to take advantage of the pooling of information across related variables. For example, the recorded information on a particular solute may help to predict the values of a different contaminant with similar molecular weight for which there is not enough data available in the same data set.

A final point of interest is the optimisation of the network design. This involves detecting wells that might be removed from the network because they produce information which is either redundant or of little value. The identification of optimal locations for new wells is a further very interesting aspect of this issue.

Appendix A

Brief summary on (semi) positive definite matrices

This appendix briefly recaps some definitions and results on (semi) positive definite matrices which mimic the behaviour of non-negative real numbers (see e.g. Sheldon, 1997; Meenakshi and Rajian, 1999)

Given matrices $\boldsymbol{A}$ and $\boldsymbol{B} \in \mathbb{R}^{m \times m}$, and $\boldsymbol{C} \in \mathbb{R}^{m \times n}$

1. $\boldsymbol{A}$ is said to be **semi-positive definite** if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0 \ \forall \boldsymbol{x} \in \mathbb{R}^m, \ \boldsymbol{x} \neq 0$
2. $\boldsymbol{A}$ is said to be **(strictly) positive definite** if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0 \ \forall \boldsymbol{x} \in \mathbb{R}^m, \ \boldsymbol{x} \neq 0$
3. As a corollary of the previous definitions, if $\boldsymbol{A}$ and $\boldsymbol{B}$ are (semi) positive definite so is $\boldsymbol{A} + \boldsymbol{B}$
4. Also, if $\boldsymbol{A}$ is positive definite and $\boldsymbol{B}$ is semi-positive definite then $\boldsymbol{A} + \boldsymbol{B}$ is positive definite
5. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric and semi-positive definite then $\boldsymbol{AB}$ is semi-positive definite if and only if $\boldsymbol{AB}$ is also symmetric
6. If $\text{rank}(\boldsymbol{A}) < m$ then $\boldsymbol{A}'\boldsymbol{A}$ is semi-positive definite (and symmetric)
7. If $\text{rank}(\boldsymbol{A}) = m$ then $\boldsymbol{A}'\boldsymbol{A}$ is positive definite (and symmetric)
8. If $m < n$ then $\boldsymbol{C}'\boldsymbol{C}$ is semi-positive definite (and symmetric)

9. If $m = n$ then but $\text{rank}(\boldsymbol{C}) < m$ then $\boldsymbol{C}'\boldsymbol{C}$ is semi-positive definite (and symmetric)

10. If $m = \text{rank}(\boldsymbol{C}) = n$ then $\boldsymbol{C}'\boldsymbol{C}$ is positive definite (and symmetric)

11. **Theorem of Spectral Decomposition**: If $\boldsymbol{A}$ is semi-positive definite then there exists $\boldsymbol{P} \in \mathbb{R}^{m\times m}$ orthogonal (i.e. $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_m$ or $\boldsymbol{P}^{-1} = \boldsymbol{P}'$) such that $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}' = \boldsymbol{\Delta} = \text{diag}(\delta_1, \ldots, \delta_m)$ with $\delta_i \in \mathbb{R},\ \delta_i \geq 0\ \forall i = 1, \ldots, m$. The δ_i's making up the diagonal of the matrix $\boldsymbol{\Delta}$ are called the *eigenvalues* of the matrix $\boldsymbol{A}$.

12. If $\boldsymbol{A}$ is positive definite, then the previous theorem holds with $\delta_i \in \mathbb{R},\ \delta_i > 0\ \forall i = 1, \ldots, m$

13. As a corollary of the definition of orthogonal matrices, we have that if $\boldsymbol{P} \in \mathbb{R}^{m\times m}$ is orthogonal then $|\boldsymbol{P}|^2 = 1$

14. If $\boldsymbol{A}$ is semi-positive definite then $\text{rank}(\boldsymbol{A})$ equals the number of strictly positive eigenvalues δ_i in the spectral decomposition

Appendix B

Execution times for the computation of the optimal penalisation parameter λ

Number of candidates for λ	Naïve computation	Efficient computation
1	19.25	28.96
2	23.79	28.97
3	28.19	29.01
4	32.84	29.05
5	37.38	29.06
6	42.00	29.05
7	46.37	29.08
8	51.16	29.11
9	55.30	29.16
10	59.98	29.14
11	64.41	29.21
12	68.93	29.21
13	73.64	29.25
14	78.18	29.22
15	82.76	29.28
16	87.53	29.31
17	91.50	29.33
18	96.16	29.31
19	100.95	29.39
20	105.52	29.39
21	109.98	29.44
22	114.42	29.41
23	118.93	29.47
24	123.36	29.51
25	128.56	29.54
26	132.67	29.53
27	137.18	29.61
28	141.29	29.53
29	145.75	29.66
30	150.40	29.64

TABLE B.1: Total execution times (in seconds) for computing the posterior densities for different numbers of candidates for the penalisation parameter λ (These data correspond to Figure 3.13)

Dimension of $\hat{\alpha}$	Naïve computation	Efficient computation
216	12.77	0.92
343	26.22	2.31
512	47.31	5.80
729	83.79	13.79
1000	150.95	29.54
1331	267.83	61.52
1728	469.68	120.16

TABLE B.2: Total execution times (in seconds) for computing the posterior densities of 30 values of λ, for different dimensions of $\hat{\alpha}$ (These data correspond to Figure 3.14)

Appendix C

Model and data used in Figure 6.4

The model used in the example to compare the Laplace-type approximation with the standard approach is $Y = 0.2X + 0.1X^2 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$ where σ_0 is assumed to be known and equal to 2. The detection limit was set to 4 and the seed used for random numbers generation was equal to 77. The estimated coefficients are $\hat{\boldsymbol{\alpha}} = (0.40, 0.24, 0.11)$.

The following table summarizes the results of the fitting process.

Obs. Number	x	True y	Observed y	Fitted y	Imputed Values	t	Weight
1	-10	8.0	6.9007	8.6952			1.0000
2	-9	6.3	8.4821	6.9052			1.0000
3	-8	4.8	6.0796	5.3286			1.0000
4	-7	3.5	5.5852	3.9656			1.0000
5	-6	2.4	2.7394	2.8161	0.9204	0.5920	0.4885
6	-5	1.5	3.7756	1.8801	0.3717	1.0599	0.3526
7	-4	0.8	-1.1411	1.1577	-0.1092	1.4212	0.2487
8	-3	0.3	0.0363	0.6487	-0.4759	1.6756	0.1829
9	-2	0.0	0.2925	0.3533	-0.6984	1.8234	0.1490
10	-1	-0.1	2.7826	0.2714	-0.7613	1.8643	0.1403
11	0	0.0	-5.8828	0.4030	-0.6605	1.7985	0.1544
12	1	0.3	-0.1857	0.7481	-0.4025	1.6259	0.1950
13	2	0.8	0.5188	1.3068	-0.0060	1.3466	0.2695
14	3	1.5	1.4347	2.0789	0.4954	0.9605	0.3820
15	4	2.4	2.9596	3.0646	1.0517	0.4677	0.5225
16	5	3.5	4.6803	4.2638			1.0000
17	6	4.8	6.8486	5.6765			1.0000
18	7	6.3	10.5146	7.3028			1.0000
19	8	8.0	8.3093	9.1425			1.0000
20	9	9.9	11.7261	11.1958			1.0000
21	10	12.0	11.4916	13.4626			1.0000

TABLE C.1: Data corresponding to Figure 6.4

Bibliography

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. *Second International Symposium on Information Theory, Akademia Kiado*, 267–281.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data*. USA: Chapman & Hall/CRC.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. New York, USA: Springer-Verlag.

Bowman, A., E. Crawford, G. Alexander, and R. Bowman (2007). rpanel: Simple interactive controls for R functions using the tcltk package. *Journal of Statistical Software 17–9*, 1–18.

Bowman, A. W., L. Evers, D. A. Molinari, W. R. Jones, and M. J. Spence (2013). Efficient and automatic methods for flexible regression on spatiotemporal data, with applications to groundwater monitoring. *Environmetrics*.

Bowman, A. W., M. Giannitrapani, and E. M. Scott (2009). Spatiotemporal models for sulphur dioxide pollution over Europe. *Journal of the Royal Statistical Society: Series C-Applied Statistics 58*, 737–752.

Box, G. and G. Tiao (1992). *Bayesian Inference in Statistical Analysis*. USA: John Wiley & Sons, Ltd.

Brezger, A. and S. Lang (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis 50–4*, 967–991.

Brezger, A. and S. Lang (2008). Simultaneous probability statements for Bayesian P-splines. *Statistical Modelling 8–2*, 141–168.

Chib, S. (1992). Bayes inference in the Tobit censored regression model. *Journal of Econometrics 51, North-Holland*, 79–99.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, New Jersey, USA: Princeton University Press.

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

Cressie, N. and C. K. Wikle (2011). *Statistics for Spatio-Temporal Data*. New York, USA: Wiley.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of Royal Statistical Society 39–1, Series B (Methodological)*, 1–38.

Denison, D. G. T., C. C. Holmes, B. K. Mallick, and A. F. M. Smith (2002). *Bayesian Methods for Nonlinear Classification and Regression*. West Sussex, England, UK: John Wiley & Sons, Ltd.

Diggle, P. and P. Ribeiro (2007). *Model-based Geostatistics*. Springer.

Durbán, M. (2009). An introduction to smoothing with penalties: P-splines. *Boletín de Estadística e Investigación Operativa 25–3*, 195–205.

Eilers, P. H. C. and B. D. Marx (1992). Generalized Linear Models with P-splines. *Advances in GLIM and Statistical Modelling - L.Fahrmeir et al. (eds.) - Springer-Verlag New York, Inc.*.

Eilers, P. H. C. and B. D. Marx (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science 11–2*, 89–102.

Eilers, P. H. C. and B. D. Marx (2010). Splines, knots and penalties. *Computational Statistics 2–6*, 637–653.

Eldén, L. (1977). Algorithms for the Regularization of Ill-Conditioned Least Squares Problems. *BIT Numerical Mathematics 17–2*, 134–145.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica 14–3*, 731–761.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression - Models, Methods and Applications*. Berlin, Germany: Springer-Verlag.

Finkenstädt, B., L. Held, and V. Isham (Eds.) (2007). *Statistical Methods for Spatio-Temporal Systems*. London, UK: Chapman & Hall/CRC.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1–3*, 385–650.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian Data Analysis* (third ed.). Chapman & Hall/CRC.

Gentle, J. E. (2002). *Elements of Computational Statistics*. New York, USA: Springer-Verlag New York, Inc.

Gentle, J. E. (2007). *Matrix Algebra - Theory, Computation and Applications in Statistics*. New York, USA: Springer-Verlag New York, Inc.

Girolami, M. and S. Rogers (2006). Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. *Neural Computation 18–8*, 1790–1817.

Golub, G. and C. Van Loan (1996). *Matrix Computations*. Baltimore, USA: The Johns Hopkins University Press.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning - Data Mining, Inference and Prediction* (second ed.). New York, USA: Springer.

Hauck, W. and A. Donner (1977). Wald's Test as Applied to Hypothesis in Logit Analysis. *Journal of the American Statistical Association 72–360*, 851–853.

Helsel, D. R. (2005). *Nondetects and Data Analysis: Statistics for Censored Environmental Data*. New York, USA: John Wiley.

Helsel, D. R. (2006). Fabricating data: How substituting values for non-detects can ruin results, and what can be done about it. *Chemosphere 65*, 2434–2439.

Helsel, D. R. (2012). *Statistics for Censored Environmental Data Using Minitab and R* (second ed.). Hoboken, New Jersey, USA: John Wiley & Sons, Inc.

Helsel, D. R. and R. M. Hirsch (2002). Statistical Methods in Water Resources. *U.S. Geological Survey Techniques of Water Resources Investigationns, Book 4, Chapter A3*.

Hurvich, C. M., J. S. Simonoff, and C. L. Tsai (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.

Husmeier, D. (2000). The bayesian evidence scheme for regularizing probability-density estimating neural networks. *Neural Computation 12*, 2685–2717.

Johnston, J. and J. diNardo (1997). *Econometric Methods* (fourth ed.). New York, USA: McGraw-Hill.

Kotz, S. and S. Nadarajah (2004). *Multivariate t-Distributions and their Applications*. Cambridge, UK: Cambridge University Press.

Lang, S. and A. Brezger (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics 13–1*, 183–212.

Lee, D. J. and M. Durbán (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statistical Modelling 11–1*, 49–69.

Lemieux, C. (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. New York, USA: Springer Science+Business Media.

McLachlan, G. J. and T. Krishnan (2008). *The EM algorithm and Extensions* (second ed.). Hoboken, New Jersey, USA: John Wiley & Sons, Inc.

Meenakshi, A. and C. Rajian (1999). On a product of positive semidefinite matrices. *Linear Algebra and its Applications 295*, 3–6.

Morrisey, E., M. Juarez, K. Denby, and N. Burroughs (2011). Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression. *Biostatistics 12–4*, 682–694.

O'Sullivan, F. (1986). A Statistical Perspective on Ill-Posed Inverse Problems. *Statistical Science 1–4*, 502–518.

Raftery, A., D. Madigan, and J. Hoeting (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association 92–437*, 179–191.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: The MIT Press.

Rényi, A. (2007). *Probability Theory*. Mineola, New York, USA: Dover Publications, Inc.

Ruppert, D., M. P. Wand, and R. Carroll (2003). *Semiparametric Regression*. London, UK: Cambridge University Press.

Schölkopf, B. and A. J. Smola (2002). *Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond*. Massachusetts, USA: The MIT Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6–2*, 461–464.

Seeger, M. (2000). Bayesian Model Selection for Support Vector Machines, Gaussian Processes and other Kernel Classifiers. *Proceedings of the 13th Annual Conference on Neural Information Processing Systems*, 603–609.

Sheldon, A. (1997). *Linear Algebra, Done Right* (second ed.). New York, USA: Springer.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York, USA: Chapman & Hall/CRC.

Simonoff, J. (1996). *Smoothing Methods in Statistics*. New York, USA: Springer.

Smola, A. J. and B. Schölkopf (2004). A Tutorial on Support Vector Regression. *Statistics and Computing 14*, 199–222.

Tanner, M. A. (1996). *Tools for Statistical Inference* (third ed.). New York, USA: Springer-Verlag New York, Inc.

Wood, S. (2006). *Generalized Additive Models: an introduction with R*. London, UK: Chapman and Hall/CRC.

Wood, S. N. (2000). Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62–2*, 413–428.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73–1*, 3–36.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B-Statistical Methodology 67 – Issue 2*, 301–320.