Changepoint analysis using Gaussian process regression: A Bayesian statistical model for the identification of lithological strata in geotechnical engineering

A thesis submitted by

Jens Bendel

— to the —

Department of Mathematics at Imperial College London

in partial fulfilment of the requirements for the
degrees of Master of Philosophy of Imperial College London and
Diploma of Imperial College London

Abstract

Around the Earth's surface it is common for the ground to be made up of layers of different soil stacked on top of each other. The field of geotechnical engineering studies these lithological layers and their effect on engineering work such as tunnelling or the construction of buildings and bridges. A standard procedure to determine a segmentation of the ground into individual layers is to take a drilling core and examine slices sampled at various depths. Typical methods to distinguish soil layers are based on microfossil content or measurements of physical quantities such as the water content of the soil. Common practice is the inspection of these properties by eye using expert engineering judgement. Such approaches lack scientific rigour and fail to address the uncertainty that is inherent to any such analysis.

This thesis discusses statistical methodology (changepoint analysis) in order to propose a reproducible scientific approach for the identification of soil layers based on measurements of water content. With a focus on uncertainty quantification the proposed approach combines a Bayesian changepoint method with a Gaussian process regression model for each soil layer. This Gaussian process changepoint method is applied to data from the construction site of underground railway tunnels. The method correctly identifies well-established layers while suggesting that a previously proposed subdivision of a particular layer (A3) is not supported by the dataset. Further results indicate that more research work is needed with regard to the collection of data as well as the development of statistical methodology. Overall, this thesis shows that the proposed focus on mathematical rigour and uncertainty quantification is very much needed.

Declarations

Originality

This thesis is the product of my own work. Work of other people is fully acknowledged in accordance with the standard referencing practices of the discipline.

Copyright

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Around the Earth's surface it is common for the ground to be made up of layers of different soil stacked on top of each other. This composition of layers can for example be the result of marine deposition over the course of thousands or millions of years. The field of geotechnical engineering investigates these lithological layers and their effect on engineering work such as tunnelling or the construction of buildings and bridges. A standard procedure to partition the ground into individual layers is to take a drilling core and examine slices sampled at various depths. Commonly, individual layers of soil are not visible to the naked eye but a partition into layers can, for instance, be based on microfossil content or physical quantities such as the water content of the soil. Different approaches to partition the ground into individual layers can generally lead to different results. This motivates a reproducible scientific approach such as the statistical method presented in this thesis.

In Section 1.1 we proceed by introducing the example of the London Clay Foundation, which will be the main example throughout this thesis. Based on this example we then continue the discussion of the segmentation of soil into layers based on water content. The presentation of geotechnical background is based on Hight et al. (2007).

Section 1.2 highlights the contributions of this thesis and gives a brief overview of the entire dissertation.

1.1 The London Clay Foundation and water content profiles

Because of the relevance for engineering works, the soil structure in a metropolis such as London is particularly well-studied (Hight et al., 2003, 2007). The formation of soil strata found there is known as the *London Clay Foundation* (LCF) and stretches across the London basin, a larger region in the south-east of England. The current geotechnical understanding of the LCF goes back to the work of King (1981) who determined a partitioning into five divisions (A to E) using biostratigraphy, that is, based on microfossils present in the layers (Pantelidou and Simpson, 2007). These divisions are still used as benchmark (Pantelidou and Simpson, 2007; Wang et al., 2014a) even today.

In the LCF the soil consists of sands, clays and silts of different coarseness, higher clay content generally meaning higher water content. More recent research (Hight et al., 2003, 2007; Wang et al., 2014a) has shown that an analysis of the water content can lead to a segmentation similar to that established by King (1981). For such an analysis a sequence of water content measurements at various depths is taken. The resulting dataset is referred to as a *water content profile*. Water content profiles of the LCF have been found to be well described by a piecewise linear model (Wang et al., 2014a). Some of the transition points from one linear segment to the next coincide with the boundary layers between different geological strata identified by King (1981), while within each soil layer the water content is well described by a linear model.

A simple but important motivation to base the segmentation on water content measurements is the comparably low cost of this method: Measuring the water content is easier and cheaper than a microscopic analysis that needs to be carried out by an expert. Also the analysis of the measurements is easier and cheaper once an analysis procedure has been established.

Further power of basing a segmentation on water content lies in geotechnical engineering applications. In the construction of Terminal 5 at London Heathrow Airport (Hight et al., 2007), the information is needed to understand the stand-up time of temporary soil slopes before failures such as land slides occur. In other contexts the information is used to understand the effects of tunnelling on the environment, for example in the construction of the Jubilee line of the London Underground network (Standing and Burland, 2006). For these geotechnical engineering applications the effects of a property like water content are likely of higher relevance than those of microfossil content. In this context Hight et al. (2007) express particular interest in potential identification of previously unidentified sub-layers. They use the example of an observed land slide in which one soil sub-layer persisted – "presumably because of its larger sand content" visible in the water content profile.

We conclude that segmenting soil based on water content is a highly relevant field of research for various applications in geotechnical engineering. Section 1.2 presents the contributions of this thesis to the field.

1.2 Contributions and structure of this thesis

Section 1.1 highlighted the potential of investigating soil strata based on their water content. First attempts of such analyses were based on engineering judgement. Personal communication with Dr. Standing[1] suggests that this usually means an analysis by eye, sometimes taking previous knowledge into account; in particular the benchmark by King (1981). Wang et al. (2014a) propose a mathematically rigorous approach

- to address the subjectivity that comes with the analysis based on the judgement of an engineer,
- to formally include uncertainty such as measurement error that goes into this judgement,
- and to hence establish a scientific method to perform the analysis.

This thesis presents such an approach. The presented method differs from that of Wang et al. (2014a) as it quantifies uncertainty using a state-of-the-art Bayesian approach. The presented analysis is transparent and reproducible. The results show that, in fact, there is a significant amount of ambiguity about possible segmentations and that the standard presented in the literature is far from definite. The results from the proposed Bayesian method partially coincide with those of King (1981) but generally suggest

[1] The project presented in this thesis is a collaboration with Dr. Jamie Standing, who is a geotechnical engineer in the Department of Civil and Environmental Engineering at Imperial College London.

a different segmentation. Some boundary layers suggested by King are not identified while the Bayesian method identifies further (sub)layers not suggested by King. These results motivate further research work to investigate the possible new layers. The presented method is readily applicable to other datasets to identify segments that show statistically *coherent* patterns. For the piecewise linear model of the water content data coherent means that within one segment the data is scattered homogeneously around a line. The transition points from one coherent segment of data to the next are not known to us in advance. We refer to those as *changepoints* and they are in fact the centre of interest in our analysis as they indicate the transition from one soil layer to the next according to our model. The pertinent statistical methodology, *changepoint analysis*, is reviewed in Section 3.

The rest of this thesis is laid out in the following way. Section 2 sets the framework for concepts and notation used in the rest of this thesis. Sections 3 and 4 review the statistical literature and the building blocks that will be used to construct the statistical model which, in Section 5, is applied to a water content profile.

Chapter 2

Statistical framework

The purpose of this section is to set the statistical framework for this thesis. We introduce a simple linear regression model and give a generally phrased introduction to Bayesian statistics. We then demonstrate the ideas by discussing how to turn the linear regression model into a Bayesian linear regression model. This facilitates the introduction of advanced statistical methodology in Sections 3 and 4.

2.1 Notation

Consider a dataset of n measurements $y_1, y_2, \ldots, y_n$ that are associated with *inputs* $x_1, x_2, \ldots, x_n$. Throughout this thesis a core sample of soil will serve as example: The measurement y_j denotes the water content of the soil measured at a depth of x_j. While spatial variables such as depth or distance might be the most common applications in the geotechnical sciences we may also think of y_j to be a stock price or a temperature at a certain point in time x_j.

2.2 Regression analysis

A common goal of a statistical analysis is to investigate the relationship between the measurements and the corresponding inputs. Regression analysis assumes that there exists a function $f(\cdot)$ such that the measurement y_j is related to its input x_j via

$$y_j = f(x_j) + \eta_j. \tag{2.1}$$

Deviation from the value $f(x_j)$ is accounted for by adding η_j, a zero-mean random variable that models measurement noise. Throughout most of this thesis we make the common assumption that $\eta_1, \eta_2, \ldots, \eta_n$ are independent and follow a normal distribution with mean zero and noise variance σ^2.[1]

A simple but powerful choice for $f(\cdot)$ is that of a line. A linear regression model assumes

$$f(x_j) = \beta_0 + \beta_1 x_j \tag{2.2}$$

for all j and the goal of a linear regression analysis is to infer parameter values β_0, β_1 that provide the best *fit* of the model to the data. Of course one might find that the assumption of a linear relationship cannot be justified and the linear model should then be rejected.

Let $\theta = (\beta_0, \beta_1, \sigma^2)$ denote the vector of all parameters. It will be useful to state models in terms of the likelihood $p(y|\theta)$ instead of the representation via random variables in equation (2.1). The likelihood is the probability density function (PDF) of the measurement given all parameters. For a single measurement y with input x and the example of the linear regression model with normally distributed noise this means

$$p(y|\theta) = N(y; \beta_0 + \beta_1 x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[y - (\beta_0 + \beta_1 x)]^2}{2\sigma^2} \right).$$

We use $N(\mu, \sigma^2)$ to denote a normal distribution with mean μ and variance σ^2, and $N(y; \mu, \sigma^2)$ to denote its PDF evaluated at y. The *joint likelihood* of all independent measurements $y_1, y_2, \ldots, y_n$ is given by the product of the individual likelihoods,

$$p(y_1, y_2, \ldots, y_n|\theta) = \prod_{j=1}^{n} p(y_j|\theta) = \prod_{j=1}^{n} N(y_j; \beta_0 + \beta_1 x_j, \sigma^2).$$

[1] Examples for models with non-Gaussian noise are briefly discussed in Section 4.6.

2.3 Bayesian statistical inference

In this thesis we are interested in methods that use a dataset of observations $\mathcal{D}$ to infer about (the structure of) the mechanism that generated this dataset. For the main contents of this thesis it will be useful to have outlined some concepts and terminology of Bayesian statistics.

2.3.1 Parametric Bayesian statistics

We start from a parametric statistical model $\mathcal{M}$ phrased in terms of a probability density function (PDF), $p(\mathcal{D}|\theta)$, where θ denotes the unknown vector of parameters that determine the distribution. To turn this statistical model into a Bayesian statistical model we consider θ as a random variable and define the *prior distribution* of the parameter. Let $p(\theta)$ denote the PDF of the prior distribution. We will refer to both, the prior distribution and its PDF $p(\theta)$, as the *prior*. This terminology originates from the idea that $p(\theta)$ describes the distribution of the parameter prior to taking the data into account. Once the data is observed Bayes' formula links the *Bayesian model* (likelihood and prior) to the *posterior distribution of* θ given the data $\mathcal{D}$:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}. \tag{2.3}$$

If the posterior is in the same family of probability distributions as the prior then the Bayesian model is referred to as a *conjugate* model. The denominator $p(\mathcal{D})$ plays a crucial role in Bayesian statistics. It is referred to as the *marginal likelihood* because it is obtained from the likelihood by marginalising over the prior distribution of the parameter θ,

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)\,\mathrm{d}\theta. \tag{2.4}$$

It is common for the prior distribution to depend on a vector of parameters itself. Denoting this vector by ψ we then write the prior distribution as $p(\theta|\psi)$ and refer to ψ as *hyperparameter* in order to highlight the different levels of hierarchy between ψ and the parameter θ. Bayes' formula can be used to summarise the model and clarify the hierarchy of the random quantities involved:

$$p(\theta|\mathcal{D},\psi) = \frac{p(\mathcal{D}|\theta)p(\theta|\psi)}{p(\mathcal{D}|\psi)}$$

with $p(\mathcal{D}|\psi) = \int p(\mathcal{D}|\theta)p(\theta|\psi)\,\mathrm{d}\theta$. If ψ is unknown, a fully Bayesian treatment requires the specification of a prior on ψ, $p(\psi)$, which is then referred to as hyperprior. With

$$p(\mathcal{D}) = \int p(\mathcal{D}|\psi)p(\psi)\,\mathrm{d}\psi \tag{2.5}$$

Bayes' formula yields

$$\begin{aligned} p(\theta,\psi|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta,\psi)p(\theta,\psi)}{p(\mathcal{D})} \\ &= \frac{p(\mathcal{D}|\theta)p(\theta|\psi)p(\psi)}{p(\mathcal{D})}. \end{aligned}$$

The first equality is simply equation (2.3) with parameters (θ,ψ) replaced by θ. The second equality emphasises the hierarchical structure of the model; in particular that the distribution of the data does not depend on ψ if θ is given: $p(\mathcal{D}|\theta,\psi) = p(\mathcal{D}|\theta)$. Of course, the procedure of hyperparametrisation can be continued, leading to hyperhyperparameters and hyperhyperpriors. But no details about such "three-stage" (Carlin et al., 1992) (or higher) levels of hierarchy will be needed in this thesis.

Bayesian analysis requires the evaluation of integrals, in particular the marginal likelihood in (2.4) or (2.5). For conjugate models these integrals are of closed form. In practice this is rarely the case and computational approximation methods are needed.

A simpler approximation that is commonly used at the level of hyperparameters is known as type II maximum likelihood (ML-II). Instead of computing its posterior distribution, ψ is estimated from the data by maximising the marginal likelihood,

$$\psi^* := \arg\max_{\psi} p(\mathcal{D}|\psi).$$

2.3.2 Model comparison

The Bayesian framework can be applied in situations where we wish to compare different statistical models $\mathcal{M}_1, \dots, \mathcal{M}_K$. Let $p(\mathcal{M}_j)$ denote the probabilities describing our prior belief about the individual models. The posterior probability of model $\mathcal{M}_j$ is given by Bayes' formula

$$p(\mathcal{M}_j|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_j)p(\mathcal{M}_j)}{p(\mathcal{D})}$$

where $p(\mathcal{D}|\mathcal{M}_j)$ is the probability of the data under the assumption of model $\mathcal{M}_j$ and $p(\mathcal{D}) = \sum_{j=1}^{K} p(\mathcal{D}|\mathcal{M}_j)p(\mathcal{M}_j)$. Two competing models $\mathcal{M}_j$ and $\mathcal{M}_k$ can be compared based on the *Bayes factor*

$$\frac{p(\mathcal{D}|\mathcal{M}_j)}{p(\mathcal{D}|\mathcal{M}_k)} = \frac{p(\mathcal{M}_j|\mathcal{D})}{p(\mathcal{M}_k|\mathcal{D})}\frac{p(\mathcal{M}_k)}{p(\mathcal{M}_j)},$$

which does not depend on the choice of the prior. The Bayes factor is identical to the ratio of the posterior probabilities in the case of equal prior probabilities.

Section 2.3.1 considers the Bayesian treatment of the parameters for one model and thus the dependence on the model is not included in the notation. Here we rewrite equation (2.5) including the model

$$p(\mathcal{D}|\mathcal{M}_j) = \int p(\mathcal{D}|\mathcal{M}_j, \psi)p(\psi|\mathcal{M}_j)\,\mathrm{d}\psi. \tag{2.6}$$

We can rewrite equation (2.4) similarly and hence refer to $p(\mathcal{D}|\mathcal{M}_j)$ as the marginal likelihood.

2.4 Example: Bayesian linear regression

For a Bayesian analysis of the linear regression model in equations (2.1) and (2.2) we assign a prior distribution $p(\beta|\psi)$ with hyperparameter ψ to the parameter $\beta = (\beta_0, \beta_1)^T$. If we assume that ψ and the noise variance σ^2 are known and we are only interested to infer about the intercept β_0 and the slope β_1 of the regression line, then the model is summarised by

$$p(\beta|y_{1:n}, \psi, \sigma^2) = \frac{p(y_{1:n}|\beta, \sigma^2)p(\beta|\psi)}{p(y_{1:n}|\psi, \sigma^2)}.$$

Denoting $\psi = (\sigma_0^2, \sigma_1^2)$ and defining a normal prior distribution

$$p(\beta_0, \beta_1 | \sigma_0^2, \sigma_1^2) = N(\beta_0; 0, \sigma_0^2) N(\beta_1; 0, \sigma_1^2)$$

leads to a normal posterior distribution and thus a conjugate model. The integral that defines the marginal likelihood $p(y_{1:n} | \sigma_0^2, \sigma_1^2, \sigma^2)$ is of closed form but rather than performing the calculations we note that using an independent Gaussian prior means that observation y_j is given as the linear combination of three Gaussian random variables,

$$y_j = \beta_0 + \beta_1 x_j + \eta_j,$$

and hence the observations $y_1, \ldots, y_n$ are also normally distributed with mean zero[2] and covariances

$$\begin{aligned} \mathrm{Cov}(y_j, y_k) &= \mathrm{Cov}(\beta_0 + \beta_1 x_j + \eta_j, \beta_0 + \beta_1 x_k + \eta_k) \\ &= \mathrm{Cov}(\beta_0, \beta_0) + x_j x_k \mathrm{Cov}(\beta_1, \beta_1) + \mathrm{Cov}(\eta_j, \eta_k) \\ &= \sigma_0^2 + x_j x_k \sigma_1^2 + \delta_{jk} \sigma^2, \end{aligned}$$

where

$$\delta_{jk} := \begin{cases} 0 \text{ iff } j \neq k, \\ 1 \text{ iff } j = k, \end{cases} \tag{2.7}$$

denotes the Kronecker delta. The predictive distribution $p(y^* | y_{1:n}, \sigma_0^2, \sigma_1^2, \sigma^2)$ for unobserved y^* with input x^* is also normal. Bayesian linear regression with normally distributed noise is well studied. We refer to Gelman et al. (2013) for further details and more advanced models.

If the noise variance σ^2 is unknown it can be included in the Bayesian inference process by assigning a prior distribution and studying its posterior. By choosing an inverse gamma prior distribution $\sigma^2 \sim IG(a, d)$ and a conditional (on σ^2) normal prior for the weights, $\beta_j | \sigma^2 \sim N(0, \sigma_j^2 \sigma^2)$, $j = 0, 1$, we obtain a conjugate model. We refer to Section A.1 for details and only give a brief overview here. The hyperparameters in this model are $\sigma_0^2, \sigma_1^2, a, d$ and the joint prior distribution $p(\beta, \sigma^2 | \sigma_0^2, \sigma_1^2, a, d)$ is a normal-inverse-gamma

[2] In the more general case of a non-centred prior, $p(\beta_0, \beta_1 | \sigma_0^2, \sigma_1^2, \mu_0, \mu_1) = N(\beta_0; \mu_0, \sigma_0^2) N(\beta_1; \mu_1, \sigma_1^2)$, the mean function is $\mathrm{E}(y) = \mathrm{E}(\beta_0 + \beta_1 x + \eta) = \mathrm{E}(\beta_0) + x\mathrm{E}(\beta_1) = \mu_0 + \mu_1 x$ and the covariance function remains unchanged.

distribution (see Section A.1). The hierarchy of parameters is summarised by Bayes' formula for the posterior distribution

$$
\begin{aligned}
p(\beta, \sigma^2 | y_{1:n}, \sigma_0^2, \sigma_1^2, a, d) &= \frac{p(y_{1:n}|\beta, \sigma^2) p(\beta, \sigma^2|\sigma_0^2, \sigma_1^2, a, d)}{p(y_{1:n}|\sigma_0^2, \sigma_1^2, a, d)} \qquad (2.8) \\
&= \frac{p(y_{1:n}|\beta, \sigma^2) p(\beta|\sigma_0^2, \sigma_1^2, \sigma^2) p(\sigma^2|a, d)}{p(y_{1:n}|\sigma_0^2, \sigma_1^2, a, d)}. \qquad (2.9)
\end{aligned}
$$

This posterior as well as the marginal likelihood and the predictive distribution are of closed form (see Section A.1) but not normal.

With the notation and concepts introduced in Section 2 we now proceed to give introductions to more advanced statistical methodology: Change-point analysis is introduced in Section 3 and Gaussian process regression in Section 4.

Chapter 3

Changepoint methods

Section 1 identified the need for a rigorous mathematical approach to identify individual soil layers in a core sample based on a water content profile. The proposed approach is based on the assumption of a piecewise linear structure of the water content data. A transition from one linear segment of measurements to the next indicates the transition from one soil layer to the next. To infer the number of soil layers and the locations of the boundaries between them is the central objective of this thesis. An appropriate statistical framework for such an inference procedure is the field of *changepoint analysis*. It is introduced in this section.

Section 3.1 introduces changepoints as a general statistical concept to describe time-series-like datasets that show abrupt changes of statistical behaviour. We then review methods to analyse such data in situations where the number or locations of changepoints are unknown. In order to introduce some basic ideas we consider the case of models with at most one changepoint in Section 3.2. These ideas can then be extended to the situation where there are multiple changepoints and where the number is unknown. The discussion of such *multiple-changepoint methods* is split between frequentist and Bayesian approaches, Sections 3.3 and 3.4. Section 3.6 concludes with a discussion of the introduced methods in the face of the application to the water content profile data in Section 5.

3.1 Introduction

We are now going to give some introductory examples for the statistical concept of a changepoint as well as a formal definition. The general idea of changepoints is visualised in Figure 3.1a. All datasets displayed in Figure 3.1 show a common feature: the statistical properties of the data change at $x = 100$, 175, and 300. Consider the example of a linear regression model in Figure 3.1a. Assuming homogeneity throughout the entire dataset and trying to fit one line to all 500 data points will not yield a useful model. Splitting the dataset at $x = 100$, 175, and 300 and fitting one line to each of the four *segments* of data will provide a much better model for the data at hand.

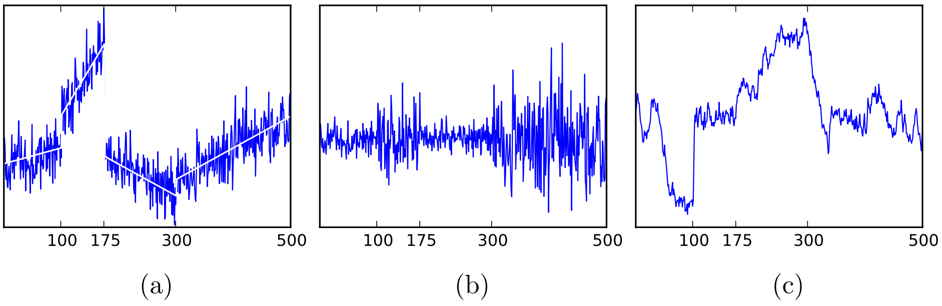


Figure 3.1: Examples for data generated from changepoint models. All datasets consist of 500 outputs with changepoints occurring at inputs 100, 175 and 300. (a) Piecewise linear mean. (b) Constant mean and piecewise constant variance. (c) AR(1) process with piecewise constant coefficients. Figure in the style of Eckley et al. (2011, Figure 10.1).

Our model will consider the data as samples from an underlying probability distribution. Whenever there is an abrupt change in this distribution from a data point (x_τ, y_τ) to the subsequent data point $(x_{\tau+1}, y_{\tau+1})$ we refer to the index τ or the input x_τ as a *changepoint location* or *changepoint*. There can also be more than one changepoint in a dataset. A collection of m changepoints $\tau_1, \tau_2, \ldots, \tau_m$ divides the dataset into $m+1$ *segments* and a given set of changepoints $\tau_1, \tau_2, \ldots, \tau_m$ is referred to as a *segmentation* of the dataset. Note that the j^{th} segment contains all measurements with indices $\tau_{j-1}+1, \tau_{j-1}+2, \ldots, \tau_j$, that is, the measurements $y_{\tau_{j-1}+1}, y_{\tau_{j-1}+2}, \ldots, y_{\tau_j}$. For indices s and t with $s \leq t$ we use the notation $y_{s:t} := (y_s, y_{s+1}, \ldots, y_t)$ such that the j^{th} segment can be written as $s_j := y_{\tau_{j-1}+1:\tau_j}$ and the entire dataset is given by $y_{1:n}$. Similarly we denote $\tau_{1:m} := (\tau_1, \tau_2, \dots, \tau_m)$. Discussing the statistical model and the computer algorithm will be more convenient in terms of indices τ_j and t_j. Discussing results however can often be more clear in terms of inputs x_{τ_j} and x_{t_j}. Due to to the one-to-one relation between indices and inputs we will use either notation when appropriate and refer to both as changepoint or changepoint location interchangeably.

Our focus is on datasets with an unknown number m of changepoints at unknown locations $\tau_1, \tau_2, \dots, \tau_m$ and the main objective is to infer about m as well as $\tau_1, \tau_2, \dots, \tau_m$. We refer to $m, \tau_1, \tau_2, \dots, \tau_m$ as *changepoint parameters.*

3.1.1 Motivational examples

Example 1. A simple example for a changepoint model is to assume a Gaussian distribution with constant variance and piecewise constant mean (Barry and Hartigan, 1993, Section 3): $y_k \sim N(\theta_j, \sigma^2)$ for all y_k in the j^{th} segment. If the variance σ^2 is unknown it can be included as a model parameter $(\theta_1, \dots, \theta_{m+1}, \sigma^2)$.

Example 2. Another example studied by Eckley et al. (2011) assumes that all observations have mean zero and that the variance is constant within each segment:

$$y_k \sim N(0, \theta_j) \tag{3.1}$$

for all y_k in the j^{th} segment. The data shown in Figure 3.1b was generated from this model based on the following parameters: $m = 3$, $\tau_1 = 100$, $\tau_2 = 175$, $\tau_3 = 300$, $\theta_1 = 1$, $\theta_2 = 5$, $\theta_3 = 1$, $\theta_4 = 10$.

Example 3. The data shown in Figure 3.1a was generated using the same changepoint parameters but a very different model for the data in each of the four segments. Each segment is described by a linear regression model

$$y = ax + b + \eta$$

with normally distributed noise η. Hence, each segment parameter θ_j consists of the slope, y-intercept and noise variance for the j^{th} segment $\theta_j =$

(a_j, b_j, σ_j^2). The likelihood for an individual segment is given by

$$p(y_{\tau_{j-1}+1:\tau_j}|\theta_j) = \prod_{k=\tau_{j-1}+1}^{\tau_j} N(y_k; a_j x_k + b_j, \sigma_j^2)$$

with

$$N(y_k; a_j x_k + b_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{[y_k - (a_j x_k + b_j)]^2}{2\sigma_j^2}\right).$$

3.1.2 A changepoint model

We now combine the concept of changepoints with a changepoint-free regression model. The changepoint parameters $m, \tau_{1:m}$ are used to split the dataset into $m+1$ segments such that each segment is homogeneous in the sense that there is no further changepoint in any of the segments. Each segment is then described by a regression model with its own *segment parameter* θ_j. In terms of the likelihood our changepoint model is given as

$$p(y_{1:n}|m, \tau_{1:m}, \theta_{1:m+1}) = \prod_{j=1}^{m+1} p(y_{\tau_{j-1}+1:\tau_j}|\theta_j), \tag{3.2}$$

where $p(y_{\tau_{j-1}+1:\tau_j}|\theta_j)$ is the segment likelihood that describes the regression model for the j^{th} segment. For notational convenience we denote $\tau_0 := 0$ and $\tau_{m+1} := n$.

Conditional on the changepoint parameters the segments are modelled independently, in other words, the data in one segment provides no information about the data in any other segment. This is the reason for the product form of the likelihood on the right-hand side of equation (3.2). Similarly, the segment parameter of each segment contains all information about the segment. Thus, the *segment likelihoods* $p(y_{\tau_{j-1}+1:\tau_j}|\theta_j)$ only depend on their respective parameter θ_j instead of all segment parameters $\theta_{1:m+1}$.

3.1.3 A changepoint method

Given a dataset of n observations $y_{1:n}$ we are interested in the number of changepoints $m \in \{0, 1, \ldots, n-1\}$ present in the data as well as their locations $\tau_1, \ldots, \tau_m$. In this section we give an overview of state-of-the-art methods that allow us to infer these parameters. A changepoint *method* is a combination of a statistical model with an algorithm that allows us to infer the changepoint parameters. The methods we discuss can be different on three levels:

(i) The statistical model used by a method: One of the most simple models for a change in mean (cf. Section 3.1.1) is given by $y_k = \mu_k + \eta_k$, where $\{\eta_k\}_{k=1}^n$ is an i.i.d. sequence of zero-mean random variables (often assumed to be Gaussian) and $\mu_k = \theta_j$ for all $\tau_{j-1} + 1 \leq k < \tau_j$. Instead of this piecewise constant mean we may consider a slightly more general model of a piecewise linear mean $y_k = (a_k x_k + b_k) + \eta_k$, where x_k is the input corresponding to measurement y_k, and the parameter (a_k, b_k) is constant for all measurements within the same segment. For both of these models all measurements are independent given the parameters. In order to introduce dependence between measurements an autoregressive model can be used. We also discuss methods that introduce a latent sequence $\{h_k\}_{k=1}^n$. Depending on the model a latent variable h_k either takes values 0 or 1 to indicate the presence of a changepoint, or indicates which segment the corresponding measurement y_k belongs to by taking one of the values $1, \ldots, m+1$. The methods then infer the value of the latent variables in order to infer the number of changepoints and their locations. Recall that, generally, any definition of a segment-wise likelihood $p(y_{\tau_{j-1}+1:\tau_j}|\theta_j)$ can be used to formulate a model.

(ii) Frequentist or Bayesian approach: frequentist approaches estimate the number of changepoints and their locations for example by maximising a penalised version of the likelihood. For most applications this maximisation requires the use of optimisation algorithms. The Bayesian approach defines a statistical model in terms of a likelihood as well as a prior distribution on the parameters of the model. For the likelihood $p(y_{1:n}|m, \tau_{1:m}, \theta_{1:m+1})$ in equation (3.2) this means that a prior $p(m, \tau_{1:m}, \theta_{1:m+1})$ needs to be specified. Similar to the frequentist approach algorithms are needed, but now to compute the posterior distribution

$$p(m, \tau_{1:m}, \theta_{1:m+1}|y_{1:n}) = \frac{p(m, \tau_{1:m}, \theta_{1:m+1})p(y_{1:n}|m, \tau_{1:m}, \theta_{1:m+1})}{p(y_{1:n})}. \quad (3.3)$$

(iii) The algorithm used by a method: As we just noted the analysis of most statistical models requires us to use computer algorithms in order to draw statistical inference. For the methods we discuss the main difference is often the algorithm that is used to either estimate the changepoint parameters or to compute the posterior distribution.

Once the model (i) is specified the objective is to find out which parameter values explain the data best. If we know that there is exactly one changepoint and we are interested in its location the problem is much simpler than the general case. Similarly we might only be interested whether one changepoint has occurred or no change is present. We start by looking at these simpler cases in Section 3.2 as they provide a good introduction to common methodology. In Sections 3.3 and 3.4 we introduce existing methods for the general problem of identifying the number of changepoints in a dataset as well as their locations.

3.2 At-most-one-changepoint methods

This section is used to introduce some basic ideas of changepoint methods. A lot of the content presented can be extended or modified to the multiple-changepoint problem that we introduced in Section 3.1 and return to in Sections 3.3 and 3.4. For this section only we assume that there is either one changepoint $\tau \in \{1, \ldots, n-1\}$ or no changepoint in the data. Hence the models we take into consideration are given by the likelihoods $p(y_{1:n}|\theta_0)$ or $p(y_{1:\tau-1}|\theta_1) \times p(y_{\tau:n}|\theta_2)$.

3.2.1 At most one change (AMOC)

If we are only interested whether there is a single changepoint in a dataset or not we can consider the changepoint problem as a hypothesis testing problem: Testing the null hypothesis "There is no changepoint." against the alternative hypothesis "There is exactly one changepoint." In order to apply a generalised likelihood-ratio test we compare the maximum likelihood value under the null hypothesis $p\left(y_{1:n}|\hat{\theta}_0\right)$ against the maximum likelihood value under the alternative hypothesis. The latter is given by $\max_{\tau \in \{1,\ldots,n\}} p\left(y_{1:\tau-1}|\hat{\theta}_1\right) \times p\left(y_{\tau:n}|\hat{\theta}_2\right)$, where the maximum likelihood estimate of the segment parameter $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ depends on the value of τ. The test statistic for the likelihood-ratio test is given by

$$\begin{aligned} D &:= -2 \ln \frac{p\left(y_{1:n}|\hat{\theta}_0\right)}{\max_{\tau \in \{1,\dots,n\}} p\left(y_{1:\tau-1}|\hat{\theta}_1\right) \times p\left(y_{\tau:n}|\hat{\theta}_2\right)} \\ &= 2\left\{\ln\left[\max_{\tau \in \{1,\dots,n\}} p\left(y_{1:\tau-1}|\hat{\theta}_1\right) \times p\left(y_{\tau:n}|\hat{\theta}_2\right)\right] - \ln p\left(y_{1:n}|\hat{\theta}_0\right)\right\}. \end{aligned}$$

This test statistic is compared to a critical value c that needs to be determined. If $D > c$ the null hypothesis is rejected (meaning that a changepoint has been found) and

$$\hat{\tau} := \arg\max_{\tau \in \{1,\dots,n\}} p\left(y_{1:\tau-1}|\hat{\theta}_1\right) \times p\left(y_{\tau:n}|\hat{\theta}_2\right)$$

is used to estimate its location.

Likelihood ratio methods make the somewhat strong assumption of a parametric form of the likelihood. They further require the evaluation of the likelihood under the null and the alternative hypotheses. For many distributions this has been done, cf. Chen and Gupta (2011); Eckley et al. (2011). They also require the determination of a critical value for the derived test statistic. Chen and Gupta (2011) provide asymptotic distributions for some test statistics but for most distributions simulations are needed. Information about the uncertainty regarding the presence of a changepoint comes with the significance level of the constructed test and does not provide information about uncertainty regarding the changepoint location.

A popular non-parametric approach to design a statistical test for the AMOC problem are cumulative sum (CUSUM) methods. In order to evaluate CUSUM test statistics only cumulative sums need to be computed. Estimation of parameters and evaluation of likelihoods are not needed. As an example consider the AMOC model for a change in mean (Section 3.1.1, Example 1 for $m \in \{0, 1\}$). We define the cumulative sums $S_k := \sum_{j=1}^{k} y_j,\ k \in \{1, \dots, n\}$ and denote $V_k := k\left(\frac{1}{k}S_k - \frac{1}{n}S_n\right)$. The non-parametric test statistic

$$T := \max_{k \in \{1,\dots,n\}} \frac{1}{\sqrt{k}} \left|V_k\right|,$$

compares the cumulative statistic S_k/k to the baseline statistic S_n/n. The definition of T becomes more perspicuous with the following observations

(Khmaladze, 2007): Under the alternative hypothesis of a single changepoint $\tau \in \{1, \dots, n\}$ the expected value of V_k is

$$\mathrm{E}_{H_1}(V_k) = \begin{cases} k(\theta_1 - \bar{\theta}), & \text{for } k \in \{1, \dots, \tau\}; \\ \tau(\theta_1 - \bar{\theta}) + (k - \tau)(\theta_2 - \bar{\theta}), & \text{for } k \in \{\tau + 1, \dots, n\}, \end{cases}$$

where $\bar{\theta} := \frac{\tau}{n}\theta_1 + \frac{n-\tau}{n}\theta_2$. Hence, for $\theta_1 > \theta_2$ a plot of the points $(k, \mathrm{E}_{H_1}(V_k)) \in \mathbb{R}^2$ yields the shape of an upside-down letter V with its peak at $k = \tau$. For $\theta_1 < \theta_2$ the shape is that of a V. Under the null hypothesis of no changepoint the expected value of V_k is simply zero for all k: $\mathrm{E}_{H_0}(V_k) = 0, k \in \{1, \dots, n\}$. Hence, if the test statistic T deviates too much from zero the null hypothesis is rejected. An estimate for the changepoint location is then given by

$$\hat{\tau} := \underset{k \in \{1, \dots, n\}}{\arg\max} \frac{1}{\sqrt{k}} |V_k| .$$

The problem of choosing a critical value c for which to reject the null hypothesis if $T > c$ and the limited treatment of uncertainty in the changepoint results persist. We refer to Khmaladze (2007) and Nam (2013), respectively, for discussions. For CUSUM methods to test a change in other model parameters than the mean we refer to Lee et al. (2006). A software implementation of CUSUM methods for changes in mean is available in the R package Killick and Eckley (2014).

3.2.2 Penalised likelihood

Up to some extent the penalised likelihood approach is similar to the likelihood ratio test discussed in Section 3.2.1. Both approaches are based on the idea of comparing likelihoods in order to decide between a model that includes a changepoint and one that does not. The key difference is the inclusion of a penalty term that penalises models with (too) many parameters. Similar to Section 3.2.1 the approach compares the penalised likelihood for the model with no changepoint

$$-2 \ln p(y_{1:n} | \hat{\theta}_0) + \dim(\theta_0) n$$

to the penalised likelihood for the model with exactly one changepoint

$$-2 \ln p(y_{1:n}|\tau, \hat{\theta}_1, \hat{\theta}_2) + (\dim(\theta_1) + \dim(\theta_2) + 1)n.$$

Generally, we can write the penalised likelihood as

$$-2 \max_{\lambda} p(y_{1:n}|\lambda) + \dim(\lambda)\phi(n),$$

where λ denotes the vector of all parameters and $\phi(n)$ the a penalty function that only depends on the number of observations. We return to possible choices for $\phi(n)$ in Section 3.3.1 where we discuss the penalised likelihood for the multiple-changepoint problem.

3.2.3 Bayesian methods

In order to obtain a Bayesian statistical model we need to specify a prior distribution for the parameters: for the number of changepoints m, the changepoint location τ, and the segment parameters $\theta_0, \theta_1, \theta_2$. The goal of the analysis is the posterior distribution of m and τ. For the at-most-one-changepoint problem the prior distribution for m is given by the probability of there being one changepoint $P(m = 1)$ as this also determines the probability of no changepoint being present $P(m = 0) = 1 - P(m = 1)$. We denote the prior distribution for the changepoint location by $p(\tau)$. An important feature of the Bayesian approach is the treatment of the model parameters θ. Instead of using a point estimate we can integrate over all possible parameter values to obtain the posterior distribution given by

$$\begin{aligned} P(m = 0|y_{1:n}) &\propto P(m = 0)p(y_{1:n}|m = 0) \\ &= P(m = 0) \int p(y_{1:n}|m = 0, \theta_0)p(\theta_0)\, \mathrm{d}\theta_0 \end{aligned}$$

and, using the notation $\theta = (\theta_1, \theta_2)$,

$$\begin{aligned} P(m = 1, \tau|y_{1:n}) &\propto P(m = 1)p(\tau)p(y_{1:n}|m = 1, \tau) \\ &= P(m = 1)p(\tau) \int p(y_{1:n}|m = 1, \tau, \theta)p(\theta)\, \mathrm{d}\theta. \end{aligned}$$

In this way any uncertainty regarding the model parameters is incorporated into our model.

In order to decide whether there is a change or not we can consider the ratio $\frac{P(m=1|y_{1:n})}{P(m=0|y_{1:n})}$ while $P(m=1,\tau|y_{1:n})$ provides information about the location with the advantages of a Bayesian approach: The full posterior distribution comprises situations where more than one changepoint location yield a good model fit and provides information regarding the uncertainty about the possible locations. General difficulties of Bayesian modelling are the choice of prior as well as computing the posterior distribution. The choice of the prior might have a significant influence on the final result of the analysis and computer simulations might be needed in order to approximate the posterior distribution. We discuss these problems and their solutions in more detail in Section 3.4.

3.3 Multiple-changepoint methods: frequentist approaches

Section 3.2 introduced several *changepoint methods* (Section 3.1.3) for scenarios in which there is either no or exactly one changepoint in a dataset. In this section we consider frequentist methods for the general case of $m \geq 0$ changepoints.

Most of the methods we discuss work in the following fashion. A statistical model is formulated, often in terms of a likelihood. A function of the data and the model parameters is introduced as a measure of fit, that is, in order to quantify how well the model fits the data for different parameter values. Due to the tight links to mathematical optimisation it is common to minimise a cost function (the negative measure of fit) rather than maximising the measure of fit. The most common form for this cost function (Killick and Eckley, 2014) is

$$\phi_1(\lambda)\phi_2(n) + \sum_{j=1}^{m+1} \mathcal{C}(y_{\tau_{j-1}+1:\tau_j}), \tag{3.4}$$

where λ denotes the vector of all parameters. Choosing $\mathcal{C}(\cdot) := -2\ln p(\cdot|\lambda)$ and $\phi_1(\lambda) := \dim(\lambda)$, for instance, leads to the penalised likelihood approach discussed in Sections 3.2.2 and 3.3.1. The sum of segment-wise cost terms is due to the product form on the right-hand side of equation (3.2).

The next step is to minimise the cost function, or more precisely, to determine the model and the parameter values that provide the best explanation for the observations at hand with respect to the cost function. With the objective of minimising a cost function we are now in a classic optimisation scenario. Hence, various methods, problems and solutions apply:

- In most applications the maximisation of the measure of fit can only be performed approximately using computer algorithms.

- Overfitting is a common problem and the penalty term $\phi_1(\lambda)\phi_2(n)$ that penalises more complex models can be used to address this.

The rest of this section introduces existing frequentist approaches that are usually given as a combination of an optimisation criterion and an algorithm. Section 3.3.1 discusses existing theory for cost functions given as the sum of a likelihood and a penalty term. Binary segmentation (Section 3.3.2), segment neighbourhood search (Section 3.3.3) and optimal partitioning (Section 3.3.4) are algorithms able to minimise cost functions given by (3.4), in particular those discussed in Section 3.3.1.

Another cost function of the form (3.4) is given by the minimum description length (MDL). It has been combined with segment neighbourhood search but was originally introduced in combination with a piecewise autoregressive model and a genetic algorithm. An overview is given in Section 3.3.5.

With an increasing size of datasets the computational complexity of algorithms becomes more and more relevant, an issue that has received an increasing amount of attention more recently. One way to reduce the complexity of an algorithms is pruning. We discuss pruned versions of segment neighbourhood search and optimal partitioning in Section 3.3.6.

3.3.1 Penalised likelihood (multivariate) as cost function

The penalised likelihood approach is a parametric frequentist approach which, for the AMOC problem was introduced in Section 3.2.2. The same idea can be extended to multiple-changepoint problems by taking a model selection view: For each possible number of changepoints m we find those locations of the m changepoints (as well as the corresponding segment parameters) that maximise the likelihood. We obtain one model for each m. In order to avoid over-fitting we add a penalty function to the likelihood in order to penalise model complexity such as a high number of parameters. The sum of likelihood and penalty term of each model can then be compared in order to determine the best model. For the at-most-one-changepoint problem this approach is similar to the likelihood ratio approach in the sense that both base their decision on a comparison of likelihoods. Common choices for the penalising function are the Bayesian information criterion (BIC), also known as Schwarz information criterion (SIC, Schwarz (1978)), and Akaike's information criterion (AIC, Akaike (1974)). The latter tends to overestimate the number of changepoints while the BIC has been shown to asymptotically estimate the correct value of m (Yao, 1988). Uncertainty quantification regarding the number of changepoints can thus be based on these asymptotic results but does not provide uncertainty results regarding their location. Comparisons of the penalised likelihood approach for different penalty functions and with other changepoint methods can be found in Eckley et al. (2011). An overview of theoretical results under different model assumptions is given by Fryzlewicz et al. (2014).

3.3.2 Binary segmentation

Binary segmentation is an algorithm that utilises any at-most-one-changepoint method to solve a multiple-changepoint problem. The at-most-one-changepoint method is applied to the entire dataset. If a changepoint is found the time series is split into two parts: the part before the changepoint and the part after. Then the at-most-one-changepoint method is applied to each of these two parts. This procedure continues until no more changepoint is found. This procedure means that binary segmentation is a greedy algorithm: At each iteration the locally optimal choice for the location of one changepoint is made. As a result not all possible changepoint configurations are explored and hence a globally optimal segmentation of the data might not be found (Killick et al., 2012). At the same time this makes the algorithm computationally fast. It is also versatile, as any at-most-one-changepoint method can be used. It is one of the most applied methods and accuracy and consistency results exist. A detailed discussion of binary segmentation is given by Fryzlewicz et al. (2014) who introduces a randomised version as *wild* binary segmentation. This modification aims to handle short distances between consecutive changepoints while requiring less application specific tuning of the algorithm.

3.3.3 Segment neighbourhood (also known as global segmentation)

Segment neighbourhood search works inductively: Based on results calculated to obtain the optimal arrangement of $m-1$ changepoints one changepoint is added to find the optimal arrangement of m changepoints amongst all possible arrangements. Optimality is determined with respect to a measure of fit, that is, by measuring how well each possible segmentation fits the data. The most common measure of fit is the segment-wise negative log-likelihood. Despite the use of a dynamic programming approach the method is computationally expensive, for example compared to binary segmentation. However, it is guaranteed to find an optimal solution as it considers all possible changepoint configurations and shows good results in simulations (Braun et al., 2000; Eckley et al., 2011; Nam, 2013).

For a detailed discussion of approximate and extended versions of segment neighbourhood search we refer to Maidstone et al. (2016) and Maidstone (2016) who consider a combination of segment neighbourhood search with the pruning method of the pruned exact linear time algorithm (Section 3.3.6).

3.3.4 Optimal partitioning

Jackson et al. (2005) proposed a dynamic programming algorithm, referred to as optimal partitioning, that is guaranteed to find the globally optimal segmentation in $O(n^2)$ time, where n denotes the number of data points. Amongst the strengths of this algorithm is the automatic determination of the optimal number of changepoints. The incremental and dynamic structure of the algorithm make it particularly suitable for online analysis. It requires an additive structure of the objective function with respect to which optimality is assessed, in other words, the cost associated to a segmentation $\{y_{1:\tau_1}, \ldots, y_{\tau_m+1:n}\}$ must be given as the sum of the cost associated to the individual segments $y_{\tau_{j-1}+1:\tau_j}$.

The pruned exact linear time algorithm (Section 3.3.6) is based on optimal partitioning but includes a pruning step in order to reduce its computational time complexity.

3.3.5 Auto-PARM using MDL

Davis et al. (2006) introduce the automatic piecewise autoregressive model (Auto-PARM), an approach that splits data into segments, each modelled by a (different) autoregressive (AR) process. While there are different AR parameters for each segment, dependence between segments is modelled by permitting AR linear dependence on previous observations also across changepoints. The approach is parametric and frequentist and assumes i.i.d. Gaussian noise. Besides the number of changepoints and their location the model includes several AR parameters. Davis et al. (2006) propose the use of a genetic algorithm that is guaranteed to find the optimal segmentation with respect to the minimum description length (MDL) criterion. The idea of MDL is to split the data into two parts: a first part explained by the model, and secondly the residual, that is, the data minus the part explained by model. It is then calculated how much space is needed to store the data using this decomposition. The model that uses the least space is considered best.

Results regarding uncertainty are available for the changepoint location, but only if the number of changepoints is assumed to be known.

Eckley et al. (2011) introduce a combination of MDL with segment neighbourhood search instead of the genetic algorithm of Davis et al. (2006).

One of the main benefits but also a potential problem is the assumption of an AR structure. The approach stands out with its capability to capture changes in autocorrelation (as well as in mean and in variance) and to model dependence between observations, even across segments. At the same time its performance might be affected if the AR assumption is not satisfied.

3.3.6 Pruned versions of optimal partitioning and segment neighbourhood search: PELT, pDPA, FPOP, SNIP

When the size n of a dataset increases the computational complexity of many algorithms eventually makes them unfeasible. While this is no problem with the relatively small dataset we are analysing in Section 5 it has motivated several recent attempts to speed up existing methods.

The pruned exact linear time (PELT) method proposed by Killick et al. (2012) is based on optimal partitioning (Section 3.3.4) but eliminates potential changepoint positions from computations at every iteration if they cannot lead to an optimal segmentation. This *pruning* can lead to a significant reduction of computational complexity: If all assumptions are satisfied a computational complexity of $O(n)$ can be achieved. If no pruning can be done the complexity $O(n^2)$ of optimal partitioning is maintained. A key assumption that needs to be satisfied is an even spread of the changepoints in the data: it assumes that the number of changepoints grows linearly with the number of data points in the dataset. Because the decision rule whether to eliminate a potential changepoint position or not is based on an inequality it is referred to as *inequality based pruning*.

The pruned dynamic programming algorithm (pDPA) proposed by Rigaill (2010) and thus also referred to as Rigaill's method is based on segment neighbourhood search (Section 3.3.3). Compared to the PELT method it also uses a different pruning criterion referred to as *functional pruning* (Rigaill, 2015). Discussions of PELT and pDPA are given by Killick et al. (2012) and Rigaill (2015).

Maidstone et al. (2016) take PELT and pDPA and exchange the combination of underlying algorithm and pruning method. This results in two new methods: the combination of functional pruning with optimal partitioning (FPOP) and the combination of segment neighbourhood search with inequality based pruning (SNIP).

3.4 Multiple-changepoint methods: Bayesian approaches

For a Bayesian statistical model we need to specify a prior distribution

$$p(m, \tau_{1:m}, \theta_{1:m+1})$$

for the parameters of the likelihood (3.2). Bayes formula (3.3) combines prior information, the model and the data into the posterior distribution

$$p(m, \tau_{1:m}, \theta_{1:m+1} | y_{1:n}) = \frac{p(m, \tau_{1:m}, \theta_{1:m+1}) p(y_{1:n} | m, \tau_{1:m}, \theta_{1:m+1})}{p(y_{1:n})}, \quad (3.5)$$

with marginal likelihood

$$p(y_{1:n}) = \int p(m, \tau_{1:m}, \theta_{1:m+1}) p(y_{1:n} | m, \tau_{1:m}, \theta_{1:m+1}) \, \mathrm{d}(m, \tau_{1:m}, \theta_{1:m+1}).$$

A point estimator similar to a maximum likelihood estimator is the maximum a posteriori (MAP) estimator

$$\underset{m,\tau_{1:m},\theta_{1:m+1}}{\arg\max}\; p(m,\tau_{1:m},\theta_{1:m+1}|y_{1:n}).$$

Other Bayesian point estimators are the posterior mean or median. However, the Bayesian approach generally provides the whole posterior distribution instead of only a point estimate. While this bulk of information is harder to summarise, the extra information is very valuable when we want to quantify the uncertainty attached to any inference we draw from a statistical analysis.

This trait of the Bayesian approach also arises if the value of the segment parameters $\theta_{1:m+1}$ is unknown and not of interest. Integrating both sides of equation (3.5) over all possible values of $\theta_{1:m+1}$ we obtain

$$p(m,\tau_{1:m}|y_{1:n}) = \frac{p(m,\tau_{1:m})p(y_{1:n}|m,\tau_{1:m})}{p(y_{1:n})} \tag{3.6}$$

and inference drawn from this posterior takes all uncertainty about the segment parameters into account.

For most models it is not possible to compute the posterior distribution analytically, a standard problem in Bayesian statistics. The following sections hence discuss methods to approximate the posterior. One difficulty in approximating the posterior PDF (3.6) is the dimension of the parameter $(m,\tau_{1:m})$ as the dimension of $\tau_{1:m}$ changes with the value of m. Reversible-jump Markov chain Monte Carlo methods address this problem and are introduced in Section 3.4.1.

Other Bayesian models avoid this problem by not inferring about the changepoint parameters m and $\tau_{1:m}$ directly. Instead a sequence of unobservable (also known as *hidden* or *latent*) variables is introduced which contains all information about m and $\tau_{1:m}$. These models are combined with sampling algorithms that approximate the posterior distribution of the latent variables. We discuss these models in more detail in Sections 3.4.2, 3.4.3 and 3.4.4.

3.4.1 Reversible-jump Markov chain Monte Carlo (RJ MCMC)

Markov chain Monte Carlo (MCMC) methods (Robert and Casella, 2004) provide a popular approach to approximate the posterior distribution in a

Bayesian statistical model. Standard MCMC methods such as the Metropolis-Hastings (MH) algorithm (Hastings, 1970) require the dimension of the parameter space to be fixed. In the model formulation given by (3.5) the dimension of the parameter $(m, \tau_{1:m}, \theta_{1:m+1})$ changes with the value of m. Reversible-jump Markov chain Monte Carlo (RJ MCMC) is an adaptation of the MH algorithm (Green, 1995). The adaptation allows the Markov chain of samples to jump between parameter spaces of different dimensions.

To illustrate this idea we outline a possible way of constructing a sampler for the model described by (3.5). This outline is based on the sampler used in Green (1995). We describe the proposal step that, based on the most recent sample $(m, \tau_{1:m}, \theta_{1:m+1})$ out of the Markov chain of samples, generates the next new sample in the chain. The first step is to randomly decide which marginal parameter will be changed. Either only a segment parameter θ_k, or only a location parameter τ_j, or the number of changepoints m. In the first two cases the second step is to, again randomly, decide which dimension k (or j) is updated and then draw a random proposed value for θ_k (or τ_j). Because the dimension of the parameter does not change the sampling is similar to that in an MH sampler. In the case where we update m the second step is to decide whether to add or remove a changepoint. If a changepoint is added we next randomly draw its location τ_{new}. The new changepoint will lie between two existing changepoints, say in the interval $(\tau_j + 1, \tau_{j+1})$. Then the segment parameter θ_{j+1} will be replaced by two randomly drawn segment parameters for the intervals $(\tau_j + 1, \tau_{\text{new}})$ and $(\tau_{\text{new}} + 1, \tau_{j+1})$. Similarly, if a changepoint is removed, two segments are joined to one and require the random draw of a segment parameter for the new segment.

Each of the above steps requires the specification of a probability distribution. These distributions must satisfy regularity conditions for the theoretical guarantee that the generated samples can be used to asymptotically approximate the posterior distribution, see Green (1995) for details. Like most MCMC algorithms RJ MCMC further requires application-specific tuning of the proposal distribution and monitoring of the convergence of the Markov chain so that successfully applying RJ MCMC is a generally difficult task. Green (1995) applied RJ MCMC to the multiple-changepoint problem. A second analysis (Green, 2003) of the same dataset showed that in the analysis of Green (1995) the Markov chain of samples had not explored the entire parameter space.

3.4.2 Product-partition models

For Gaussian observations with constant variance and a change in piecewise constant mean, $y_k \sim N(\mu_k, \sigma^2)$, Barry and Hartigan (1992, 1993) propose a formulation of the changepoint model as a product-partition model (Hartigan, 1990). Let $S_1, S_2, \ldots, S_b$ denote a partition of a set S_0 with n elements. A probability distribution over all possible partitions is called a *product-partition distribution* (PPD) if the probability of any given partition $S_1, S_2, \ldots, S_b$ can be written as a product $K \prod_{j=1}^{b} c(S_j)$, where K is a normalising constant. In the changepoint context we only consider $S_0 = \{1, \ldots, n\}$ and require that any partition not defining a segmentation (as defined in Section 3.1) is assigned probability zero. That is, $c(S_j) = 0$ if S_j is not a consecutive sequence of indices $s, s+1, \ldots, t-1, t$ resulting in well-defined $y_{s:t}$. In the notation of changepoint locations $\tau_{1:m}$ we have $b = m + 1$ and $S_j = \{\tau_{j-1}, \tau_{j-1} + 1, \ldots, \tau_j\}$. A PPD combined with the assumption of independent segments conditional on the parameters is referred to as *product-partition model* (PPM). In the changepoint context this latter assumption is given by equation (3.2). PPMs offer the advantage that computations can be performed segment-wise. Barry and Hartigan (1992, 1993) propose an exact implementation of their changepoint method at a cost of $O(n^3)$ as well as an approximation at a cost of $O(n)$. The method has been implemented in the `R` package `bcp` (Erdman et al., 2007) and applied and compared in Erdman and Emerson (2008).

For the implementation – instead of modelling the number of changepoints and their locations directly – Barry and Hartigan (1992, Section 4.1) introduce a latent process $h_{1:n}$, where $h_k \in \{0, 1\}$ indicates whether there is a changepoint at y_{k+1} or not. This formulation of changepoint locations is equivalent to that using indices $\tau_{1:m}$ but has the advantage that the distribution of $h_{1:n}$ is of dimension n. Hence, instead of RJ MCMC (Section 3.4.1), a standard Gibbs sampler (Geman and Geman, 1984; Casella and George, 1992) can be used to sample from the posterior distribution of $h_{1:n}$.

The approach by Fearnhead (2006) (discussed in Section 3.4.4) can be seen as a product-partition model. In Section 3.4.3 we discuss hidden Markov models which also introduce a latent sequence $h_{1:n}$. For PPMs the latent variables are independent but conditioning on the latent variables does not imply independence of the observations. For hidden Markov models the latent variables are not independent, but the observations are independent given the latent sequence.

3.4.3 Hidden Markov models (HMMs)

Hidden Markov models (HMMs) offer a versatile framework for time series modelling which has also been applied to the changepoint problem. The following definition of an HMM is based on Cappé et al. (2006) and suited for the application to the changepoint problem. For a general introduction and overview we refer to Cappé et al. (2006). A random variable is called *hidden* (or *latent*) if its value is unobservable. In order to give a definition of an HMM we, together with the observations $\{y_k\}_{k\in\mathbb{N}}$, define a stochastic process $\{h_k\}_{k\in\mathbb{N}}$. The stochastic process $\{(h_k, y_k)\}_{k\in\mathbb{N}}$ is an HMM iff

- $\{h_k\}_{k\in\mathbb{N}}$ are hidden,
- $\{h_k\}_{k\in\mathbb{N}}$ form a Markov chain,
- given h_k the distribution of y_k is fully determined and
- given $\{h_k\}_{k\in T}$, for some index set $T \subset \mathbb{N}$, the corresponding observations $\{y_k\}_{k\in T}$ are independent.

In order to fully define the Markov chain $\{h_k\}_{k\in\mathbb{N}}$ we need to specify the *transition probabilities* for the Markov chain to transition from a state z_i to state z_j, that is

$$P(h_{k+1} = z_j | h_k = z_i) \tag{3.7}$$

for $k \geq 2$ and $z_i, z_j \in \mathbb{R}$. The initial distribution of h_1 is specified without conditioning. The distribution of the observations $\{y_k\}_{k\in\mathbb{N}}$ is specified in terms of the *emission probability* (distribution) $p(y_k|h_k)$. The random variable h_k can be seen as a parameter that determines the distribution of the observation y_k.

In order to apply the HMM to the changepoint problem Chib (1998) considers the Markov chain to take values in the finite state space $\{1, 2, \ldots, m+1\}$ and uses h_k to indicate what segment the observation y_k is in. In terms of the Markov chain this means that if h_k takes a value i then h_{k+1} can only stay in the same state (take the same value) or move to the next state (take the value $i+1$). HMMs that restrict the way in which the latent process can move are referred to as *constrained* HMMs. For the model of Chib (1998) the initial distribution of h_1 is trivial with $P(h_1 = 1) = 1$ and the transition probabilities (3.7) are zero unless $z_j = z_i$ or $z_j = z_i + 1$. The model also requires that all segments are visited, so that $P(h_n = m + 1) = 1$. The formulation of Chib (1998) allows a Bayesian treatment via a recursive MCMC algorithm developed for HMMs (Chib, 1996). This algorithm samples from the joint posterior $p(h_{1:n}, \theta_{1:m+1}|y_{1:n}, m)$ of the latent state sequence $h_{1:n}$ and the model parameters $\theta_{1:m+1}$ given the observations $y_{1:n}$ and the number of changepoints m. Thus, the number of changepoints m needs to be estimated first and inference about the changepoint locations is drawn conditional on this estimate. Chib (1998) suggests model comparison based on Bayes factors in order to estimate m, that is, considering the ratio of the marginal likelihood $p(y_{1:n}|m)$ for two different values of m at a time. While the marginal likelihoods can rarely be computed analytically the employed MCMC method provides an approximation as a by-product. An approximation of the posterior $p(m|y_{1:n})$ can also be obtained.

The method of Chib (1998) is a sophisticated changepoint method with only few shortcomings. The MCMC method still needs to be customised and tuned and uncertainty quantification for the number and the locations of changepoints is performed separately.

A more recent approach based on a constrained HMM (Luong et al., 2013) expresses uncertainty through confidence intervals for the individual changepoint locations. However, the approach assumes the number of changepoints to be known in advance and requires an a priori estimate of the changepoint locations. This prior information heavily influences the final result and thus the approach can be seen as a method to analyse the uncertainty for a specific arrangement of changepoints at hand.

An unconstrained HMM studied by Aston et al. (2012) leads to posterior distributions that can be computed exactly, given an estimate of the model parameter $\theta_{1:m+1}$. More precisely, given the data $y_{1:n}$ and an estimate $\hat{\theta}_{1:m+1}$ of the model parameter, the following two posterior distributions can be computed without introducing error caused by approximation or sampling: the distribution of the number of changepoints $p(m|\hat{\theta}_{1:m+1}, y_{1:n})$, and $P(t \in \tau_{1:m}|\hat{\theta}_{1:m+1}, y_{1:n}) := P(\exists j : \tau_j = t|\hat{\theta}_{1:m+1}, y_{1:n})$, the probability of a changepoint at time t. A key benefit of this exact approach is that no sampling of the latent state variables h_t is needed. Thus, the corresponding approximation error and computational cost are avoided. Aston et al. (2012, Figure 3b) show that large numbers of samples are needed to obtain good approximations to the exact result if a sampling scheme is used. A key drawback of the approach by Aston et al. (2012) is that it is conditional on (an estimate of) the model parameter $\theta_{1:m+1}$. In practice $\theta_{1:m+1}$ thus needs to be estimated and any uncertainty with regards to this estimate is not accounted for.

Nam et al. (2012) and Nam (2013) extend the approach of Aston et al. (2012) to fix this shortcoming. Using a sequential Monte Carlo (SMC) method they draw samples $\{\theta_{1:m+1}^{(i)}\}_{i=1}^{N}$ from the model parameter posterior $p(\theta_{1:m+1}|y_{1:n})$ and approximate the posterior changepoint distribution: $P(t \in \tau_{1:m}|y_{1:n}) \approx \sum_{i=1}^{N} w_i P(t \in \tau_{1:m}|\theta_{1:m+1}^{(i)}, y_{1:n})$. With the results by Aston et al. (2012) all $P(t \in \tau_{1:m}|\theta_{1:m+1}^{(i)}, y_{1:n})$ can be computed exactly without simulating the latent variables h_t. Sampling error is only introduced when sampling $\theta_{1:m+1}^{(i)}$. As a by-product of the method these samples can be used to approximate the model parameter posterior $p(\theta_{1:m+1}|y_{1:n})$. Having (approximately) integrated with respect to the model parameter posterior the uncertainty regarding the model parameter θ has been fully accounted for. Further extensions considered include model selection as well as uncertainty quantification for changes in the autocovariance structure of the observations $y_{1:n}$.

Nam et al. (2012) assume the number of latent states of the HMM to be known in advance. Nam et al. (2014) propose a parallel SMC sampler to estimate this number. Their method approximates the posterior distribution of the number of underlying states and causes no additional computational cost. Nam et al. (2015) modify the method of Nam et al. (2012) to quantify uncertainty for changes in autocorrelation. Nam et al. (2012) note that their assumption of a time-homogeneous HMM implies a geometric prior distribution for the segment length which might be inappropriate for many applications. In order to specify different prior distributions Nam et al. (2012) suggest Hidden semi-Markov models and Variable Transition HMMs.

3.4.4 Exact sampling (from the posterior) via recursions

Fearnhead (2006) introduces a method to sample from the joint posterior distribution of the number of changepoints and their locations directly, that is, without approximation errors as for example those associated with MCMC methods. The corresponding algorithm also does not require application-specific tuning and the results obtained are not conditional on the model parameters. For these desirable properties of the method several (often restrictive) model assumptions must be made. The key to the exact sampling is an independence assumption: Conditional on the number of changepoints and their locations the segment parameters are modelled to be independent under the posterior distribution. This independence enables the use of a recursive procedure similar to the forward-backward algorithm for HMMs (Baum et al., 1970; Fearnhead, 2008; Scott, 2002). The method further assumes that partial (and thus segment-wise) marginal likelihoods can be computed, either due to a conjugate model or numerically, with the latter leading to an increase in computational cost. The approach is fully Bayesian and two ways of specifying a prior can be used. The first way is to take a point-process view and assign a prior to the segment lengths, indirectly specifying a prior for the number and the locations of the changepoints. This is a special case of product-partition models (Section 3.4.2). The second option, based on Green (1995), is to specify a prior for the number of changepoints and then a conditional prior for the locations given the number.

Modifications of the approach have been considered: Fearnhead (2006, Section 4.2) shows that the method can be used within an MCMC method even if the independence assumption does not hold. Integrals which can be calculated explicitly under the independence assumption then need to be approximated at additional computational cost. Fearnhead and Liu (2007) present an online version of the method. They further provide empirical results showing that computationally cheaper approximation to the method can be feasible; small probabilities that otherwise would have to be saved and processed by the algorithm are *pruned*.

Following the neat exploration in Eckley et al. (2011) we now present the method in more detail. We start with the changepoint model phrased in terms of its likelihood (equation 3.2). For a Bayesian model we need to specify a prior distribution on all parameters. For each of the segment parameters θ_j we assume the same prior $p(\theta|\psi)$, where ψ denotes a vector in which we collect all parameters used to define the prior distribution. For any observations $y_{s:t}$ from one segment with parameter θ we assume that the segment marginal likelihood

$$Q(s, t; \psi) := p(y_{s:t}|\psi) = \int p(y_{s:t}|\theta)p(\theta|\psi)\,\mathrm{d}\theta \tag{3.8}$$

can be computed analytically. For the number of changepoints and their locations we define the prior indirectly by specifying a prior distribution for the length $\tau_{k+1}-\tau_k$ of a segment $y_{\tau_{j-1}+1:\tau_j}$. We define g to be the probability mass function of this prior distribution, $g(t; \psi) := P(\tau_{k+1} - \tau_k = t|\psi)$, and S to be its survival function, $S(t; \psi) := P(\tau_{k+1} - \tau_k \geq t|\psi)$. We obtain

$$p(m, \tau_{1:m}|\psi) = S(\tau_{m+1} - \tau_m; \psi) \prod_{k=1}^{m} g(\tau_k - \tau_{k-1}; \psi), \tag{3.9}$$

where ψ, besides the parameters of the prior for θ, now also includes all parameters of the prior distribution for m and $\tau_{1:m}$. We obtain the posterior probability

$$\begin{aligned} p(m, \tau_{1:m}|y_{1:n}, \psi) &\propto p(y_{1:n}|m, \tau_{1:m}, \psi)p(m, \tau_{1:m}|\psi) && (3.10)\\ &= p(s_{m+1}|\psi)S(\tau_{m+1} - \tau_m; \psi) \prod_{k=1}^{m} p(s_k|\psi)g(\tau_k - \tau_{k-1}; \psi). && (3.11) \end{aligned}$$

In order to generate independent samples from this posterior distribution we introduce a latent process $C_{1:n}$: We define the random variable C_t as the most recent changepoint before t. Hence, C_t takes values in $\{0, 1, \ldots, t-1\}$, where $C_t = 0 = \tau_0$ indicates that no changepoint has occurred in the data before t. Given $C_{t-1} = i$ the next variable C_t can only take two values: $C_t = i$, which means that there is no changepoint at $t - 1$, or $C_t = t - 1$ iff there is a changepoint at $t - 1$. The sampling is performed using an algorithm similar to the forward-backward algorithm for hidden Markov models.[1] In the forward part of the algorithm we compute the probabilities

[1] The forward-computations presented here are identical to those in the original forward-backward algorithm. The original forward-backward algorithm then backward-computes probabilities that would correspond to $P(C_t = i|y_{1:n})$. An introduction to hidden Markov models that includes a discussion of the forward-backward algorithm can be found in Jurafsky and Martin (2009).

$\gamma_t(i) := P(C_t = i|y_{1:t})$ for $t \in \{1, \ldots, n\}$ and $i \in \{0, \ldots, t-1\}$, based on the iterative formulas

$$\begin{aligned}\gamma_t(i) &\propto \gamma_{t-1}(i) P(y_t|C_t = i, y_{1:t-1}, \psi) P(C_t = i|C_{t-1} = i, \psi)\\ &= \gamma_{t-1}(i) \frac{Q(i+1, t; \psi)}{Q(i+1, t-1; \psi)} \frac{S(t-i; \psi)}{S(t-1-i; \psi)}\end{aligned}$$

for $i \in \{0, \ldots, t-2\}$ and

$$\gamma_t(t-1) \propto Q(t, t; \psi) \sum_{j=0}^{t-2} \gamma_{t-1}(j) \frac{g(t-1-j; \psi)}{S(t-1-j; \psi)}$$

derived in Section A.3. In the backward part we generate samples from the posterior distribution (3.10) as follows. We start by drawing the last changepoint τ_m from the distribution $P(C_n|y_{1:n})$. Then, iteratively backward, we draw τ_k, given $\tau_{k+1} = t$, from $P(C_t|y_{1:t}, C_{t+1} = t)$ given by

$$\begin{aligned}P(C_t = i|C_{t+1} = t, y_{1:n}, \psi) &\propto P(C_t = i|y_{1:n}, \psi) P(C_{t+1} = t|C_t = i, y_{t+1:n}, \psi)\\ &= P(C_t = i|y_{1:t}, \psi) P(C_{t+1} = t|C_t = i, \psi)\\ &= \gamma_t(i) \frac{g(t-i; \psi)}{S(t-i; \psi)}. \qquad (3.12)\end{aligned}$$

The backward iteration is complete when $t = 0$ is drawn as a changepoint. One run of the backward iteration yields one sample from the posterior distribution of $(m, \tau_{1:m})$. The forward part of computing (and storing) probabilities only needs to be performed once. With these probabilities stored it is cheap to run the backward iteration to generate samples.

Software implementation Pseudo code describing the changepoint algorithm of Eckley et al. (2011) as it was implemented for this thesis is given as Algorithm 1. The weight function W used in the algorithm is given by

$$W(y_{s:t}, \psi) = \begin{cases} Q(y_{s:t}, \psi) \text{ if length}(y_{s:t}) = 1; \\ \frac{Q(y_{s:t}, \psi)}{Q(y_{s:t-1}, \psi)} \text{ if length}(y_{s:t}) > 1. \end{cases} \qquad (3.16)$$

In the pseudo code in Eckley et al. (2011) the index t runs from 1 to n while i runs from 0 to $t-1$. This is in consistency with the mathematical

Algorithm 1 Algorithm for simulating from the posterior distribution of changepoint positions. Adapted from Eckley et al. (2011).

```
Input:  A set of data of the form $(y_1, \dots, y_n)$.
        The weight function $W(\cdot;\psi)$ defined in equation (3.16).
        Survival function for segment length $S(\cdot;p)$.
        Values for the hyperparameters $\psi, p$.
        The number $N$ of samples that we wish to generate.
Forwards computations:
Initialise: Set $\gamma_0(0) := 1$.
Iterate: For $t \in \{1, \dots, n-1\}$:
    1. For $i \in \{0, \dots, t-1\}$:
           $\gamma_t(i) := \gamma_{t-1}(i) \dfrac{S(t-i+1,p)}{S(t-i,p)} W(y_{i:t}, \psi).$    (3.13)
    2. $s := 0$.
    3. For $j \in \{0, \dots, t-1\}$:
           $s := s + \gamma_{t-1}(j) \dfrac{S(t-j,p) - S(t-j+1,p)}{S(t-j,p)}.$    (3.14)
    4. $\gamma_t(t) := W(y_t, \psi) \times s$.
    5. Normalise: Define $A := \sum_{i=0}^{t} \gamma_t(i)$ and for all $i \in \{0, \dots, t\}$ set $\gamma_t(i) := \gamma_t(i)/A$.
Backwards simulations:
Initialise: For all $j \in \{0, \dots, N-1\}$ and all $t \in \{0, \dots, n-1\}$ set $c_j^t := 0$.
Iterate: For $j \in \{0, \dots, N-1\}$:
    1. Draw a random element $t$ from the set of indices $0, \dots, n-1$ using $\gamma_{n-1}$ as vector of probability masses.
    2. If $t > 0$ set $c_j^{t-1} := 1$.
    3. While $t > 0$:
           Draw a random element $u$ from the indices $\{0, \dots, t-1\}$ with probability masses
               $P(u = i) \propto \gamma_{t-1}(i) \; \dfrac{S(t-i,p) - S(t-i+1,p)}{S(t-i,p)}.$    (3.15)
           Set t := u.
           If $t > 0$: set $c_j^{t-1} := 1$.
Output: $N$ vectors $c_j = (c_j^0, \dots, c_j^{n-1})$, $j \in \{0, \dots, N-1\}$, where each $c_j$ represents a sample of changepoints.
```

derivation of the formulas in Eckley et al. (2011) as well as in this thesis. I implemented the algorithm in Python, which uses index origin zero. I therefore chose to provide the pseudo code in Algorithm 1 with index origin 0 for all indices. As a result particular care needs to be taken with the computation of the probability masses in equations (3.13), (3.14), (3.15): Letting $\gamma_t(i)$ denote the quantities and indices in Eckley et al. (2011) and $\beta^{\cdot}_{\cdot}$ the quantities and indices used Algorithm 1 we have

$$\begin{aligned}\gamma_t(i) &= P(C_t = i | y_{1:t}) \\ &= P(\tilde{C}_{t-1} = i | \tilde{y}_{0:t-1}) \\ &=: \beta_i^{t-1} =: \beta_i^s,\end{aligned}$$

with $s := t-1 \in \{0, \ldots, n-1\}$. In equation (3.15) we use $\beta_i^{r-1} = \gamma_r(i)$, where r is the changepoint location drawn in the previous step. Equation (3.15) results from equation (3.12). Therefore we must multiply $\beta_i^{r-1} = \gamma_r(i) = P(C_r = i|y_{1:t})$ by a factor $\frac{g(r-i)}{S(r-i)} = \frac{S(r-i)-S(r-(i-1))}{S(r-i)}$ and *not* shift the index to be $\frac{g((r-1)-i)}{S((r-1)-i)}$.

3.5 Further reading

Before we conclude with a discussion (Section 3.6) of the presented changepoint methods we point to further work that is beyond the scope of this thesis.

3.5.1 Existing simulation studies

Killick et al. (2012) compare PELT, binary segmentation and optimal partitioning. Fryzlewicz et al. (2014) proposes wild binary segmentation and compares his method with other methods that are available implemented in `R` packages, including binary segmentation and PELT. Eckley et al. (2011) compare the Bayesian approach of Fearnhead (2006) with several frequentist methods such as segment neighbourhood in combination with MDL and penalised likelihood as cost functions, and binary segmentation based on the likelihood-ratio test and a test based on Bayes factors.

3.5.2 Changepoint methods considered in this thesis versus other existing methods

The examples from Section 3.1.1 and the statistical model described in Section 3.1.2 are based on several assumptions that suit the problem posed in Section 1. The idea of a changepoint problem is more general: Given an ordered sequence of observations $y_1, \ldots, y_n$ we wish to identify indices $\tau_1, \ldots, \tau_m$ such that the observations within each segment $y_{\tau_{j-1}+1:\tau_j}$ were drawn from the same probability distribution.

Different cases of this very general problem occur in various fields of research. As a result various methods exist but are often specific to a particular case of the changepoint problem. We now point out the scope of the methods considered in this thesis and provide references for information on content that is not covered here.

While our dataset is ordered by the depth at which the measurements were made any ordered set of data points (*time series*) might be searched for changepoints. We only consider one-dimensional indices and one-dimensional measurements. Changepoint problems for multivariate measurements in an offline scenario have been considered by Matteson and James (2014).

We mostly discuss parametric statistical models. A starting point for literature on non-parametric changepoint methods are the article by Pettitt (1979) as well as the recent book by Brodsky and Darkhovsky (2013). Schmidt and Morup (2013) is a recent example for a non-parametric Bayesian approach.

Most parametric models we discuss assume one family of probability distributions with different parameter values for each segment. Generally different distributions for each segment and non-parametric models can be employed.

Changepoint problems can be split into *online* and *offline* problems. In the online scenario the data is obtained and analysed sequentially. It is often desirable to detect a changepoint as soon as possible but a single new observation might not be sufficient yet in order to detect a change. In the offline scenario the dataset is analysed retrospectively, in other words, once all observations have been made. Of course, any online method can be applied in an offline scenario. The dataset described in Section 5.1 is available offline.

The assumption of independence between measurements from different segments facilitates computations and is essential for most of the methods discussed in this thesis. While this assumption is reasonable for the present application in Section 5 there are applications for which it does not hold. We refer to Nam (2013, Section 6.1) and Fryzlewicz et al. (2014) for further information.

3.6 Discussion

We have reviewed state-of-the-art methods to infer the number and locations of changepoints in a dataset. The discussion was split into frequentist methods that provide an estimate of the optimal number and arrangement of changepoints with respect to a cost function, and Bayesian methods that approximate the posterior distribution of a Bayesian model. Amongst the frequentist methods binary segmentation is a simple and fast algorithm but not guaranteed to find the optimal segmentation of a dataset. Its greedy procedure does not check all possible arrangements of changepoints. Segment neighbourhood search checks all possible segmentations which is of course associated with a high computational complexity. If the corresponding assumptions are satisfied the PELT method provides the exactness of segment neighbourhood at the computational complexity of binary segmentation.

With the dataset at hand (Section 5.1) we are not concerned about computational complexity. The main focus is to quantify the uncertainty that comes with the result of a changepoint method. For the frequentist methods uncertainty quantification can be at most provided using asymptotic arguments. The Bayesian approach addresses the issue of uncertainty quantification naturally and is therefore employed for the analysis in Section 5.

Chapter 4

Gaussian process regression

Section 3 introduced changepoint methods as an appropriate way to model and identify soil layers based on a water content profile of a core sample. Changepoint models require the specification of a regression model for each changepoint-free segment of data. In particular does the Bayesian method of Eckley et al. (2011), introduced in Section 3.4.4 and recommended as a result of the discussion in Section 3, require a Bayesian model for which we can compute the marginal likelihood. In this thesis we introduce the combination of the changepoint method of Eckley et al. (2011) with a Gaussian process model. The potential of this formulation lies in the graphical intuition of Gaussian processes and the use of the kernel trick (Section 4.4.2). The latter means that the piecewise linear model used in Section 5 is immediately applicable for datasets with non-linear segment structures.

This section introduces Gaussian processes as a state-of-the-art non-parametric Bayesian approach to regression analysis. This motivates the use of a Gaussian process model in the context of Bayesian changepoint analysis. Section 4.1 gives a probabilistic definition of a Gaussian process. On the basis of examples we provide some insight into how different choices of the covariance structure affect the data generated or described by a Gaussian process model. In Sections 4.2 and 4.3 we present how Gaussian processes can be used to perform Bayesian regression analysis. Section 4.4 gives an alternative derivation of Gaussian process models as Bayesian linear regression models. We use this view to make further conceptual and practical remarks. Sections 4.5 and 4.6 discuss the application of Gaussian process models to big datasets or datasets for which an assumption of normally distributed noise is unsuitable. These sections give an account of topical research in the field of Gaussian process regression and indicate that the approach presented in this thesis can be extended to datasets of bigger size and models with non-Gaussian likelihood. Section 4.7 summarises the potential of a Gaussian process approach to regression. We conclude that it is a potent choice for the regression part of a changepoint analysis, an example of which is given in Section 5.

4.1 Gaussian processes

A stochastic process is a collection of random variables with a continuous index set. Let $\{f(x)\}_{x\in\mathcal{X}}$ denote a stochastic process, so that each $f(x)$ is a random variable and $\mathcal{X}$ a continuous subset of $\mathbb{R}$. A stochastic process is fully described by consistently specifying the joint distribution of $f(x_1), f(x_2), \ldots, f(x_n)$ for any collection of $n \in \mathbb{N}$ indices $x_1, \ldots, x_n$.

A special case of stochastic processes are Gaussian processes, which can be seen as an extension of the multivariate Gaussian distribution. A stochastic process is called a Gaussian process if any collection $f(x_1), f(x_2), \ldots, f(x_n)$ follows a multivariate Gaussian distribution. Similarly to a Gaussian distribution a Gaussian process is fully defined by its mean function $m(\cdot)$ and covariance function $k(\cdot,\cdot)$:

$$\begin{aligned}
&m(x) := \mathrm{E}[f(x)],\\
&k(x,x') := \mathrm{E}[(f(x)-m(x))(f(x')-m(x'))],
\end{aligned}$$

with $x, x' \in \mathcal{X}$.

Let $(\Omega, \mathcal{A}, P)$ be the probability space on which we define the Gaussian process. Most of the time it is a notational overload to express the dependence of $f(x)$ on $\omega \in \Omega$ and hence omitted. But for clarity in the definition let us briefly consider the random variable $f_\cdot(x) : \Omega \to \mathbb{R},\ \omega \mapsto f_\omega(x)$. Each realisation $\{f_\omega(x)\}_{x\in\mathcal{X}}$ of a Gaussian process defines a function via $f_\omega : \mathcal{X} \to \mathbb{R},\ x \mapsto f_\omega(x)$. In this sense it is often said that a Gaussian process specifies a probability distribution over a class of functions. The mean function and, in particular, the covariance function of the Gaussian process determine what class of functions is assigned high or low probability.

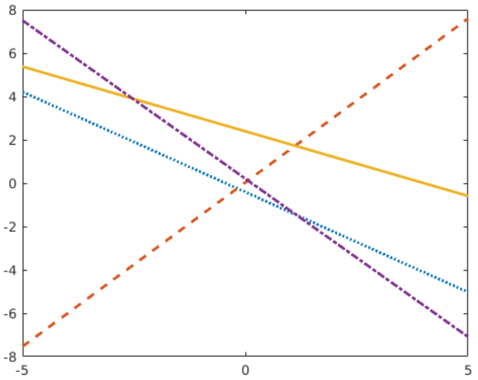


Figure 4.1: Four functions drawn from a Gaussian process with mean zero and linear covariance function (4.1) with parameters set to $\sigma_0^2 = \sigma_1^2 = 1$.

4.1.1 Examples for covariance functions

A Gaussian process with mean and covariance function given by

$$m(x) := 0, \; k(x, x'; \sigma_0, \sigma_1) := \sigma_0^2 + \sigma_1^2 x x' \tag{4.1}$$

yields a probability distribution over linear functions of the form $x \mapsto w_1 x + w_0$ with $w_0, w_1 \in \mathbb{R}$. An example of samples randomly drawn from such a Gaussian process is shown in Figure 4.1. It turns out that the Gaussian process defined by (4.1) is equivalent to randomly drawing a y-intercept w_0 and a slope w_1 from independent Gaussian distributions with mean zero and variances σ_0^2 and σ_1^2, respectively: $w_j \sim N(0, \sigma_j^2), \; j \in \{0, 1\}$.

This equivalence between Bayesian linear regression and the Gaussian process model specified by (4.1) is discussed more generally by Williams (1998) who also shows that the potential of Gaussian processes goes far beyond this simple linear example. Different choices for the covariance function result in distributions over various different classes of functions that can be used for (non-)linear Bayesian regression. Properties such as differentiability, stationarity or periodicity can be imposed by the choice of the covariance function. While the construction of a valid covariance function is generally no easy task, a variety of examples and applications can be found in the literature. We refer to Rasmussen and Williams (2006) for a comprehensive treatment and only consider here the covariance functions that are most commonly used in the literature.

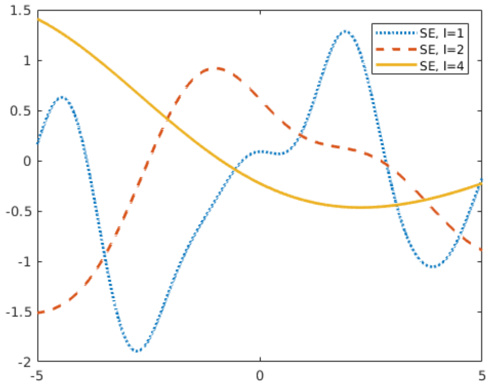


Figure 4.2: Three functions drawn from a Gaussian process with mean zero and squared exponential covariance function (4.2) using parameter values $\ell \in \{1, 2, 4\}$.

The standard example for a covariance function is the squared exponential (SE) covariance function

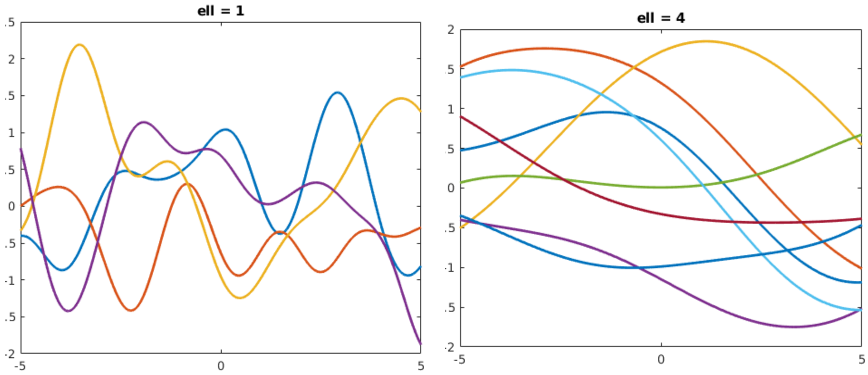
$$k_{\text{SE}}(x, x') = \exp\left(-\frac{(x - x')^2}{2\ell^2}\right). \tag{4.2}$$

Its prevalence is most likely due to having a simple form while still being sufficiently versatile to yield satisfactory results in a wide range of applications.[1] In the form stated in (4.2) the SE covariance function only has one parameter ℓ. For similar values of x and x', that is, sufficiently small $(x-x')^2$, the covariance is close to one. It decreases as $(x-x')^2$ gets larger. The *characteristic length-scale* parameter ℓ determines the scale of this decrease as visible in Figure 4.2: Consider nearby x_1, x_2; for larger values of ℓ we observe higher (positive) correlation between the function values $f(x_1), f(x_2)$. The same observation can be made in all plots shown Figures 4.3 and 4.4.

Covariance functions that depend on the inputs x, x' only through the difference $d := x - x'$ are called *stationary* and often written as a function of d, for example $k_{\text{SE}}(d) = \exp\left(-\frac{d^2}{2\ell^2}\right)$.

Realisations of a zero-mean Gaussian process with SE covariance func-

[1] During the first chapters Rasmussen and Williams (2006) almost exclusively use the SE covariance function as example before introducing various other choices in chapter 4. Throughout their book the SE covariance function keeps being used as reference. Wilson and Adams (2013, section 4.2) comment on strengths and weaknesses of the SE covariance function and compare its performance in several applications.



(a) $\ell = 1$.

(b) $\ell = 4$.

Figure 4.3: Functions drawn from Gaussian processes with zero mean function, $m(x) = 0$ for all x, and squared exponential covariance (4.2) function. (a) The characteristic length-scale was set to $\ell = 1$ and four functions were drawn. (b) The characteristic length-scale was set to $\ell = 4$ and eight functions were drawn.

tion are smooth (infinitely differentiable). It can be derived from the Matérn family of covariance functions,

$$k_{\text{Matern},\nu}(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}d}{\ell} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}d}{\ell} \right),$$

as the smoothness parameter $\nu > 0$ tends to infinity. Here Γ denotes the Gamma function and K_ν denotes the modified Bessel function of the second kind. For values of ν that are of the form $p + \frac{1}{2}$ for some non-negative integer p the Gamma and the Bessel functions take simpler forms. In particular, we have (Abramowitz and Stegun, 1965):

$$\begin{aligned} k_{\text{Matern},1/2}(d) &= \exp\left(-\frac{d}{\ell}\right); \\ k_{\text{Matern},3/2}(d) &= \left(1 + \frac{\sqrt{3}d}{\ell}\right) \exp\left(-\frac{\sqrt{3}d}{\ell}\right); \\ k_{\text{Matern},5/2}(d) &= \left(1 + \frac{\sqrt{5}d}{\ell} + \frac{5d^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}d}{\ell}\right), \end{aligned} \tag{4.3}$$

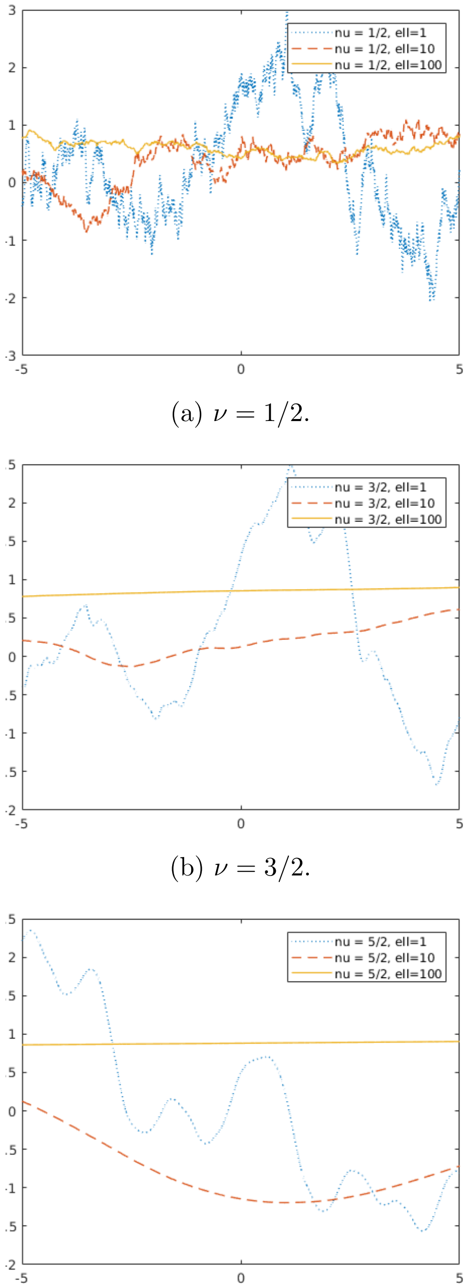
where $k_{\text{Matern},1/2}$ is also known as the exponential covariance function.

The characteristic length-scale parameter $\ell > 0$ plays the same role as it does for the squared exponential covariance function. The parameter $\nu > 0$ controls the smoothness of the sample functions drawn from a Gaussian process with a Matérn covariance function. For increasing ν the functions become smoother, illustrated in Figure 4.4. For $\nu = 5/2$ sample functions already look similar to those of the squared exponential covariance function in Figures 4.2 and 4.3. Rasmussen and Williams (2006) argue that in applications with noisy data it might be hard to decide which value $\nu \geq \frac{7}{2}$ best models the data, and that even the distinction from an SE covariance function might not be possible. Hence, the cases $\nu \in \left\{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\right\}$ given in (4.3) are most commonly implemented in practice, for example in the *GPML* Matlab software package by Rasmussen and Nickisch (2005 – 2017) that was used to create Figures 4.1 to 4.4.

By definition of the covariance function we have $\operatorname{Var}(f(x)) = k(x, x)$ with $d = x - x = 0$. For the SE and the Matérn covariance function this implies a variance of $k(0) = 1$. Alternative conventions include the *signal variance* σ^2 as a parameter of the covariance function, for example by defining the squared exponential covariance function as $\sigma^2 \exp\left(-\frac{d^2}{2\ell^2}\right)$.

4.1.2 Visualisation of functions randomly drawn from a Gaussian process

Figures 4.1 to 4.4 show functions that were randomly drawn from Gaussian processes. The usual procedure to plot a deterministic function f is to evaluate the function at finitely many inputs $x_1, \ldots, x_n$, then plot the points $(x_j, f(x_j))$, and interpolate between them. Plotting a function randomly drawn from a Gaussian process works similarly, except that evaluating f means to draw $(f(x_1), \ldots, f(x_n))$ from a multivariate normal distribution. We start by choosing a mean function m, a covariance function k and parameter values. As an example, consider the zero mean function and the squared exponential covariance function with $\ell = 1$ in Figure 4.3a. We choose an interval $[a, b] \subset \mathcal{X}$ as domain for the plot and a partition of the interval, $a = x_1 < x_2 < \cdots < x_n = b$. Figure 4.3a uses the interval $[a, b] = [-5, 5]$ and $n = 500$ equidistant inputs x_j. Next we compute the mean vector $\mathbf{m} := (m(x_1), \ldots, m(x_n))$ and the $n \times n$ covariance matrix $K = (K_{j,k})_{j,k=1,\ldots,n}$ given by $K_{j,k} := k(x_j, x_k)$. For Figure 4.3a, the mean vector is the zero vector of length $n = 500$ and K is a 500×500 matrix with $K_{j,k} = k_{SE}(x_j, x_k)$. We then draw a sample $(y_1, \ldots, y_n)$ from the n-



(a) $\nu = 1/2$.

(b) $\nu = 3/2$.

(c) $\nu = 5/2$.

Figure 4.4: Sample functions drawn from a Gaussian process with mean zero and Matérn covariance function. Each figure corresponds to one value of ν and shows three samples. Each sample was drawn from a GP with length-scale parameter $\ell \in \{1, 10, 100\}$, respectively. A discussion of the parameters ℓ and ν is given in the text.

dimensional normal distribution with mean $\mathbf{m}$ and covariance matrix K. Finally, we plot the points (x_j, y_j) and linearly interpolate between them.

For rough functions such as those drawn from a Matérn covariance function with $\nu = 1/2$ (Figure 4.4a) a finer partition of $n = 1000$ or $n = 2000$ might be needed. If the partition is not fine enough for the resolution of the plot, the interpolation can have the effect that the drawn rough functions look smoother than they are. For the plots based on the linear covariance function in Figure 4.1 it is sufficient to use $n = 2$.

4.2 Gaussian process regression

Section 4.1 introduced the probabilistic concept of Gaussian processes. In Section 4.2.1 we use this concept to define a Gaussian process regression model. Observations y are modelled to be given by an unobserved function f plus observational noise. The unobserved function f is modelled as realisation of a Gaussian process. Adding Gaussian noise results in a Gaussian process description of the observations y. Section 4.2.2 discusses the treatment of the parameters of the Gaussian process.

4.2.1 Function-space view

We start with a dataset $\mathcal{D} = \{(x_i, y_i) | i = 1, \dots, n\}$, where (x_i, y_i) is an input x_i with corresponding observation y_i. Our regression model assumes that the relation between inputs and corresponding observations is given by

$$y_i = f(x_i) + \eta_i$$

for an underlying function $f : \mathcal{X} \to \mathbb{R}$ and i.i.d. Gaussian observational noise: $\eta_i \sim N(0, \sigma^2)$ for all $i = 1, \dots, n$. Hence, letting I denote the $n \times n$ identity matrix and defining $\mathbf{f} := (f(x_i))_{i=1,\dots,n}$ we can write

$$y_{1:n} | \mathbf{f}, \sigma^2 \sim N(\mathbf{f}, \sigma^2 I). \tag{4.4}$$

Gaussian process regression is considered non-parametric in the sense that we do not infer about the unobserved "parameters" $f(x_i)$ by computing their posterior distribution. Instead $\mathbf{f}$ is assigned a prior distribution and integrated out as part of the regression analysis. More precisely, we assign a Gaussian process prior to f, that is, for any selection of $x_1, \ldots, x_n$ we assume

$$(f(x_i))_{i=1,\ldots,n}|\psi \sim N(\mathbf{0}, \mathbf{K}) \tag{4.5}$$

with mean zero $\mathbf{0} \in \mathbb{R}^n$ and covariance matrix $\mathbf{K}$ given by $\mathbf{K}_{i,j} = k(x_i, x_j)$ for some covariance function $k(\cdot, \cdot)$ with parameters ψ. Given (ψ, σ^2) we have a conjugate model with marginal likelihood

$$y_{1:n}|\sigma^2, \psi \sim N(\mathbf{0}, \mathbf{K} + \sigma^2 I). \tag{4.6}$$

Effectively, we have defined a Gaussian process $\{y(x)\}_{x \in \mathcal{X}}$ with mean function zero and covariance function

$$k_y(x_j, x_k) = k(x_j, x_k) + \sigma^2 \delta_{jk}, \tag{4.7}$$

where δ_{jk} denotes the Kronecker delta defined in (2.7).

We recapitulate the hierarchy between parameters by writing down the above model in terms of Bayes' formula for the posterior distribution of all parameters:

$$\begin{aligned} p(\sigma^2, \psi, \mathbf{f}|y_{1:n}) &\propto p(y_{1:n}|\sigma^2, \psi, \mathbf{f})p(\mathbf{f}|\sigma^2, \psi)p(\sigma^2, \psi) \\ &= p(y_{1:n}|\sigma^2, \mathbf{f})p(\mathbf{f}|\psi)p(\sigma^2)p(\psi). \end{aligned}$$

Integrating out $\mathbf{f}$ yields the posterior distribution over the hyperparameters only:

$$p(\sigma^2, \psi|y_{1:n}) \propto p(y_{1:n}|\sigma^2, \psi)p(\sigma^2)p(\psi). \tag{4.8}$$

4.2.2 The (hyper)parameters of the covariance function

In the literature on Gaussian process regression the parameters of the covariance function are often referred to as hyperparameters. The reason for this terminology lies in the construction of the model in Section 4.2.1, where ψ denotes the parameter of the prior distribution on the parameters $\mathbf{f}$. The noise variance is often included as a hyperparameter of the covariance function k_y. Besides the more concise notation and terminology, another reason for this is the similar treatment of σ^2 and ψ in the analysis.

Because σ^2 and ψ are unlikely to be known in practice a fully-Bayesian approach would be to: assign a prior distribution to both parameters and analyse their posterior distribution (4.8). In Section 2.4 the normal-inverse-gamma model is used to assign a prior distribution to σ^2. While this approach still yields a conjugate model it disrupts the “pure” Gaussianity that is maintained in the Gaussian process model. Similarly, assigning a prior distribution to ψ leads to a collapse of the practical, elegant but fragile end-to-end Gaussian structure. Quantities of interest such as marginal likelihood, posterior, and the predictive distribution no longer have a closed form. We then rely on numerical methods to approximate these.

A common approach in practice is the estimation of the hyperparameters using the ML-II approach, that is, by maximising the marginal likelihood (4.6) with respect to the hyperparameters and setting

$$(\widehat{\sigma^2}, \widehat{\psi}) := \arg\max_{\sigma^2, \psi} p(y_{1:n}|\sigma^2, \psi).$$

The partial derivatives of the marginal likelihood with respect to the hyperparameters can be computed and hence gradient-based optimisers can be used. For details, also on alternative methods such as cross-validation, we refer to Rasmussen and Williams (2006, Section 5.4).

4.2.3 Gaussian process regression with linear covariance function

Consider the following three ways of performing linear regression with Gaussian noise in a Bayesian fashion:

(A) A Gaussian process model with linear covariance function (4.1) with parameters $\sigma_0^2, \sigma_1^2, \sigma^2$;

(B) Bayesian linear regression with known variance σ^2 discussed at the beginning of Section 2.4;

(C) Similar to (B) but with an inverse gamma prior on the unknown noise variance as in equation (2.8).

We have seen that models (A) and (B) are equivalent in the following sense: In both models the observations $y_{1:n}$ follow a multivariate Gaussian distribution. Hence, they are fully described by means and pairwise (co)variances. Sections 2.4 and 4.2.1 show that these characteristics are identical for both models. Thus, if we only consider the statistical model for the observations

$y_{1:n}$, the models (A) and (B) are equivalent. The difference between the two is how we arrive at this final description of $y_{1:n}$. This relationship is an example of the more general relationship between the *weight-space view* and the *function-space view* on Gaussian processes and topic of Section 4.4.

The parametric hierarchical structure for each of the models is different and each comes with its own advantages and disadvantages. The weight-space view might seem more intuitive because it is a somewhat simple extension of the standard linear regression model. Once we consider more complex structures it might become easier to take the function-space view to combine covariance functions and tune their parameters. Even the seemingly simple example of a squared exponential covariance function corresponds to a non-trivial infinite linear combination of Gaussian basis functions (Rasmussen and Williams, 2006, Section 4.2.1). From a practitioner's perspective it is easier to think about what regression function might be suitable rather than to think about what linear combination of basis functions to choose.

In comparison to model (C), models (A) and (B) do not assign a prior distribution to the noise variance σ^2 but instead treat it as known or estimate its value from the data. As discussed in Section 2.4 assigning a prior to σ^2 as in (C) is closer to the fully Bayesian approach but disrupts the Gaussianity of the model.

4.3 Gaussian process prediction and imputation

Suppose we are in the situation of Section 4.2.1. For a new input x^* it is a matter of interest to predict the value of $f(x^*)$ based on the observed data $\mathcal{D}$, in other words, we are interested in the distribution of $f(x^*)$ given $\mathcal{D}$. For known fixed values for the parameters σ^2 and ψ we firstly have (Rasmussen and Williams, 2006, Section 2.2):

$$f(x^*)|\mathcal{D}, \sigma^2, \psi \sim N\left(\mathbf{k}_*^T(\mathbf{K}+\sigma^2 I)^{-1} y_{1:n},\ \mathbf{k}_{**} - \mathbf{k}_*^T(\mathbf{K}+\sigma^2 I)^{-1}\mathbf{k}_*\right), \quad (4.9)$$

where $\mathbf{k}_* = (k(x^*, x_i))_{i=1,\dots,n}$ and $\mathbf{k}_{**} = k(x^*, x^*)$. In practice σ^2 and ψ will be unknown and we approximate

$$p(f(x^*)|\mathcal{D}) = \int p(f(x^*)|\mathcal{D}, \sigma^2, \psi) p(\sigma^2, \psi|\mathcal{D})\, \mathrm{d}\left(\sigma^2, \psi\right).$$

For a full Bayesian treatment Monte Carlo methods can be employed. A simpler approximation is given by

$$p(f(x^*)|\mathcal{D}) \approx p(f(x^*)|\mathcal{D}, \sigma_0^2, \psi_0), \tag{4.10}$$

where σ_0^2, ψ_0 denote point estimates of σ^2, ψ. This corresponds to the approximation of $p(\sigma^2, \psi|\mathcal{D})$ by a singular distribution with mass in (σ_0^2, ψ_0) only. A common choices is the ML-II estimate. This simple approximation does not take uncertainty regarding σ^2, ψ into account but the right hand side of equation 4.10 can be computed analytically via (4.9).

4.4 The weight-space view

In Section 4.2.1 we introduced Gaussian process models as a way of defining a prior distribution over functions. The approach we follow there is commonly known as taking the function-space view (Williams, 1998). For the choice of a linear covariance function $k(\cdot, \cdot)$ Section 4.2.3 discusses the equivalence to a Bayesian linear regression model. Consider the more general linear regression model given by

$$f(x) = \sum_{i=1}^{p} w_i \phi_i(x) = w^T \phi(x),$$

with a vector of weights $w := (w_1, \ldots, w_p)^T$ and a vector of basis functions $\phi_1, \ldots, \phi_p$ evaluated at x, $\phi(x) := (\phi_1(x), \ldots, \phi_p(x))^T$. We assign a Gaussian prior to the weights, $w \sim N(\mathbf{0}, \Sigma_w)$. While this procedure is different from the function-space approach we can already see that this defines a prior distribution over some class of functions: With the basis functions fixed we obtain, for each realisation of w, a function f. As a linear combination of normally distributed w_i, the distribution of $f(x_j)$ is also normal. Hence, we have defined a (potentially singular) Gaussian process with mean zero and covariance function

$$\begin{aligned} k(x_j, x_k) = \mathrm{Cov}(f(x_j), f(x_k)) &= \sum_{i=1}^{p} \sum_{l=1}^{p} \phi_i(x_j) \phi_l(x_k) \mathrm{Cov}(w_i, w_l) \\ &= \phi(x_j)^T \Sigma_w \phi(x_k). \end{aligned} \tag{4.11}$$

This shows how we can change our view from weight-space to function-space.

The reverse direction, namely that Gaussian processes defined in function space also have a representation in weight space, is a consequence of Mercer's theorem (Section A.2). The theorem, simply speaking, guarantees that covariance functions (such as the examples discussed in Section 4.1.1) can be written in the form of (4.11). For this view note that for positive definite Σ_w the covariance function (4.11) is an inner product via $\langle a, b\rangle_{\Sigma_w} := b^T \Sigma_w a$ for $a, b \in \mathbb{R}^p$.

4.4.1 Singularity

For linearly dependent basis functions or for $p < n$ the multivariate normal distribution of $\mathbf{f}$ is singular: The space spanned by p basis functions is at most of dimension p. Hence, if we consider $n > p$ points $f(x_1), \dots, f(x_n)$ in this space they must be linearly dependent.

The Bayesian linear regression model (Section 2.4) is obtained as the special case $p = 2$, $w = (\beta_0, \beta_1)$, $\phi_1 \equiv 1$, $\phi_2(x) = x$, and $\Sigma_w = \operatorname{diag}(\sigma_0^2, \sigma_1^2)$. This model defines a prior distribution over lines, which are fully determined by two points that lie on it. Any further points on the line will be a linear combination of the other two. In consistency with the above observation the linear covariance function leads to a singular model as soon as there are $n > 2$ observations. In practice this issue is irrelevant when we add the noise term to model the observations $y(x_j) = f(x_j) + \eta_j$. In weight space this can be represented by basis functions $\phi_1 \equiv 1$, $\phi_2(x) = x$, $\phi_i(x_j) = \delta_{i=j+2}$ for $3 \le i \le n+2$, $1 \le j \le n$, and prior covariance matrix $\Sigma_w = \operatorname{diag}(\sigma_0^2, \sigma_1^2, \sigma^2, \sigma^2, \dots, \sigma^2) \in \mathbb{R}^{(n+2)\times(n+2)}$. The diagonal structure of Σ_w implies independence of the weights which, combined with the linearly independent basis functions, guarantees non-singularity of $\mathbf{f}$.

4.4.2 The kernel trick

The duality between weight and function space is often referred to as the *kernel trick*. The "trick" is that, by working with covariance functions given by

$$k(x, x') = \langle \phi(x), \phi(x')\rangle_{\Sigma},$$

we can use the interpretation and structure of the weight space without explicitly dealing with the representation in weight space. Depending on the situation authors stress different views on this idea:

Schölkopf et al. (2002, Remark 2.8) for instance write: “Given an algorithm which is formulated in terms of a positive definite kernel k, one can construct an alternative algorithm by replacing k by another positive definite kernel $\tilde{k}$.”. The Gaussian process changepoint model introduced in this thesis (Section 5) can be looked at this way: The GP regression model with linear covariance function can be derived from Bayesian linear regression. However, none of the discussed formulas for inference and prediction depend on the specific choice of linear k. We have already seen several examples for kernels that impose non-linear structures in Section 4.1.1. The Gaussian process changepoint model in Section 5 is derived and implemented in Python with a piecewise linear model in mind. Simply by changing the covariance function k we have the model formulation and software for a variety of changepoint models readily available.

A second view on the kernel trick stresses that we do not need to know the explicit weight-space representation of a covariance function. In practice it can be easier to choose a covariance function that leads to a prior over a suitable class of regression functions (for instance from the examples in Section 4.1.1) than it is to select a suitable set of basis functions. As long as the covariance function is positive definite the Gaussian process prior is well defined.

Rasmussen and Williams (2006, Section 2.1) use the term to stress that computational savings can be made when we take advantage of high-dimensional structures in weight space at the cost of low-dimensional evaluations of the covariance function: For $n << p$, in particular for $p = \infty$, we achieve computational savings if we can avoid explicit evaluations of the basis functions $\phi(x_1), \dots, \phi(x_n)$ because the inner product is of a simpler closed form.

4.4.3 Kernel methods

The ideas discussed in Section 4.4.2 are known and applicable in a much wider context. A variety of optimisation methods are based on kernels and from this point of view GP methods are only one example of *kernel methods*. A *kernel function* (short *kernel*) is a symmetric bivariate function $k : \mathcal{X}^2 \to \mathbb{R}$. To generalise the idea of the weight-space view we define the *feature map* $\phi : \mathcal{X} \to \mathcal{F}$, $x \mapsto (\phi_1(x), \dots, \phi_p(x))$ that transforms the inputs into a (usually higher- and often infinite-dimensional) *feature space* $\mathcal{F}$. For many optimisation problems the basic idea is to find $\mathcal{F}$ and ϕ so that the optimisation becomes easier for the features $\phi(x_1), \ldots, \phi(x_n)$ than it is for the original data $x_1, \ldots, x_n$. Equation (4.11) generalises to

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}, \tag{4.12}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ denotes an inner product on the feature space $\mathcal{F}$.

In this thesis kernels occur in form of covariance functions that define the structure between two observations $y(x_j)$, $y(x_k)$ in a regression context. Further examples include classification where a kernel can describe the similarity between two observations. We refer to the book by Schölkopf et al. (2002) and restrict our remarks with the role of kernels in GP regression in mind.

Now suppose we have a kernel method using a kernel k. What kind of functions $\tilde{k}$ are able to do the kernel trick? And when does a kernel k have a representation as in (4.12)? We have seen the answer to the first question in the context of GP regression: Whether or not a covariance function $\tilde{k}$ leads to a well-defined (that is, non-singular) GP model is equivalent to $\mathbf{K} = (k(x_j, x_k))_{j,k}$ being positive definite for all $x_1, \ldots, x_n$. The above quote by Schölkopf et al. (2002) reveals that this observation generalises to kernel methods, that is, kernels need to be positive definite to do the kernel trick. The answer to the second question is given by Mercer's theorem. We refer to Appendix A.2 for details.

4.5 Computational cost and methods to reduce it

The dataset studied in Section 5 consists of $n = 159$ observations $(x_i, y_i) \in \mathbb{R}^2$. In the era of big data (Liu et al., 2018) this could be considered as tiny. Computational complexity is not an issue in Section 5 but the method as it is presented would become unfeasible if it was applied to bigger datasets.

The favourable properties (flexibility, interpretability, and in particular uncertainty quantification) of the Bayesian non-parametric GP approach are reflected in its computational complexity, which, generally speaking, is of order $O(n^3)$. In practice this means that datasets from the size of thousands (Hensman et al., 2013) or ten thousands (Liu et al., 2018) are already considered big data for a Gaussian process model. To scale[2] Gaussian process models up to big datasets all approaches replace exact expensive computations with cheaper approximative ones. Peng et al. (2017) perform Gaussian process regression on a dataset with over one billion trip records of taxi journeys in New York City. Other big-data applications of Gaussian processes include Hensman et al. (2013) and Deisenroth and Ng (2015) who use up to 700,000 training and 100,000 test points to predict airline delays, and Hensman et al. (2017) who study the full dataset of almost 6 million data points, using two thirds for training and one third for testing. Liu et al. (2019) compare state-of-the-art scalable Gaussian process models on datasets that describe the 3D structure of proteins, the kinematics of a robotic arm, and data from the *Sloan Digital Sky Survey* (SDSS) which "[...] has created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects."[3] The rest of this section gives a quick overview of the work that has been done in order to scale GP models to still be applicable for bigger datasets. We start by briefly explaining the computational complexity of Gaussian process modelling.

A parametric statistical model uses available data to estimate parameters of the likelihood. These estimates then form the basis for further analysis such as prediction. Gaussian process models are considered non-parametric Bayesian models in the sense that the data is not used to estimate the parameters $\mathbf{f}$. Instead $\mathbf{f}$ is integrated out of the model and the marginal likelihood as well as the predictive distribution depend directly on all data. The computational complexity of Gaussian process regression is dominated by this dependence on the data. More precisely, looking at equations (4.6) and (4.9), the data enters the PDFs through the covariance matrix $K + \sigma^2 I$. The required computation of the determinant and the inverse of this $n \times n$ matrix means that the analysis has a computational complexity of $O(n^3)$. Several approaches exist to reduce this computational complexity, mostly under the name of *scalable Gaussian process methods.* Following Liu et al. (2018) we split the discussion between global and local approximations.

[2] A statistical model is called *scalable* if it stays applicable even when the size of the dataset is increased / "scaled up".

[3] `https://www.sdss.org/`

4.5.1 Local approximations

Local approximations split the dataset $\mathcal{D}$ into m local disjoint subsets $\mathcal{D}_k$, $k = 1, \dots, m$. For each of the m subsets a Gaussian process model is fitted and referred to as "local expert". The separate treatment reduces the computational complexity for the training to $O(n^3 m^{-2})$. Prediction is performed at a complexity of $O(n^2 m^{-1})$ by combining the predictions of the individual experts. The product-of-experts (POE) model (Hinton, 2002) assumes independence of the expert predictions via the product form predictive PDF

$$p(y^*|\mathcal{D}, x^*) = \prod_{k=1}^{m} [p_k(y^*|\mathcal{D}_k, x^*)]^{\beta_k} \,,$$

where $\beta_k = 1$ for all k. As m increases the variance of this predictive distribution goes to zero leading to over-confident predictions (Deisenroth and Ng, 2015). Generalisations that choose β_k adaptively in order to counterbalance this issue go under the name of *generalised-product-of-experts* models (Cao and Fleet, 2014).

Tresp (2000) proposed *Bayesian committee machines* (BCM), an extension of the POE to ensure that the predictive distribution of y^* falls back to the prior when x^* is far from the observed data $\mathcal{D}$. A combination of BCM and the generalised POE model was proposed by Deisenroth and Ng (2015) under the name of *robust Bayesian committee machines.*

Liu et al. (2019) review local approximations and compare them to sparse approximations discussed below.

4.5.2 Global approximations

Global approximations approximate the full $n \times n$ covariance matrix K_y and we outline the most common approaches.

Tapering Kaufman et al. (2008) study *tapering* methods, also known as *sparse kernel* methods, to reduce the cost of maximum likelihood estimation of the parameters of the covariance function. The approach is based on the idea that the correlation between two observations y_j, y_k is low if they lie far from each other, more precisely, if $|x_j - x_k|$ is larger than some threshold parameter γ. The corresponding entries of the covariance matrix are set to zero, producing a covariance matrix that is sparse, with entries tapering off as we move away from the diagonal. Existing methods can then be used that, for sparse matrices, perform matrix operations at a lower computational cost, leading to a complexity of $O(\alpha n^2)$ for some $0 < \alpha < 1$.

Sparse approximations Many of the most popular methods are grouped under the name of *sparse approximations*. The underlying idea is to introduce m latent variables that enable a sparse approximation of covariance matrices. More precisely, the random latent variables $f(\tilde{x}_1), \ldots, f(\tilde{x}_m)$ (referred to as "inducing variables" or "support variables") are modelled as marginals of the Gaussian process prior at *inducing points* $\tilde{x}_1, \ldots, \tilde{x}_m$. Quiñonero-Candela and Rasmussen (2005) introduced a unifying view showing that various methods suggested in the literature can be seen as different ways of approximating the prior $p(\mathbf{f}, f(x^*))$. Their formulation of "exact inference with an approximated prior" stresses the difference to more recent approaches such as Titsias (2009) and Hensman et al. (2013). Both of the latter also introduce latent variables but the approximation is made when computing the posterior $p(\mathbf{f}, f(x^*)|\mathbf{y})$. Generally, the computational cost of sparse approximations is of the order $O(m^2 n)$, though combinations with stochastic optimisation and variational methods have recently achieved $O(m^3)$ (Hensman et al., 2017). For prediction all approaches scale at the order of $O(m^2)$. A review focussed on sparse approximation methods is given by Liu et al. (2019), showing that, compared to other methods, the low complexity of the aforementioned stochastic variational GP methods comes at the price of being sensitive to the initial setting of hyperparameters and some restrictive assumptions such as a heteroscedastic noise variance.

Spectral approximations / Fourier features A group of methods that are not based on sparse approximations, is instead based on a spectral representation of stationary covariance functions: $k(x, x') = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega|x-x'|} s(\omega) \, \mathrm{d}\omega$, where s denotes the Fourier transform of k. The approach taken by Rahimi and Recht (2008), named *random Fourier features*, uses a Monte Carlo sum to approximate this integral. The points (/ frequencies) ω_k for this approximation are randomly drawn from a PDF proportional to $s(\omega)$. Hensman et al. (2017) provide a useful review of spectral approximations. Instead of randomly drawing Fourier features they propose a regularly spaced grid. They combine this with a variational approach to sparse approximation like the one of Titsias (2009). Opper and Archambeau (2009) study a similar non-sparse approach that can be applied to Gaussian process models with non-Gaussian likelihood (see Section 4.6).

4.5.3 Discussion and further comments

How well the discussed scalable GP models perform in practice depends on the dataset at hand and can largely be said to depend on whether the underlying model assumptions match the data. In the study by Liu et al. (2019) for example, the sparse approximation of the prior by Snelson and Ghahramani (2006) is the only global model that is capable to correctly learn and predict the variance for data with heteroscedastic noise. In the analysis of data from the SDSS, uncertainty quantification is the main motivation to employ the computationally complex GP approach (Almosallam et al., 2016). Hence, all other global approximations would be considered unsuitable. However, if we only consider the quality of the mean prediction (leaving the correct estimation of the noise variance aside), stochastic variational methods (Titsias, 2009; Hoang et al., 2015, both are global sparse approximations of the posterior) outperform the heteroscedastic model, despite the incorrect assumption of homoscedastic noise. The posterior approximations converge to the exact posterior distribution as the number of inducing points m increases (Hoang et al., 2015). This, of course, is only of limited significance because (i) computational savings are only achieved for small m, and (ii) the exact posterior is still that of a model assuming homoscedastic noise. For fair comparisons of the quality of approximations the scaling parameters (that is, the number of inducing points for sparse approximations or the number of experts for local approximations) are chosen so that all compared approximations run at the same computational cost. In such settings it is interesting to observe the different strengths of global vs. local approximations. Global methods spread inducing points across the input domain and hence, when the number of inducing points is small, they miss patterns that are only visible on a local level. Local methods tend to better capture such patterns but suffer from local over-fitting, leading to worse predictions away from the local domains.

For a comparison of local approximations we refer to Deisenroth and Ng (2015). Most of this section is based on Liu et al. (2019) whose review is focussed on sparse approximations (global) versus local methods, using simulated as well as multiple real-world datasets. A review of spectral approximations is given by Hensman et al. (2017).

4.6 Gaussian process models with non-Gaussian likelihood

A Gaussian process prior with a Gaussian likelihood as specified by (4.4) leads to a conjugate model. The marginal likelihood and the predictive distribution are of closed form and given by equations (4.6) and (4.9). But Gaussian process models are not restricted to a Gaussian likelihood. Examples for non-Gaussian distributions of the noise η_i are the Student's t-distribution (Neal, 1997) and the beta distribution (Jensen et al., 2013). For these likelihoods we lose the conjugacy of the model and approximation methods are used for the analysis.

The *GPML* Matlab software package (Rasmussen and Nickisch, 2005 – 2017) includes implementations of various likelihoods as well as suitable approximation methods. The likelihoods can be summarised into three categories: standard regression with output domain $\mathbb{R} \ni y_i$; classification with $y_i \in \{-1, +1\}$; and generalised linear regression with one of the three output domains $\mathbb{R}_{>0}$, $\mathbb{N}$, or $(0, 1)$.

4.7 Summary and conclusion (Motivation for the GP approach)

Gaussian process models provide a framework to fit a regression function to a dataset in a Bayesian manner. The Bayesian approach allows us to encode prior knowledge about the class of regression functions considered suitable and results in a predictive distribution, addressing uncertainty quantification in a state-of-the-art fashion. Different covariance functions (Section 4.1.1) can be chosen to encode properties such as smoothness or periodicity. While it is not easy to invent covariance functions, there is a toolbox of well-known functions readily available. These can be combined to build more complex covariance functions as needed.

The parameters of the covariance function are used to further tune the properties of the regression functions and can be learned from the data (Section 4.2.2). Like the form of the covariance function itself, the parameters are interpretable. Examples include the noise variance, smoothness parameters, or length-scale parameters. The latter control the level of correlation between observations based on the distance between them. Different Gaussian process models[4] can be compared in a fully probabilistic framework using Bayesian model selection or cross-validation (Rasmussen and Williams (2006, Section 5.4)).

Gaussian process models are less user-friendly on big datasets and in high dimensions when there is little prior understanding of the structure and when computational cost becomes an issue. Reducing the computational cost of Gaussian process models when more structure or information is available is a topical field of research (Section 4.5).

The dataset considered in this thesis is small, the inputs are one-dimensional and due to previous research (Wang et al., 2014a) we have a good understanding of the structure. Hence, the Gaussian process framework is suitable and allows us to provide empirical evidence for our final choice of model.

[4] Following Rasmussen and Williams (2006, Section 5.1) the term model includes the choice of mean function, covariance function and parameter values.

Chapter 5

Changepoint analysis of a single core sample

Section 1 introduced the problem of soil layer identification and motivated the statistical analysis of a water content profile as a scientific way to approach this problem. This section presents such an analysis based on the method proposed by this thesis, a Gaussian process changepoint model: As a conclusion of Section 3 the model is based on the Bayesian changepoint method of Eckley et al. (2011), combined with Gaussian process models for each segment of data. Gaussian process models are an intuitive formulation of a Bayesian regression model and discussed in Section 4.

Section 5.1 describes the dataset which will be analysed. The full statistical model is recapped in Section 5.2. Results of the analysis are presented in Section 5.3 with a discussion in Section 5.4.

5.1 The dataset

The dataset at hand is a water content profile of the London Clay Formation (LCF) taken from borehole 1 at St James's Park, London (Hight et al., 2003). The dataset consists of $n = 159$ values (x_j, y_j), $j \in \{1, \dots, n\}$, where y_j denotes the water content measured at a depth of x_j. Water content is given in percent and depths are given in metres. The data is ordered by depth: $x_1 < x_2 < \cdots < x_n$ with $x_1 = 9.85$ and $x_n = 39.9$ metres. Measurements are taken approximately every 0.15 to 0.2 metres. Water content measurements $y_1, \dots, y_n$ lie between approximately 18.22% and 33.37%. The dataset is visualised in Figure 5.1.

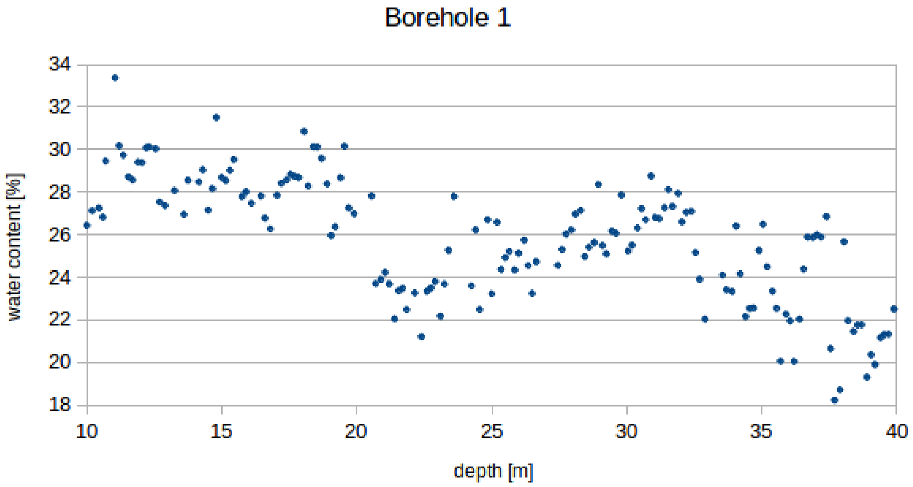


Figure 5.1: Moisture contents from split U4 samples, borehole 1, as discussed by Wang et al. (2014a) and Wang et al. (2014b). Data provided by Dr. Jamie Standing.

Around a depth of 20 metres, we see an abrupt drop of the water content from approximately 28% to approximately 24%. A similar shift is visible around a depth of 32 metres. Both of these changes have been found to coincide with boundaries between two lithological strata (Wang et al., 2014a). Between the two changes an upwards trend is visible and the data could be described as scattered around a line with positive slope. This exemplifies the structure that leads us to the description of the dataset by a piecewise linear model. Under the assumption of a piecewise linear changepoint model we would expect both of these changes to be identified as changepoints. For the segments from 10 to 20 metres and from 32 to 40 metres it is less clear where further changepoints occur and what segments of data might be well-described by a line. The statistical approach presented in this section addresses assessment of the data in a mathematically rigorous way.

The terminology and notation used throughout Section 5 was introduced in Section 3.1. Based on the present application the mathematical parameters now have the following interpretation: The soil in the core sample is made up of an unknown number of $m+1$ layers and the j^{th} segment of data $s_j = y_{\tau_{j-1}+1:\tau_j}$ contains those measurements that were taken from the j^{th}

soil layer. The first and last measurements in segment j were taken at a depth of $x_{\tau_{j-1}+1}$ and x_{τ_j}, respectively.

5.1.1 A priori information

The benchmark established by King (1981) based on microfossil content divides the LCF into 5 layers, A to E, with subdivisions A1, A2, A3, B1, B2, C1, C2, C3, D1, D2, E (Hight et al., 2003). This full profile extends over a depth of about 120 metres. The dataset analysed in this thesis only covers about 30 metres and we cannot expect all layers to occur. Depending on the location where a core sample is taken some of the top layers might not be present. In general we are thus not able to know which part of the full profile we expect to find in the data. In order to enable the comparison of my results I analyse a dataset that has been studied in the literature. In particular I restrict the analysis to be based solely on water content. Hight et al. (2003) suggest that the dataset covers layers A2, A3 and B, without further division into B1 and B2, stating that "The step changes in water content with depth were confirmed by King to be the boundaries to the lithological units [. . .]" (Hight et al., 2003, Section 3.4). A further subdivision of A3 into A3i and A3ii is shown in some but not all figures and not mentioned in the text. Standing and Burland (2006) suggest the presence of layers A2, A3, B1, B2 as well as the further subdivision of A3 into A3i and A3ii. The latter can hardly be justified if solely based on the water content profile itself but is the result of a broader analysis of the core sample of soil as well as additional core samples from nearby boreholes.

Wang et al. (2014a), in a first mathematical analysis, mainly identify the same layers as Hight et al. (2003) but also suggest one (or two) potential new layers towards the deep end (and the "beginning") of the core sample.

Hight et al. (2003) also study further datasets, in particular water content profiles from boreholes around London. Based on a water content profile from the construction site of London Heathrow Airport's Terminal 5, about 25 kilometres from the source of the present dataset in St James's Park, Hight et al. (2003, p. 869) propose: "In unit B the water content profile suggests that it would be possible to subdivide the layer into at least three parts [. . .]".[1] Due to the thickness of the individual layers being "remarkably

[1] Based on a biostratigraphical analysis De Freitas and Mannion (2007) endorse this division.

authors	SB06	K81	H03	H03T5
identified layers				B3
	B2	B	B	B2
	B1			B1
	A3ii	A3ii*	A3ii*	A3ii
	A3i	A3i*	A3i*	A3i
	A2	A2	A2	A2
estimate for m	4	3	3	≥ 4

Table 5.1: Simplified overview of lithological strata identified based on water content, see discussion in the text. The authors are Standing and Burland (2006) (SB06), King (1981) (K81), Hight et al. (2003) (H03). H03T5 refers to the results of Hight et al. (2003) for data from Terminal 5 rather than St James's Park. The asterisk (*) indicates that the split of layer A3 into A3i and ii is not properly addressed by the authors.

uniform across the London area" Hight et al. (2003, p. 869) one might expect to find a similar structure in the water content profile analysed in this thesis. An overview of the discussed findings is given in Table 5.1.

As also discussed in Section 1, these analyses are based on engineering judgement and usually done by eye. From a scientific point of view such approaches lack mathematical rigour. A Bayesian approach, such as the one taken in this thesis, is one way to perform in a mathematically sound manner the incorporating of a priori knowledge as done by the engineers.

5.2 Practical implementation

Section 5.3 presents the results of a statistical analysis of the water content profile introduced in Section 5.1. In line with the conclusions of Section 3 the analysis uses the changepoint method of Eckley et al. (2011) discussed in detail in Section 3.4.4. We now briefly recap the statistical model to provide details about my practical implementation, including the inputs of Algorithm 1: the number of samples generated, the marginal likelihood $Q(\cdot;\psi)$ required for the computation of the weight function $W(\cdot;\psi)$, the survival function $S(\cdot;p)$, the treatment of the parameters ψ and p.

The presented results are a discussion of $N = 10000$ independent samples drawn from the posterior $p(m, x_{\tau_1}, \ldots, x_{\tau_m}|y_{1:n})$. As previously explained, drawing a larger number of samples poses no problem. But this was found to not change the qualitative aspects of the presented results.

For my implementation I follow Eckley et al. (2011) by choosing a geometric prior distribution, that is,

$$S(t;p) = (1-p)^{t-1} \tag{5.1}$$

in equation (3.9). In Section A.4 we show that this is equivalent to assigning the following hierarchical prior distribution to the number of changepoints and their locations: First, a binomial distribution is assigned to the number of changepoints, $p(m) \sim B(n-1, p)$, and then, conditional on m, a uniform distribution to the changepoint locations, $p(\tau_{1:m}|m) = 1/\binom{n-1}{m}$.

For this particular choice of the survival function the fractions in equations (3.13), (3.14), (3.15) simplify:

$$\frac{S(t-i+1,p)}{S(t-i,p)} = \frac{(1-p)^{t-i+1}}{(1-p)^{t-i}} = 1-p$$

and similarly

$$\frac{S(t-j,p) - S(t-j+1,p)}{S(t-j,p)} = 1 - \frac{S(t-j+1,p)}{S(t-j,p)} = 1-(1-p) = p,$$

both for all (occurring) values of t, i, j.

In line with the conclusions of Section 4 we combine the changepoint method of Eckley et al. (2011) with a Gaussian process model for each segment. Hence, the marginal likelihood $Q(\cdot)$ implemented for the algorithm (see equations (3.8) and (3.16)) is

$$Q(y_{s:t};\psi) = p(y_{s:t}|\psi) = N(y_{s:t};\mathbf{0}, K_y),$$

cf. equation (4.4). We use mean zero and a linear covariance function

$$k_y(x_j, x_k) = \psi_0 + \psi_1 x_j x_k + \delta_{jk}\psi_2$$

from equations (4.1) and (4.7) with parameters $\psi = (\psi_0, \psi_1, \psi_2)$ to define

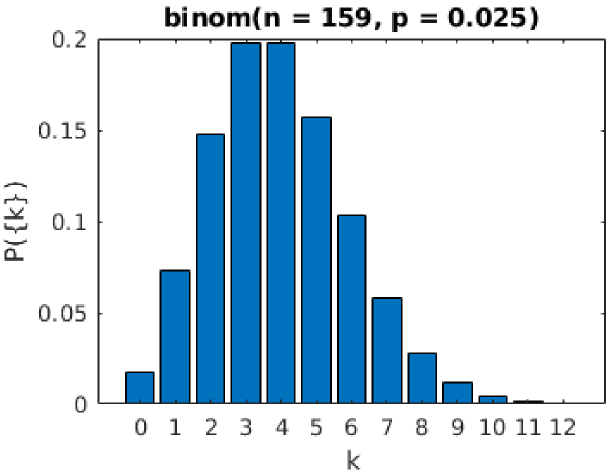


Figure 5.2: Probability mass function of a binomial distribution with parameters $n = 159$ and $p = 0.025$.

the covariance matrix K_y for our analysis. The noise variance σ^2 is now denoted by ψ_2 for notational uniformity.

Our treatment of p, the parameter of the prior survival function, is based on the discussion of prior information in Section 5.1.1: We set $p := 0.025$, defining the binomial prior distribution shown in Figure 5.2. The mean np and mode $\lfloor (n+1)p \rfloor$ of this prior distribution approximately match the result of $m = 4$ boundaries (that is, 5 segments) suggested by Standing and Burland (2006). An almost identical prior probability is assigned to the case of $m = 3$ when the subdivision of layer B is not assumed. Positive but decreasing probabilities are assigned to segmentations with fewer or more boundaries, the latter being suggested for instance by the analysis of the potentially similar water content profile from Heathrow Terminal 5 (Hight et al., 2003).

The parameters ψ_0, ψ_1, ψ_2 of the covariance function enter the algorithm through the marginal likelihood $p(y_{s:t}|\psi)$. Whenever this likelihood needs to be evaluated we first execute an optimisation step to find the value of ψ that maximises $p(y_{s:t}|\psi)$ for the current segment $y_{s:t}$ of data. Hence, the value of ψ depends on, and is generally different for, each segment of data $y_{s:t}$.

5.3 Results

This section presents the results of the statistical method recapped in Section 5.2, which provides us with $N = 10000$ independent samples from the posterior distribution of m and $\tau_1, \dots, \tau_m$. The generated samples contain a substantial amount of information. For the presentation of results we rely on summary statistics to grasp this information. The reader of this thesis might have interest in particular details or questions that can be addressed using these samples. Because it is impossible to foresee and cover these I focus on those that seem most interesting or are common in the literature. We start by considering the posterior distribution $p(m|y_{1:n})$ of the number of changepoints. Because m is a key quantity of interest this is presumably the most obvious choice and, due to being univariate, also the simplest quantity to consider. Then we discuss how to draw inference about the changepoint locations which is a much more involved issue.

5.3.1 Marginal posterior of only the number of changepoints

The histogram in Table 5.2 approximates the marginal posterior distribution $p(m|y_{1:n})$ of the number of changepoints m. It is visualised in Figure 5.3 where each bar shows the number of samples with the respective number of changepoints. Since m is a one-dimensional variable the histogram presented in Figure 5.3 provides a good summary of its posterior distribution. The mode shows that most samples (2869 out of 10000) are found to have 7 changepoints,

Table 5.2: Approximate posterior probabilities $p_k := P(m = k|y_{1:n})$ in % for different numbers of changepoints in the data. Probabilities are zero for $k < 2$ and for $k > 13$. Approximations are based on $N = 10000$ samples. Numbers are visualised in Figure 5.3.

k	2	3	4	5	6	7	8	9	10	11	12	13
p_k	0.15	1.59	5.67	16.64	28.12	28.69	12.85	4.80	1.24	0.20	0.04	0.01

$$P(m = 7|y_{1:n}) \approx \frac{2869}{10000} = 28.69\%,$$

making $\hat{m} = 7$ the MAP estimate for m. The mean and the median of the posterior distribution $P(m|y_{1:n})$ are given by 6.4171 and 6, respectively.

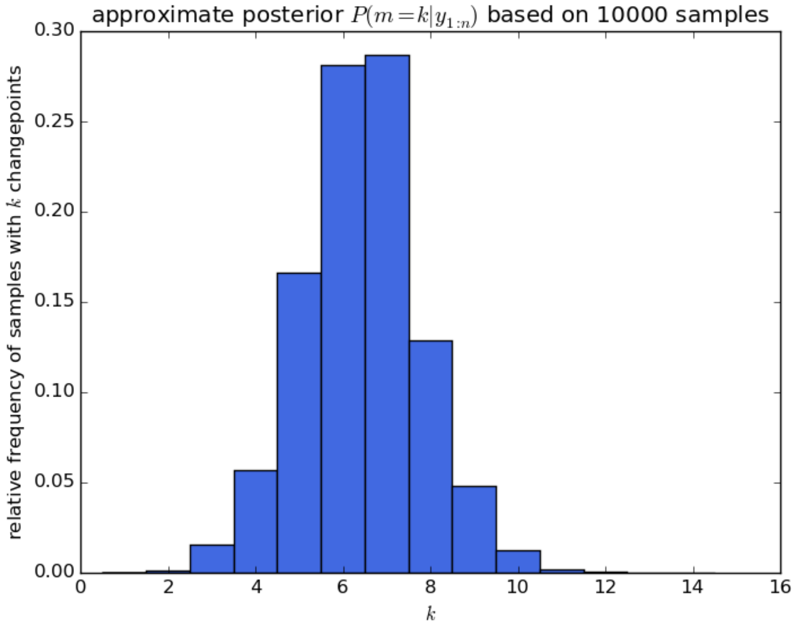


Figure 5.3: A visualisation of Table 5.2: A histogram showing the relative frequencies of samples with $m = k$ changepoints, giving an approximation to the posterior distribution $p(m|y_{1:n})$.

5.3.2 Posterior distribution of the changepoint locations

Analysing the posterior distribution of the changepoint locations $\tau_{1:m}$ is difficult. The dimension of $\tau_{1:m}$ changes with the value of m and even when we fix the value of m it is difficult to summarise m-dimensional distributions for $m > 2$. Note that the difficulty we are discussing is regarding the question of how to summarise the comprehensive information comprised by the samples. Answering particular questions or obtaining individual quantities such as the joint MAP estimate for $(m, \tau_{1:m})$ is straightforward. However, only considering a MAP estimate skips the procedure of understanding the information about m and $\tau_{1:m}$ that is available to us. The MAP estimate is the result of considering all possible values of $(m, \tau_{1:m})$ and selecting the one value of $(m, \tau_{1:m})$ that has the largest posterior probability. While all available information goes into this selection process, the only result is a single value whereas most of the information is discarded.

Instead of the indices $\tau_1, \dots, \tau_m$ the following presentation often uses the equivalent input depths $x_{\tau_1}, \dots, x_{\tau_m}$ of the dataset. The latter has the advantage of being interpretable while the former provides simpler notation.

A one-dimensional summary statistic: the probability of a changepoint at t

Following Nam (2013, Section 3.2) we consider

$$P(t \in \{\tau_1, \tau_2, \ldots, \tau_m\} | y_{1:n}), \tag{5.2}$$

the posterior probability of there being a changepoint at x_t, in Figure 5.4. This one-dimensional function of t (or x_t) is useful to get a first idea regarding the distribution of changepoints in the dataset but ignores any dependence structure between changepoint locations. As an example, consider two indices t_1, t_2 with high probabilities $P(t_j \in \{\tau_1, \tau_2, \ldots, \tau_m\} | y_{1:n})$, for example $x_{t_1} = 10.6$ and $x_{t_2} = 20.55$. Despite the high probabilities for each index, they might have zero probability of occurring as changepoints together. This would mean that, while it is very likely to have a boundary

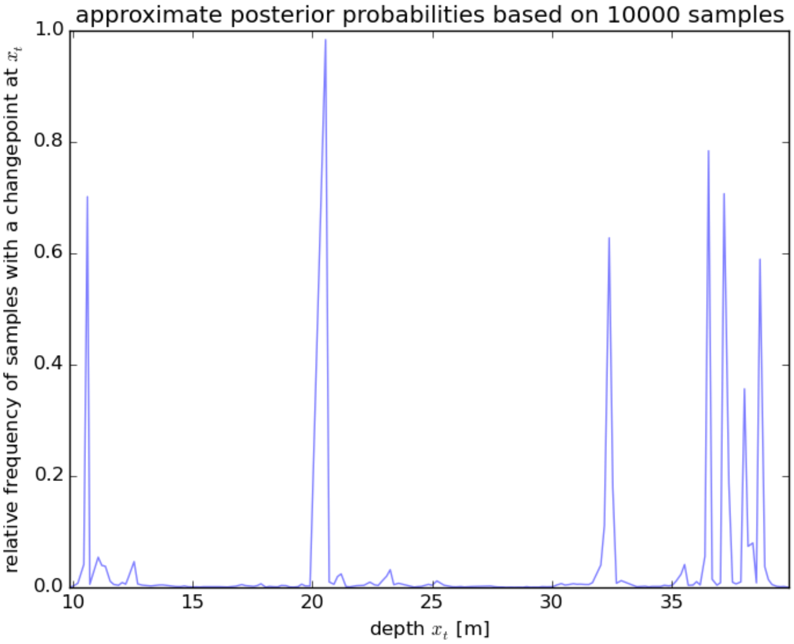


Figure 5.4: A summarising representation of the posterior distribution of the changepoint locations $\tau_{1:m}$. The horizontal axis shows the depth x_t, the vertical axis the approximate posterior probabilities $P(t \in \{\tau_1, \tau_2, \ldots, \tau_m\} | y_{1:n})$ of a changepoint occurring at x_t.

at each of the two locations $x_{t_1} = 10.6$ and $x_{t_2} = 20.55$, it is very unlikely that we have boundaries at both locations.

Results presented by conditioning on different values of m

In order to break down the information comprised by the samples we proceed by considering information about the locations $\tau_{1:m}$ for one fixed value of m at a time.

Figure 5.5 shows $P(t \in \{\tau_1, \tau_2, \ldots, \tau_m\} | m = k, y_{1:n})$, the aforementioned probability (5.2), but conditional on a fixed number of changepoints, $|m = k$. We present results for those values k that have the highest posterior probability $P(m = k|y_{1:n})$ as seen in Figure 5.3.

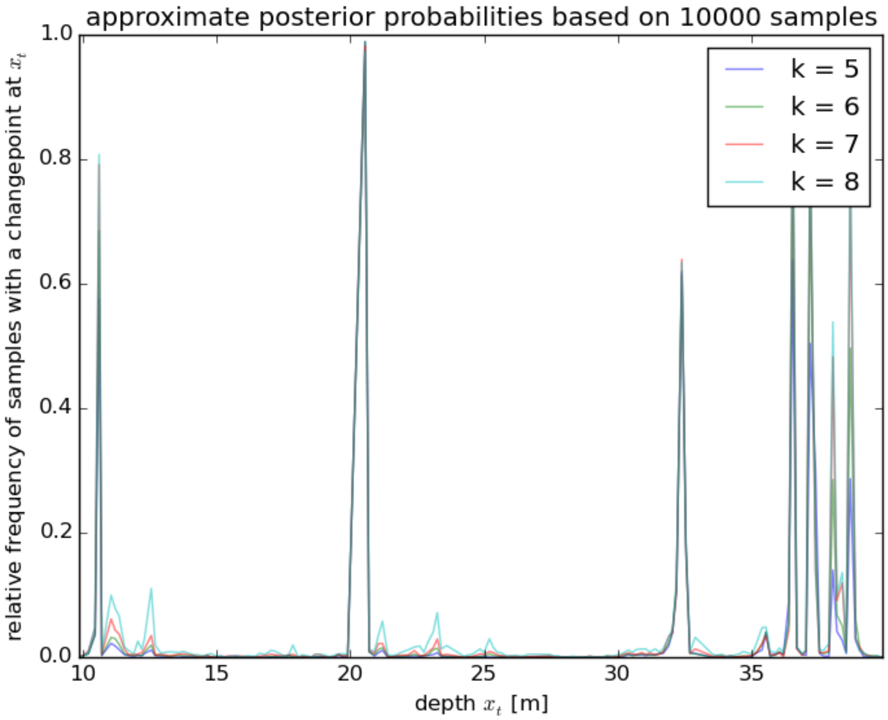


Figure 5.5: A summarising representation of the posterior distribution of the changepoint locations $\tau_{1:m}$. The horizontal axis shows the depth x_t, the vertical axis shows the conditional probabilities $P(t \in \{\tau_1, \tau_2, \ldots, \tau_m\}|m = k, y_{1:n})$, displayed for $k \in \{5, 6, 7, 8\}$.

Starting with the most common value $m = 7$ (cf. Figure 5.3) we consider only those samples with 7 changepoints in Table 5.3. These samples

Table 5.3: Description of the 2869 (out of 10000) samples that have 7 changepoints. This table presents the 8 most common configurations of 7 changepoints together with the number of times ("count") each configuration $x_{t_{1:7}}$ occurred. It also shows the relative frequency (in %) obtained as the fraction of count divided by 2869.

count	relative frequency	x_{t_1}	x_{t_2}	x_{t_3}	x_{t_4}	x_{t_5}	x_{t_6}	x_{t_7}
474	16.52	10.6	20.55	32.4	36.55	37.2	38.05	38.7
131	4.57	10.6	20.55	32.55	36.55	37.2	38.05	38.7
112	3.90	10.6	20.55	32.4	36.55	37.2	38.4	38.7
104	3.62	10.6	20.55	32.2	36.55	37.2	38.05	38.7
83	2.89	10.6	20.55	32.4	36.55	37.2	38.2	38.7
40	1.39	10.6	20.55	32.55	36.55	37.2	38.4	38.7
36	1.25	10.45	20.55	32.4	36.55	37.2	38.05	38.7
32	1.12	10.6	20.55	32.05	36.55	37.2	38.05	38.7

alone provide an approximation to the m-dimensional conditional posterior distribution $p(\tau_{1:m}|y_{1:n}, m = 7)$. Once we have studied this distribution for different values of m we can use the gained insights and compare our findings between different values of m.

From Table 5.3 we obtain the following results: The most commonly sampled configuration of changepoints $x_{t_{1:7}} = (10.6, 20.55, 32.4, 36.55, 37.2, 38.05, 38.7)$ occurs 474 out of 10000 times yielding the joint posterior probability

$$P(m = 7, \tau_{1:m} = t_{1:7}|y_{1:n}) \approx \frac{474}{10000} = 4.74\%;$$

or the conditional (on $m = 7$) posterior probability

$$P(\tau_{1:m} = t_{1:7}|m = 7, y_{1:n}) \approx \frac{474}{2869} \approx 16.52\%.$$

The latter could also be computed as

$$\begin{aligned} P(\tau_{1:m} = t_{1:7}|m = 7, y_{1:n}) &= \frac{P(m = 7, \tau_{1:m} = t_{1:7}|y_{1:n})}{P(m = 7|y_{1:n})} \\ &\approx \frac{474/10000}{2869/10000}. \end{aligned}$$

While there are approximately $\binom{n}{m} = \binom{159}{7} > 10^{11}$ possible values for $\tau_{1:m}$ only 1280 of these values occur amongst the 10000 samples. The 8 most

Table 5.4: Description of the 2812 (out of 10000) samples that have 6 changepoints. This table presents the 6 most common configurations of 6 changepoints together with the number of times ("count") each configuration $x_{t_{1:6}}$ occurred. It also shows the relative frequency (in %) obtained as the fraction of count divided by 2812.

count	relative frequency	x_{t_1}	x_{t_2}	x_{t_3}	x_{t_4}	x_{t_5}	x_{t_6}
317	11.27	10.6	20.55	32.4	36.55	37.2	38.7
164	5.83	10.6	20.55	32.4	36.55	37.2	38.05
159	5.65	20.55	32.4	36.55	37.2	38.05	38.7
97	3.45	10.6	20.55	32.55	36.55	37.2	38.7
67	2.38	10.6	20.55	32.2	36.55	37.2	38.7
63	2.24	10.6	20.55	32.4	36.55	37.2	38.9

common values are presented in Table 5.3. The 8 presented values are the only values that were sampled at a relative frequency of at least 1%, while most values (1071 out of 1280, that is, 83.67%) occur only once.

Based on the plot of the raw data in Figure 5.1 it is already possible to identify the depths of approximately 11, 20 and 32 metres as candidates for changepoint locations. Figure 5.5 endorses this guess. It comes as no surprise that the most common changepoints configurations in Tables 5.3 (and 5.4) reflect this insight.

Most sampled configurations of changepoints are very similar. The first and second most common value in Table 5.3, for example, only differ in the third coordinate x_{t_3}: 32.4 vs. 32.55. This observation indicates uncertainty in regard to the exact location of $x_{\tau_3}|m = 7$, which can, if it should be of interest, be inferred about in more detail.

Table 5.4 shows similar details as Table 5.3 but for the median $m = 6$ of the posterior distribution. The event of $m = 6$ changepoints also has the second largest posterior probability $P(m = 6|y_{1:n}) \approx \frac{2812}{10000} = 28.12\%$ as visible in Figure 5.3. Out of the $\binom{159}{6}$ possible values of $\tau_{1:m}|m = 6$ only 1001 have positive probability based on the samples. Again, most values (773 out of 1001 $\approx$ 77%) occur just once.

The third most common configuration of 6 changepoints puts no boundary layer at a depth of 10.6 metres. This is particularly remarkable as this configuration has a total probability of $\frac{159}{10000}$, higher than for any configuration with $m = 7$ changepoints except the most common one with $\frac{474}{10000}$.

5.4 Discussion

This section presented a practical application of the Gaussian process changepoint method introduced in this thesis. The state-of-the-art approach is mathematically rigorous and explicitly addresses the uncertainty that is inherent to any such analysis. In order to enable the comparison of my results with those of existing analyses my analysis was performed on an established dataset, the water content profile of the core sample from borehole 1 at St James's Park. Following common engineering practice my method incorporates a priori knowledge about the number of changepoints in the data. In line with the overall mathematically rigorous method and treatment of uncertainty this knowledge is encoded in the prior distribution of the Bayesian model.

The presented results demonstrate that the proposed focus on mathematical rigour and uncertainty quantification is very much needed. The key findings include:

- In agreement with all existing analyses the rather striking changes around the depths of 20 and 32 metres are captured.

- Our results propose that the number of changepoints m is higher than all previous analyses suggest.

- The variability of the posterior distribution $p(m|y_{1:n})$ (Figure 5.3) shows that there is a significant amount of uncertainty regarding the number of changepoints. This suggests that the standard presented in the literature (Section 5.1.1) is questionable.

- We conclude that a further subdivision of layer A3 (=20 to 32 metres) as it has been suggested by several authors is not supported by the present dataset. In particular is the possible changepoint at a depth of 27.9 metres in the results of Wang et al. (2014a) (explicitly alluded to by Wang et al. (2014b)) no tenable statistical result.

- Our analysis draws attention to a change at around 10.6 metres. This change in the data, also visible by eye, has not been regarded by other authors. However, it might be possible to link it to further subsections identified by Hight et al. (2007) in their analysis of a water content profile from the site of Heathrow Terminal 5 at which more of the top

B2 soil layer is present in the ground (that is, further measurements at depths $< x_1 = 9.85$ metres are available).

- Our method provides an unprecedented amount of information that can be considered in detail. One example is the third most probable segmentation in Table 5.4, which is the fifth most probable segmentation overall (that is, under the posterior distribution $p(m, \tau_{1:m}|y_{1:n})$) and does not include a changepoint at a depth of $x_t = 10.6$, cf. discussion at the end of Section 5.3.2.

5.4.1 Future work

Besides the changepoint at 10.6 metres our results suggest multiple closely spaced changepoints towards the deep end of the dataset (36.55 to 38.9 metres). An inspection of the data in this region by eye gives little insight. This suggests two lines for future work: On the one hand our results motivate further engineering research to scrutinise the soil and its geotechnical properties in these regions. On the other hand the assumptions that the presented results are based on should be investigated further. The estimation of the parameter ψ for example (described at the end of Section 5.2) is based on maximisations of the likelihood and thus takes the data into account. This optimisation step might lead to overfitting and could have impacted results such as the closely spaced changepoints between 36.55 and 38.9 metres.

Further future work, constituting a major step in geotechnical engineering as well as statistical research, is to establish a joint model for sequences of data obtained either from multiple nearby boreholes or multiple sequences of measurements obtained from the same borehole. The detail of interest is that the number of soil layers is identical and that the thicknesses of most layers are similar across sequences. Taking additional core samples into account provides further information that can improve our analysis. In particular we can improve our estimates of m and $\tau_{1:m}$ as well as our quantification of the uncertainty in the estimates. For example, sharing information between individual boreholes can help to identify outliers that might otherwise be identified as changepoints. Similarly, Wang et al. (2014b) point out that there could be a changepoint that is only detectable in some of the core samples.

Bibliography

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Dover Publications, 1965.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

I. A. Almosallam, M. J. Jarvis, and S. J. Roberts. GPz: non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 462(1):726–739, 2016.

J. A. D. Aston, J. Y. Peng, and D. E. K. Martin. Implied distributions in multiple change point problems. *Statistics and Computing*, 22(4):981–993, 2012. doi: 10.1007/s11222-011-9268-6.

D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

D. Barry and J. A. Hartigan. A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421):309–319, 1993.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

J. V. Braun, R. K. Braun, and H. G. Müller. Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87:301–314, 2000.

E. Brodsky and B. S. Darkhovsky. *Nonparametric methods in change point problems*, volume 243. Springer Science & Business Media, 2013.

Y. Cao and D. J. Fleet. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014.

I. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models.* Springer Science & Business Media, 2006.

B. P. Carlin, A. E. Gelfand, and A. F. M. Smith. Hierarchical Bayesian analysis of changepoint problems. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):389–405, 1992.

George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167, 1992.

J. Chen and A. K. Gupta. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance.* Springer Science & Business Media, 2011.

S. Chib. Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, 75(1):79–97, 1996.

S. Chib. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241, 1998.

R. A. Davis, T. C. M. Lee, and G. A. Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239, 2006.

M. De Freitas and W. Mannion. A biostratigraphy for the London Clay in London. *Géotechnique*, 57(1):91–99, 2007.

M. P. Deisenroth and J. W. Ng. Distributed Gaussian processes. In *International Conference on Machine Learning*, 2015. URL `http://proceedings.mlr.press/v37/deisenroth15.pdf`.

I. A. Eckley, P. Fearnhead, and R. Killick. Analysis of changepoint models. *Bayesian Time Series Models*, pages 205–224, 2011.

C. Erdman and J. W. Emerson. A fast Bayesian change point analysis for the segmentation of microarray data. *Bioinformatics*, 24(19):2143–2148, 2008.

C. Erdman, J. W. Emerson, et al. bcp: an R package for performing a Bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13, 2007.

P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2):203–213, 2006.

P. Fearnhead. Computational methods for complex stochastic systems: a review of some alternatives to MCMC. *Statistics and Computing*, 18(2): 151–171, 2008.

P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

P. Fryzlewicz et al. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721, 1984.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

P. J. Green. Trans-dimensional Markov chain Monte Carlo. *Oxford Statistical Science Series*, pages 179–198, 2003.

J. A. Hartigan. Partition models. *Communications in Statistics – Theory and Methods*, 19(8):2745–2756, 1990.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97, 1970.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence, UAI 2013*, 2013.

J. Hensman, N. Durrande, and A. Solin. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18:151–1, 2017.

D. W. Hight, F. McMillan, J. J. M. Powell, R. J. Jardine, and C. P. Allenou. Some characteristics of London clay. *Characterisation and engineering properties of natural soils*, 2:851–946, 2003.

D. W. Hight, A. Gasparre, S. Nishimura, N. A. Minh, R. J. Jardine, and M. R. Coop. Characteristics of the London Clay from the Terminal 5 site at Heathrow Airport. *Géotechnique*, 57(1):3–18, 2007.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

T. N. Hoang, Q. M. Hoang, and B. K. H. Low. A unifying framework of anytime sparse Gaussian process regression models with stochastic variational inference for big data. In *International Conference on Machine Learning*, pages 569–578, 2015.

B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai. An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE*, 12(2):105–108, 2005.

B. S. Jensen, J. B. Nielsen, and J. Larsen. Bounded gaussian process regression. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pages 1–6. IEEE, 2013.

D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Pearson Prentice Hall, 2009. ISBN 9780131873216.

C. G. Kaufman, M. J. Schervish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008.

E. Khmaladze. Statistical analysis of electricity prices. *Journal of Data Science*, 5(1):103–129, 2007.

R. Killick and I. Eckley. changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19, 2014.

R. Killick, P. Fearnhead, and I. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

C. King. *The stratigraphy of the London Clay and associated deposits*. Backhuys, 1981.

S. Lee, Y. Nishiyama, and N. Yoshida. Test for parameter change in diffusion processes by cusum statistics based on one-step estimators. *Annals of the Institute of Statistical Mathematics*, 58(2):211–222, 2006.

H. Liu, Y.-S. Ong, X. Shen, and J. Cai. When Gaussian process meets big data: A review of scalable GPs. *arXiv preprint arXiv:1807.01065*, 2018.

H. Liu, J. Cai, Y.-S. Ong, and Y. Wang. Understanding and comparing scalable Gaussian process regression for big data. *Knowledge-Based Systems*, 164:324–335, 2019.

T. M. Luong, Y. Rozenholc, and G. Nuel. Fast estimation of posterior probabilities in change-point analysis through a constrained hidden Markov model. *Computational Statistics & Data Analysis*, 68:129–140, 2013.

R. Maidstone. *Efficient analysis of complex changepoint problems*. PhD thesis, Lancaster University, 2016.

R. Maidstone, T. Hocking, G. Rigaill, and P. Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, pages 1–15, 2016.

D. S. Matteson and N. A. James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

S. E. Minson, J. R. Murray, J. O. Langbein, and J. S. Gomberg. Real-time inversions for finite fault slip models and rupture geometry based on high-rate GPS data. *Journal of Geophysical Research: Solid Earth*, 119 (4):3201–3231, 2014.

C. Nam. *The uncertainty of changepoints in time series.* PhD thesis, University of Warwick, 2013.

C. F. H. Nam, J. A. D. Aston, and A. M. Johansen. Quantifying the uncertainty in change points. *Journal of Time Series Analysis*, 33(5):807–823, 2012.

C. F. H. Nam, J. A. D. Aston, and A. M. Johansen. Parallel sequential Monte Carlo samplers and estimation of the number of states in a Hidden Markov Model. *Annals of the Institute of Statistical Mathematics*, 66(3): 553–575, 2014.

C. F. H. Nam, J. A. D. Aston, I. A. Eckley, and R. Killick. The uncertainty of storm season changes: Quantifying the uncertainty of autocovariance changepoints. *Technometrics*, 57(2):194–206, 2015.

R. M. Neal. Monte carlo implementation of gaussian process models for bayesian regression and classification. Technical report, Technical Report 9702, Department of Statistics, 1997.

A. O'Hagan. *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference.* 1994. ISBN 0 340 52922 9.

M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

H. Pantelidou and B. Simpson. Geotechnical variation of London Clay across central London. *Géotechnique*, 57(1):101–112, 2007.

H. Peng, S. Zhe, X. Zhang, and Y. Qi. Asynchronous distributed variational Gaussian process for regression. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2788–2797, 2017.

A. N. Pettitt. A non-parametric approach to the change-point problem. *Applied statistics*, pages 126–135, 1979.

J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

C. E. Rasmussen and H. Nickisch. Gaussian process regression and classification toolbox version 4.1 for GNU Octave 3.2.x and Matlab 7.x. `http://www.gaussianprocess.org/gpml/code/matlab/doc/`, 2005 – 2017.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006. ISBN 9780262182539.

G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *arXiv preprint arXiv:1004.0887*, 2010.

G. Rigaill. A pruned dynamic programming algorithm to recover the best segmentations with 1 to Kmax change-points. *arXiv preprint arXiv:1004.0887v2*, 2015.

Christian Robert and George Casella. *Monte Carlo statistical methods.* Springer, 2nd ed. edition, 2004.

M. N. Schmidt and M. Morup. Nonparametric bayesian modeling of complex networks: An introduction. *IEEE Signal Processing Magazine*, 30(3):110–128, 2013.

B. Schölkopf, A. J. Smola, F. Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2002.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

S. L Scott. Bayesian methods for hidden Markov models. *Journal of the American Statistical Association*, 97(457):337–351, 2002. doi: 10.1198/016214502753479464.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

J. R. Standing and J. B. Burland. Unexpected tunnelling volume losses in the Westminster area, London. *Géotechnique*, 56(1):11–26, 2006.

M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

V. Tresp. A Bayesian committee machine. *Neural computation*, 12(11): 2719–2741, 2000.

Y. Wang, K. Huang, and Z. Cao. Bayesian identification of soil strata in London Clay. *Géotechnique*, 64(3):239–246, 2014a. doi: 10.1680/geot.13.T.018.

Y. Wang, K. Huang, Z. Cao, J. R. Standing, and B. Calderhead. Bayesian identification of soil strata in London Clay. *Géotechnique*, 64(12):1014–1016, 2014b. doi: 10.1680/geot.14.D.004.

C. K. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models*, pages 599–621. Springer, 1998.

A. Wilson and R. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

Y. Yao. Estimating the number of change-points via Schwarz'criterion. *Statistics & Probability Letters*, 6(3):181–189, 1988.

Appendix A

Mathematical details

A.1 Standard Bayesian linear regression with unknown variance

Following Minson et al. (2014) consider the linear model

$$y = X\beta + \eta,$$

with observations vector $y \in \mathbb{R}^{n\times 1}$, design matrix $X \in \mathbb{R}^{n\times r}$, weights $\beta \in \mathbb{R}^{r\times 1}$, and random noise vector $\eta \in \mathbb{R}^{n\times 1}$. We assume

$$\eta \sim N(\mathbf{0}, \sigma^2 D),$$

that is, a normal linear model with zero-mean and covariance matrix $D \in \mathbb{R}^{n\times n}$. We thus have a parametric model $y|\theta \sim N(X\beta, \sigma^2 D)$ with model parameter $\theta = (\beta, \sigma^2)$.

Taking a fully Bayesian approach we assume the weights β and the noise variance σ^2 to be unknown and specify a prior distribution. To obtain a conjugate model we employ a conditional normal prior

$$p(\beta|\sigma^2) = N(\beta; \mu, \sigma^2 V)$$

with mean vector μ and a positive definite covariance matrix $V \in \mathbb{R}^{r\times r}$, as well as an inverse-gamma prior $IG(a, d)$ with positive shape parameter

$a > 0$ and rate parameter $d > 0$:

$$\begin{aligned} p(\sigma^2) &= IG(\sigma^2; a, d) \\ &= \frac{(a/2)^{d/2}}{\Gamma(d/2)} \left(\sigma^2\right)^{-\frac{d+2}{2}} \exp\left(-\frac{a}{2\sigma^2}\right). \end{aligned}$$

The joint prior distribution

$$\begin{aligned} p(\theta) &= p(\beta, \sigma^2) \\ &= p(\beta|\sigma^2) p(\sigma^2) \\ &= N(\beta; \mu, \sigma^2 V) IG(\sigma^2; a, d) \\ &=: NIG(\beta, \sigma^2; a, d, \mu, V) \end{aligned}$$

is known as a *normal-inverse-gamma* (NIG) distribution[1] with parameters a, d, μ, V. As a result of the conjugate model the posterior distribution is (Minson et al., 2014, Appendix B)

$$p(\theta|y) = NIG(\beta, \sigma^2; \tilde{a}, \tilde{d}, \tilde{\mu}, \tilde{V})$$

with posterior parameters

$$\begin{aligned} \tilde{V} &:= (V^{-1} + X^T D^{-1} X)^{-1} \in \mathbb{R}^{r \times r}, \\ \tilde{\mu} &:= \tilde{V}(V^{-1}\mu + X^T D^{-1} y) \in \mathbb{R}^{r \times 1}, \\ \tilde{a} &:= a + \mu^T V^{-1} \mu + y^T D^{-1} y - \tilde{\mu}^T \tilde{V}^{-1} \tilde{\mu} \in \mathbb{R}_{>0}, \\ \tilde{d} &:= d + n. \end{aligned}$$

Only $\tilde{a}$ depends on the observations y while all other quantities are constant with respect to y.

The marginal likelihood is given by

$$\begin{aligned} p(y) &= \frac{1}{(2\pi)^{n/2} \det(D)^{1/2}} \frac{\det(\tilde{V})^{1/2}}{\det(V)^{\frac{1}{2}}} \frac{(a/2)^{d/2}}{(\tilde{a}/2)^{\tilde{d}/2}} \frac{\Gamma(\tilde{d}/2)}{\Gamma(d/2)} \\ &= \left(\frac{1}{\pi^n \det(D)} \frac{\det(\tilde{V})}{\det(V)} \frac{a^d}{\tilde{a}^{\tilde{d}}} \right)^{1/2} \frac{\Gamma(\tilde{d}/2)}{\Gamma(d/2)}. \end{aligned}$$

[1] A different parametrisation of the inverse-gamma distribution is given by a scaled-inverse-chi-squared distribution. The resulting normal-inverse-chi-squared (NIX) distribution is also used in the literature as the conjugate model for a normal linear regression.

For independent noise ($D = I$) the predictive distribution of unobserved outputs $y^* = (y_1^*, \ldots, y_q^*)^T$ corresponding to a design matrix $X^* \in \mathbb{R}^{q \times r}$ is (O'Hagan, 1994, Section 9.39)[2] a multivariate t-distribution,

$$\begin{aligned} p(y^*|X^*, X, y) &= t_\nu(y^*; \mu^*, \Sigma) \\ &= \frac{\Gamma((\nu+q)/2)}{\Gamma(\nu/2)\nu^{q/2}\pi^{q/2}|\Sigma|^{1/2}} \left(1 + \frac{1}{\nu}(y^* - \mu^*)^T \Sigma^{-1}(y^* - \mu^*)\right)^{-\frac{\nu+q}{2}} \\ &= \frac{\Gamma((\nu+q)/2)}{\Gamma(\nu/2)\pi^{q/2}|\nu\Sigma|^{1/2}} \left(1 + (y^* - \mu^*)^T (\nu\Sigma)^{-1}(y^* - \mu^*)\right)^{-\frac{\nu+q}{2}}, \end{aligned} \tag{A.1}$$

with $\nu = \tilde{d}$ degrees of freedom, location parameter $\mu^* = X^*\tilde{\mu}$ and covariance parameter $\Sigma = \frac{\tilde{a}}{\tilde{d}}(\mathbf{I}_q + X^*\tilde{V}(X^*)^T))$. Mean and (co)variances of the predictive distribution are given by

$$E(y^*|X^*, X, y) = \mu^* = X^*\tilde{\mu} \text{ and } \mathrm{Cov}(y_j^*, y_k^*|X^*, X, y) = \tilde{d}/(\tilde{d} - 2)\Sigma_{j,k}.$$

For the engineering problem considered in Section 5 the equations simplify to

$$y_j = \beta_0 + \beta_1 x_j + \eta_j$$

with independent noise ($D = I_n$) and a zero-mean independent normal prior on the weights $\beta = (\beta_0, \beta_1)^T$, that is, $\mu = \mathbf{0}$ and $V = \mathrm{diag}(\sigma_0^2, \sigma_1^2)$:

$$p(\beta|\sigma^2) = N(\beta; \mathbf{0}, \sigma^2 \left(\begin{bmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}\right)).$$

[2] O'Hagan (1994) uses the last representation (A.1) to define a multivariate t-distribution with covariance parameter $C := \nu\Sigma$. For this parametrisation the (co)variances are given by $\frac{1}{\nu-2}C_{j,k}$.

As a result the above formulas simplify to

$$\begin{aligned}\tilde{V} &:= (V^{-1} + X^T X)^{-1} \in \mathbb{R}^{2\times 2},\\ \tilde{\mu} &:= \tilde{V} X^T y \in \mathbb{R}^{2\times 1},\\ \tilde{a} &:= a + y^T y - \tilde{\mu}^T \tilde{V}^{-1} \tilde{\mu}\\ &= a + y^T y - \tilde{\mu}^T X^T y\\ &= a + (y - X\tilde{\mu})^T y \in \mathbb{R}_{>0},\\ \tilde{d} &:= d + n\end{aligned}$$

and

$$p(y) = \left(\frac{1}{\pi^n} \frac{\det(\tilde{V})}{\det(V)} \frac{a^d}{\tilde{a}^{\tilde{d}}} \right)^{1/2} \frac{\Gamma(\tilde{d}/2)}{\Gamma(d/2)}$$

with

$$\begin{aligned}\det(V) &= \sigma_0^2 \sigma_1^2,\\ V^{-1} &= \operatorname{diag}(1/\sigma_0^2, 1/\sigma_1^2),\\ X &= \begin{pmatrix} 1 & \dots 1 \\ x_1 & \dots x_n \end{pmatrix}^T,\\ X^T X &= \begin{pmatrix} n & \sum_{j=1}^n x_j \\ \sum_{j=1}^n x_j & \sum_{j=1}^n x_j^2 \end{pmatrix},\\ X^T y &= \begin{pmatrix} \sum_{j=1}^n y_j \\ \sum_{j=1}^n x_j y_j \end{pmatrix},\\ y^T y &= \sum_{j=1}^n y_j^2.\end{aligned}$$

Finally, in the notation of Section 5, let $\theta = (w_0, w_1, \sigma^2)$, $\psi = (a, d, \sigma_0^2, \sigma_1^2)$,

$$p(y_{s:t}|\theta) = \prod_{k=s}^{t} \phi(y_k; w_0 + w_1 x_k, \sigma^2)$$

and

$$\begin{aligned} p(\theta|\psi) &= p(w_0, w_1, \sigma^2|a, d, \sigma_0^2, \sigma_1^2) \\ &= NIG\left((w_0, w_1)^T; a, d, \mu = \mathbf{0}, \operatorname{diag}(\sigma_0^2, \sigma_1^2)\right). \end{aligned}$$

We obtain the marginal likelihood

$$\begin{aligned} Q(y_{s:t}; \psi) &= \int p(y_{s:t}|\theta) p(\theta|\psi) \,\mathrm{d}\theta \\ &= \left(\frac{1}{\pi^l} \frac{\det(\tilde{V}_l)}{\det(V)} \frac{a^d}{\tilde{a}_l^{\tilde{d}_l}} \right)^{1/2} \frac{\Gamma(\tilde{d}_l/2)}{\Gamma(d/2)} \end{aligned}$$

required for the algorithm, where now

$$\begin{aligned} l &:= t - s + 1, \\ X_{s:t} &:= \begin{pmatrix} 1 & \dots & 1 \\ x_s & \dots & x_t \end{pmatrix}^T, \\ \tilde{V}_l &:= (V^{-1} + X_{s:t}^T X_{s:t})^{-1} \in \mathbb{R}^{r \times r}, \\ \tilde{\mu}_l &:= \tilde{V}_l X_{s:t}^T y_{s:t} \in \mathbb{R}^{r \times 1}, \\ \tilde{a}_l &:= a + y_{s:t}^T y_{s:t} - \tilde{\mu}_l^T \tilde{V}_l^{-1} \tilde{\mu}_l \\ &= a + y_{s:t}^T y_{s:t} - \tilde{\mu}_l^T X_{s:t}^T y_{s:t} \in \mathbb{R}_{>0}, \\ \tilde{d}_l &:= d + l \end{aligned}$$

and the simple form of $\det(V)$, V^{-1}, $X^T X$, ... still hold. For the implementation we note

$$\begin{aligned} \ln Q(y_{s:t}; \psi) = \frac{1}{2} &\left(-\ln \pi + \ln \det(\tilde{V}_l) - \ln \det(V) + d \ln a - \tilde{d}_l \ln \tilde{a}_l \right) \\ &+ \ln \Gamma(\tilde{d}_l/2) - \ln \Gamma(d/2). \end{aligned}$$

A.2 Mercer's theorem

For any continuous positive definite kernel k Mercer's theorem states the existence of a feature map ϕ so that $k(x, x')$ is given by the inner product $\langle \phi(x), \phi(x') \rangle$. Following Schölkopf et al. (2002) let $(\mathcal{X}, \mathcal{A}, \mu)$ denote a finite measure space and let $L_2(\mathcal{X})$ denote the space of all measurable functions $f : \mathcal{X} \to \mathbb{R}$ for which $\int_{\mathcal{X}} f(x)^2 \,\mathrm{d}\mu(x)$ is finite. Let further $L_\infty(\mathcal{X}^2)$ denote the space of measurable functions $k : \mathcal{X}^2 \to \mathbb{R}$ for which there exist a finite $u \in \mathbb{R}$ such that $|k(x, x')| < u$ for μ^2-almost all $(x, x') \in \mathcal{X}^2$.

Theorem 1 (Theorem 2.10 in Schölkopf et al. (2002)) *Suppose that $k \in L_\infty(\mathcal{X}^2)$ is a symmetric real-valued function such that the integral operator*

$$T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X}),\ (T_k f)(x) := \int_{\mathcal{X}} k(x, x') f(x') \,\mathrm{d}\mu(x')$$

is positive definite*; that is, for all $f \in L_2(\mathcal{X})$, we have*

$$\int_{\mathcal{X}^2} k(x, x') f(x) f(x') \,\mathrm{d}\mu(x) \,\mathrm{d}\mu(x') \geq 0.$$

Let $\psi_j \in L_2(\mathcal{X})$ be the normalised orthogonal eigenfunctions of T_k associated with the eigenvalues $\lambda_j > 0$, sorted in non-increasing order. Then

1. *the supremum $\sup_{j \in \mathbb{N}} |\lambda_j|$ is finite;*
2. *$k(x, x') = \sum_{i=1}^{\mathcal{N}_\mathcal{H}} \lambda_j \psi_j(x) \psi_j(x')$ holds for almost all (x, x'). Either $\mathcal{N}_\mathcal{H} \in \mathbb{N}$, or $\mathcal{N}_\mathcal{H} = \infty$; in the latter case, the series converges absolutely and uniformly for almost all (x, x').*

The feature map to write the kernel as an inner product is then given by $\phi(x) := (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \dots)$.

A.3 Derivation of formulas for Fearnhead's method

This section derives the formulas for the probabilities $\gamma_t(i)$ stated in Section 3.4.4. For presentational simplicity we drop the conditioning on ψ from our notation. In order to derive the iterative formulas used in the forward part of the algorithm let $t = 2, \dots, n$ and $i = 0, \dots, t-2$. The event $C_t = i$ implies $C_{t-1} = i$ and all computations are for non-trivial events $P(C_t = i) > 0$. Fearnhead and Liu (2007) refer to "standard filtering recursions" when stating the relation

$$\begin{aligned}\gamma_t(i) &:= P(C_t = i | y_{1:t}) \\ &= P(C_t = i | y_t, y_{1:t-1}) \\ &\propto P(y_t | C_t = i, y_{1:t-1}) P(C_t = i | y_{1:t-1}),\end{aligned}$$

but it can also be seen as a direct instance of Bayes formula conditional on $y_{1:t-1}$. In order to establish a recursive formula for $\gamma_t(i)$ we express the prior term $P(C_t = i|y_{1:t-1})$ as the product of the prior probability $\gamma_{t-1}(i) = P(C_{t-1} = i|y_{1:t-1})$ and the transition probability $P(C_t = i|C_{t-1} = i)$,

$$\begin{aligned}P(C_t = i|y_{1:t-1}) &= P(C_{t-1} = i|y_{1:t-1})P(C_t = i|C_{t-1} = i, y_{1:t-1})\\ &= P(C_{t-1} = i|y_{1:t-1})P(C_t = i|C_{t-1} = i),\end{aligned}$$

giving

$$\gamma_t(i) \propto \gamma_{t-1}(i)P(y_t|C_t = i, y_{1:t-1})P(C_t = i|C_{t-1} = i).$$

The last two factors can be written as

$$\begin{aligned}P(y_t|C_t = i, y_{1:t-1}) &= \frac{P(y_{1:t}|C_t = i)}{P(y_{1:t-1}|C_t = i)}\\ &= \frac{P(y_{1:i}|C_t = i)P(y_{i+1:t}|C_t = i)}{P(y_{1:i}|C_t = i)P(y_{i+1:t-1}|C_t = i)}\\ &= \frac{P(y_{i+1:t}|C_t = i)}{P(y_{i+1:t-1}|C_t = i)}\\ &= \frac{Q(i+1, t; \psi)}{Q(i+1, t-1; \psi)}\end{aligned}$$

and

$$\begin{aligned}P(C_t = i|C_{t-1} = i) &= \frac{P(C_t = i, C_{t-1} = i)}{P(C_{t-1} = i)}\\ &= \frac{P(C_t = i)}{P(C_{t-1} = i)}\\ &= \frac{S(t-i)}{S(t-1-i)},\end{aligned}$$

where in both cases we do not use the specific choice for likelihood and the prior until the last equality. Similarly, now for $i = t-1$,

$$\begin{aligned}
\gamma_t(t-1) &= P(C_t = t-1|y_{1:t}) \\
&\propto P(y_t|C_t = t-1, y_{1:t-1})P(C_t = t-1|y_{1:t-1}) \\
&= P(y_t|C_t = t-1)\sum_{j=0}^{t-2} P(C_t = t-1|C_{t-1} = j, y_{1:t-1})P(C_{t-1} = j|y_{1:t-1}) \\
&= Q(t,t;\psi)\sum_{j=0}^{t-2}\gamma_{t-1}(j)\frac{g(t-1-j;\psi)}{S(t-1-j;\psi)}.
\end{aligned}$$

A.4 Derivation of an equivalent prior specification

In Section 5.2 we use a geometric distribution to define the prior on the length of a segment via its survival function $S(\cdot)$. We claim that this is equivalent to assigning hierarchical prior distribution to the number of changepoints and their locations. We will now prove this claim. From equation (5.1) we derive the joint prior distribution (3.9) as

$$\begin{aligned}
p(m, \tau_{1:m}) &= S(\tau_{m+1} - \tau_m)\prod_{k=1}^{m} g(\tau_k - \tau_{k-1}) \\
&= (1-p)^{\tau_{m+1}-\tau_m-1}\prod_{k=1}^{m} p(1-p)^{\tau_k-\tau_{k-1}-1} \\
&= p^m(1-p)^{-(m+1)}\prod_{k=1}^{m+1}(1-p)^{\tau_k-\tau_{k-1}} \\
&= p^m(1-p)^{-(m+1)}(1-p)^{\tau_{m+1}-\tau_0} \\
&= p^m(1-p)^{n-m-1}.
\end{aligned}$$

If follows that

$$\begin{aligned}
p(m) &= \sum_{\tau_{1:m}} p(m, \tau_{1:m}) \\
&= \sum_{\tau_{1:m}} p^m(1-p)^{n-m-1} \\
&= \binom{n-1}{m} p^m(1-p)^{n-1-m},
\end{aligned}$$

since $n = \tau_{m+1}$ is not amongst the changepoints $\tau_{1:m}$. This is a binomial distribution, $p(m) \sim B(n-1, p)$. Further,

$$\begin{aligned} p(\tau_{1:m}|m) &= \frac{p(m, \tau_{1:m})}{p(m)} \\ &= \frac{p^m(1-p)^{n-m-1}}{\binom{n-1}{m} p^m (1-p)^{n-1-m}} \\ &= \frac{1}{\binom{n-1}{m}}, \end{aligned}$$

meaning that we have a conditional uniform distribution for the changepoint locations, $p(\tau_{1:m}|m) = 1/\binom{n-1}{m}$.