NTNU
Norwegian University of
Science and Technology

Spatial Dependency in Methylation Data

A Bayesian Approach with R-INLA

Haakon Egdetveit Nustad

Master of Science in Physics and Mathematics
Submission date: March 2016
Supervisor: Ingelin Steinsland, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

DNA methylation is a chemical process that regulates gene transcription and is known to interact with development and differentiation of the DNA. It affects almost exclusively CpG sites, and with the Illumina HumanMethylation450k BeadChip we are able to measure the methylation level for more than 450000 CpG sites in the human DNA. The locations of these CpG sites have been accurately measured to a base pair resolution, making it possible to look into spatial dependencies.

In this paper, we investigate differences in mean between two groups of people by taking the spatial dependency into account. The investigations and analysis is done on a data set containing methylation data from 62 persons classified as having Schizophrenia and 33 Healthy persons. An exploratory analysis have been done, to investigate which assumptions that should be made when analyzing methylation data. Through auto correlation analysis, correlation estimates and regression evaluations, we have seen that the data is influenced by spatial dependencies. With Bayesian regression with Integrated Nested Laplace Approximations(INLA), we have investigated different models to be able to quantify the spatial dependency structure, and in general the underlying structure of the methylation data at a part of chromosome 6. The model that obtained the best fit included spatial dependency and an independently, identically distributed random effect in the linear predictor. The model was optimized using a likelihood that assumed a location independent precision parameter ϕ.

Through simulations, we have seen that a test for differently methylated positions that builds on a model which utilizes the spatial dependency, might lead to better results than a T-test. Still, further studies are required. Some of the results obtained by the simulations deviates from those obtained by the case study, which might indicate the presence of an underlying structure in the methylation data that is not yet quantified.

Sammendrag

DNA metylering er en kjemisk prosess som regulerer gentranskripsjon, og er kjent for å påvirke utviklingen og differensieringen av DNAet. Selve metyleringsprosessen påvirker CpG posisjoner, og med Illumina HumanMethylation-450k BeadChipen har vi mulighet til å måle metyleringsverdien til mer enn 450000 CpG posisjoner i menneskets DNA. Lokasjonen til disse CpG posisjonene i DNAet har blitt målt ned til et basepar nivå, noe som gjør det mulig å undersøke romlige avhengigheter i metyleringsdataen.

I denne artikkelen har vi undersøkt forskjeller i forventningsverdi mellom to grupper i et datasett, ved å ta hensyn til romlige avhengigheter. Undersøkelsene og analysene er utført på et eksempelstudie som inneholder metyleringsdata fra 62 personer karakterisert som schizofrene, og 33 friske. Gjennom en utforskende analyse har vi funnet ut hvilke antagelser vi må gjøre når vi skal modellere metyleringsdataen. Ved å estimere korrelasjoner og analysere regresjoner, har vi sett at metyleringsdataen vi har undersøkt er påvirket av romlige avhengigheter. I et kvalitetsstudie av en del av metyleringsdataen fra kromosom 6, har vi utført Bayesiansk regresjon med Integrerte Nøstede Laplace Approksimasjoner(INLA). Vi har undersøkt flere forskjellige modeller for den lineære prediktoren, for å kvantifisere den romlige avhengighets-strukturen og for å finne den modellen som passer dataene best. Modellen for den lineære prediktoren som passet best inneholdt romlige avhengigheter, samt en uavhengig, identisk distribuert tilfeldig effekt. Modellen ble optimalisert med en likelihood funksjon som antok en lokasjonsuavhengig presisjonsparameter ϕ.

Ved hjelp av simuleringer, har vi sett at en test som bygger på en model som tar hensyn til romlige avhengigheter, kan lede til bedre resultater for å finne forskjellige metylerte posisjoner enn en T-test. Flere simuleringer og videre studier er nødvendig. Noen av resultatene fra simuleringene avviker fra resultatene fra kvalitetsstudiet, noe som kan medføre at metyleringsdataen har en underliggende struktur som ikke enda er kvantifisert.

Contents

Preface

This thesis concludes my Master of Science degree in Applied Physics and Mathematics with specialization in Industrial Mathematics. The work on this thesis has been carried out in the months from October 2015 to March 2016, at the Norwegian University of Science and Technology (NTNU).

First of all, I would like to thank my supervisor Ingelin Steinsland for great support and motivation during these months. Her positive attitude, great ideas and feedback have been very helpful. Further, I would like to thank Heidi Lind-Tviberg for answering many biology related questions, since my knowledge on the topic was limited. Lastly, many thanks to researchers at Marmalaid forum for great support concerning packages for preprocessing methylation data. Recognition should also be given to Jabir Ali Ouassou, for providing the LaTeX template that was used in this thesis.

Finally, I would like to thank Madeleine Vikebø, as well as my family and friends, for encouragement and support throughout the years of study.

Haakon Egdetveit Nustad
March 2016

1 Introduction

DNA methylation(Section 2.1) is a chemical modification of the DNA that plays a key role in regulating gene expression(Mirella Gonzalez-Zulueta and Nguyen 1995). It involves the addition of a methyl group to a cytosine(C) base, which can occur when this base is directly followed by a guanine(G) base along one of the strands in a DNA molecule. This is visualized in Figure 2.2. Places where a C is directly followed by a G are called CpG sites, which tends to cluster in regions to form CpG islands(Gardiner-Garden and Frommer 1987). These islands are strongly connected to gene promoters, which control the activation or inactivation of a gene. Alterations in the DNA methylation at these islands are thought to play an important role in suppression or expression of the associated gene, which is essential in normal development and in development of different diseases(Feinberg 2007).

As a result of DNA methylation's role in development of diseases, it is of great interest to measure the DNA methylation at a comprehensive genomic scale. With the Illumina HumanMethylation450k BeadChip(Bibikova et al. 2011), we are able to measure the methylation level of more than 450000 CpG sites in the human DNA. This method is one of the most used technologies for obtaining methylation data(Aryee et al. 2014), and several projects such as The Cancer Genome Atlas and Marmal-aid have made data publicly available through online data bases. This makes it easy for researchers to obtain data from different experiments, and creates the foundation for rapid advances in methylation research.

The methylation values are in the range of $(0,1)$, denoting averages of methylation at the CpG sites for a given person. A common assumption is that the methylation values for different persons are beta distributed around a CpG site-specific mean, which is equal for persons of the same class. A class denotes people with the same qualities, such as sex, age and disease type. This is further described in Section 2.5.

Several papers have studied the impact of different qualities on the methylation data. An example is a study of Inflammatory Bowel Disease(McDermott et al. 2015) where an algorithm called dmpFinder[1](Hansen and Aryee 2015) is used to find differently methylated CpG sites between a group of people with the disease and a healthy control group. The algorithm is based on a Fisher test, and treats each CpG site individually and independent(Hansen and Aryee 2015).

In this paper, we look into the possibility that the methylation values are affected by spatial dependency. The sequence of base pairs in the human DNA have been

[1] differently methylated positions Finder

very accurately determined by the Human Genome Project(Human Genome Sequencing Consortium 2004), enabling the possibility of locating each of the CpG sites in the DNA along a specific chromosome. This makes it possible to look into the spatial dependency between CpG sites located along a chromosome, which might influence which sites that are differently methylated. We focus on a data set containing methylation values of people classified as having Schizophrenia or being Healthy, and through Bayesian regression we explore the spatial dependency and its effect.

To clarify the assumption of spatial dependency, we consider two different groups of people which we have divide by an illness; one group has Schizophrenia, the other is Healthy. The true mean methylation value at CpG site i for each of these groups are then defined as $\mu_{1,i}$ and $\mu_{2,i}$. If the data from each person are considered as independent of the location i, differences can be easily found by an algorithm such as the dmpFinder or a Student T-test, as described in Section 3.2. However, if each location is not independent, but rather affected by some underlying dependency structure, we have access to a new source of information which is not utilized by these tests. Therefore, we want to investigate the assumption of spatial dependency, and examine the effect of taking it into account with regression.

The main focus when taking the spatial dependency into account, is a randomly chosen part of chromosome 6 of the Schizophrenia data set, described in Section 2.2. By using Bayesian regression with the Integrated Nested Laplace Approximations(INLA)(Rue, Martino, and Chopin 2009), we model the mean of the data through a latent Gaussian model(Rue, Martino, and Chopin 2009) of the linear predictor. To take the spatial dependency into account, we use a Stochastic Partial Differential Equation(SPDE)(Lindgren, Rue, and Lindström 2011) approach to include a spatially dependent effect in the latent Gaussian model. This is further described in Chapter 4. We then look into how this changes the estimates of the differences between the groups, and the fit to the data.

INLA is a deterministic algorithm for Bayesian inference(George E. P. Box 1992), and an alternative to the simulation based Markov Chain Monte Carlo(MCMC) algorithms. It is especially designed for latent Gaussian models, and it provides accurate results in shorter computational time, compared to the MCMC. With the SPDE extension in INLA, we are able to approximate continuous dependency structures(Gaussian Random Fields)(Cressie 1993) by using a discretely indexed spatial random process(Gaussian Markov Random Fields)(Rue and Held 2005a), and hence computational efficient inference is available.

This paper is organized as follows. In Chapter 2, we give background material concerning methylation and data accessibility, as well as presentation of the data used and basic notations. We also review common assumptions when evaluating

methylation data, and the most used statistical models. In Chapter 3, we do an exploratory analysis to find out which features to use when modeling the methylation data. Here we also develop a test for finding differently methylated CpG sites. In Chapter 4, we provide background material relevant for developing models with spatial dependency, as well as a summarized explanation of Bayesian inference and INLA. In Chapter 5, we specify the models we are investigating and relevant information concerning prior specifications. The results obtained by evaluating the models on the Schizophrenia data set are given in Chapter 6. In Chapter 7, a simulation study is provided to explore the results obtained from the case study. Lastly, a discussion and conclusion is given in Chapter 8.

2 Background; Methylation and presentation of data and models

In this Chapter, we give results and explanations from other sources that are central to the paper. Some of these explanations are developed in the specific case of methylation data, while others are more general.

2.1 DNA methylation

The DNA carries most of the genetic information concerning development, functioning and reproduction of all known living organisms. The DNA molecule consists of two biopolymer strands coiled around each other to form a double helix. These strands consist of repetitive nucleotides, containing a phosphate group, a five-carbon sugar and a nitrogen-containing base. The phosphate group and the sugars create

Figure 2.1: DNA; two biopolymer strands coiled around each other to form a double helix. Four unique bases; Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). Picture downloaded from www.astrochem.org.

the backbone of each strand of the DNA double helix. These strands are linked together by the nitrogen-containing bases. There are four types of nitrogen bases

associated with nucleotides in the DNA; cytosine (C), guanine (G), adenine (A), and thymine (T). For each of the bases along one of the strands, there is only one possible complementary base on the other. This results in four distinct combinations; CG, GC, AT and TA. Figure 2.1 shows a visual interpretation of the DNA structure.

Figure 2.2: CpG, 5′-Cytosine-phosphate-Guanine-3′ site on one DNA strand read by the directionality 5′ → 3′(left in picture). Complementary C-G base pairing on two DNA strands(right). Picture downloaded from Wikimedia Commons.

Within the DNA, a process called DNA methylation can occur. This is a process that typically occurs at CpG sites, which are regions in the DNA sequence along one of the strands where a cytosine(C) base is directly followed by a guanine(G) base. CpG stands for 5′-Cytosine-phosphate-Guanine-3′, where phosphate binds any two bases together along a DNA strand, while 5′ → 3′ stands for the end-to-end chemical orientation of the single strand of bases. This is a way of denoting the directionality of the way the DNA molecule is read, meaning the process of chemically measure the sequence of bases in the DNA. Figure 2.2 shows a visualization of a CpG site.

The chemical process of DNA methylation involves a methyl group to be added to a cytosine base, at a CpG site, to form 5-methylcytosine. The change of the molecule cytosine can be seen in Figure 2.3. This change can have an effect on the processing of DNA, and is known to play a central role in various biological processes, such as stem cell differentiation(Meissner 2010), genomic imprinting(Barlow 2011) and inflammation(Martin and Herceg 2012). DNA methylation is probably best understood in the context of cancer biology(Baylin and Jones 2011), where it is clear that aberrant gains and losses of methylation, at tumor suppressor genes[1] and oncogenes[2] respectively, almost always accompany the initiation and progression of tumors(Feinberg and Tycko 2004). This can be related to the fact that high methylation values in promoters can result in silencing(Mirella Gonzalez-Zulueta and Nguyen 1995) of the tumor suppressor genes, which can inhibit the suppression of tumor growth, resulting in tumor progression. The opposite is true for low methylation values, which can result in overexpression of oncogenes that can be related to tumor growth. Characteristic changes in the DNA methylation have also been reported to interact with the development of other diseases(Feinberg 2007), such as Schizophrenia(Rukova et al. 2014). The last mentioned is what we investigate

[1] genes associated with prevention of tumor growth
[2] genes associated with tumor growth

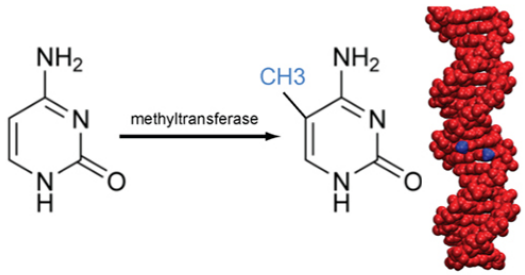


Figure 2.3: Unmethylated(left) and methylated(right) cytosine molecule. Picture downloaded from www.cincinnatichildrens.org.

further, although the main focus is not on finding CpG sites throughout the genome for which the methylation value seems to be different, but on how the differently methylated sites are found.

2.2 Presentation of data

This paper uses measurements from the Illumina Infinum HumanMethylation450k BeadChip. By using bisulfite-converted DNA, two signals of interest are record by interrogating the methylation status of 485512 CpG sites. One signal measures the amount of methylated DNA(Meth), while the other signal measures the amount of unmethylated DNA(Unmeth). In principal, the proportion Meth/(Meth + Unmeth) is the reported methylation value in the population of cells from which the DNA are extracted. This is referred to as a beta value, and is in the range of $(0,1)$. A more extensive description is given in (Bibikova et al. 2011). Illumina Infinum is currently one of the most used technologies for obtaining methylation data(Aryee et al. 2014).

We study methylation data loaded from the online database Marmalaid, which is described in more detail in Section 2.3. The sequence of the data are reorganized, such that we can investigate the abilities of methylation data at subsequent locations along chromosomes. When organized in such a way, the data will be denoted as time series. The locations s_i are measured at a base pair resolution, where the RnBeads(Assenov et al. 2014) package(in R) is used to map CpG sites to their associated location on a given chromosome. The chromosomes that are studied are chromosome $1-22$, where some are studied more than others.

Throughout the paper, the iterator i is used for different CpG sites, where s_i is the base pair location of CpG site i. The iterator j is used for different persons. The main focus is on a data set containing methylation data from $N = 95$ different persons, where $N_1 = 62$ of these have been diagnosed with Schizophrenia and $N_2 = 33$ have been classified as Healthy. n is used to denote the number of CpG sites

considered. We are going to use both location and position to denote the different CpG sites, where location points to the base pair number s_i and position to the index i.

To make the notation clear, an example is given. If we consider chromosome 6, we have $n = 36614$ CpG sites where the methylation value is measured for each person j. The location of each of these sites along chromosome 6 are then loaded from the RnBeads package, giving the location vector $\boldsymbol{s} = (s_1, ..., s_{36614})$. This makes us able to know the distance between each of the CpG sites(in base pairs), which we later use to define models for taking spatial dependency into account. In this paper, we model and evaluate data from the different chromosomes independently.

2.3 The MARMAL-AID database

Marmal-aid is a combined database and R package that allows for investigation of the methylation state of regions of interest across the genome. The database is a collection of the majority of the publicly available Illumina HumanMethylation450 data, which has been reorganized such that it follows the same setup. This makes it easy to download and use data from different experiments, which previously could be difficult due to inconsistent annotation.

Marmal-aid provides processed data that has undergone a normalization procedure as an attempt to decrease the variability across experiments. Missing values have also been imputed based on a nearest neighbor algorithm. In addition to the processed data, the developers are currently working on providing raw, unrefined data. More information concerning Marmal-aid and the normalization procedure they use, are found in (Lowe and Rakyan 2013).

Because of the variability across experiments, small methylation differences should not be weighted heavily when evaluating data from different experiments. In this paper, we therefore focus on data from the same experiment.

2.4 CpG islands

Along the chromosomes, there are regions with higher density of CpG sites than others. These regions are called CpG islands, but the definition of such an island has some variations. The usual formal definition(Takai and Jones 2002) is a region consisting of at least 200 base pairs, a GC percentage greater than 50% and with an observed-to-expected CpG ratio that is greater than 60%. To clarify this definition, the following example is given.

Let us consider a region of 300 bp, which contains 80 G bases, 80 C bases, 70 A bases and 70 T bases. The GC percentage of this region is then calculated as

follows:

$$\frac{G+C}{G+C+A+T} \cdot 100 = 53.33\%. \tag{2.1}$$

Since the GC percentage is higher than 50%, this region is a possible CpG island. For the next step, we need to calculate the observed-to-expected CpG ratio, which is defined as the number of observed CpG sites divide by the expected number of CpG sites. The expected number is usually calculated as the number of Cs times the number of Gs, divided by the length of the sequence. If we were to observe 17 CpG sites in this interval, our observed-to-expected CpG ratio would be

$$\frac{17}{\frac{80 \cdot 80}{300}} = 0.797, \tag{2.2}$$

such that our region would be classified as an CpG island. The RnBeads package (Assenov et al. 2014), which uses the definition given, has been used to extract which CpG sites that is part of a CpG island.

Many genes in the human genome have CpG islands associated with the start of the gene, called promoters. In most instances, the CpG sites in the CpG islands of promoters are unmethylated if the genes are expressed. This observation has led to beliefs that methylation of these regions might lead to silencing(Mirella Gonzalez-Zulueta and Nguyen 1995) of the following genes. To try to find only those CpG islands that are associated with promoter regions, (Takai and Jones 2002) developed a slightly different definition; sequence length at least 500 base pairs, GC percentage at least 55% and observed-to-expect ratio of 0.65. In this paper, the CpG island association is not of specific interest, such that which definition we use is not of great importance.

2.5 Common assumptions, and alternative parametrization of the Beta-distribution

Common assumptions for the methylation data, when evaluating differences between groups of people, are that the observations y_{ij} for each person are independently beta distributed between the locations. Here y stands for the methylation value, i is the iterator deciding which CpG site that is evaluated and j stands for person j. People belonging to the same group are expected to have the same expectation value at a CpG site, such that the distribution can be seen as

$$y_{ij} \sim \text{Beta}(\mu_i, \phi_i). \tag{2.3}$$

$\mu_i = E_j[Y_i]$ stands for the expectation at CpG site i for the persons $j = 1, ..., N$ of the same group and ϕ_i is the precision parameter in the distribution which determines the spread around the mean. ϕ is denoted with an i, since it can change for the different locations, or be held constant such that $\phi_i = \phi$ for all i. What assumptions

to be made for the methylation data studied in this paper are evaluated in the exploratory analysis in Chapter 3. Here we have used an alternative parametrization of the beta distribution, which is defined as follows.

For a random variable Y, the Beta(a, b)-distribution has the following density

$$\pi(y) = \frac{1}{\mathrm{B}(a,b)} y^{a-1}(1-y)^{b-1}, \quad 0 < y < 1, \quad a > 0 \quad b > 0, \tag{2.4}$$

where $\mathrm{B}(a, b)$ is the Beta-function

$$\mathrm{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{2.5}$$

and $\Gamma(y)$ is the Gamma-function. An alternative parametrization for the Beta-distribution is given by Beta(μ, ϕ), where μ is the expected value and ϕ is the precision parameter. In terms of a and b, they are defined as

$$\begin{aligned} \mu &= \frac{a}{a+b} \quad 0 < \mu < 1 \\ \phi &= a + b \qquad \phi > 0. \end{aligned} \tag{2.6}$$

The expected value and the variance in terms of μ and ϕ are then

$$\mathrm{E}[Y] = \mu \quad \text{and} \quad \mathrm{Var}[Y] = \frac{\mu(1-\mu)}{1+\phi}. \tag{2.7}$$

ϕ is known as the precision parameter since for fixed μ, larger ϕ results in smaller variance of Y. The INLA package uses this parametrization for the beta distribution.

To clarify the term group or class, which we use extensively, we mean a gathering of people with the same qualities. If we want to look into the effect methylation has on the mental illness Schizophrenia, this would be the quality which divides our group in two. Since age has shown to be a significant factor for differences in the methylation mean for many CpG sites(Horvath 2013), this should be taken into account when evaluating group differences. However, since the patients in the data set evaluated, both the Healthy and the Schizophrenia, have a high variation of age, this should not have a great impact on the differences. The mean age of the persons with Schizophrenia is 32, while it is 29 for the Healthy ones.

2.6 Generalized linear models; Beta regression

Generalized linear models(Hilbe 1994) are generalizations of ordinary linear regression that allows the response variable to have a distribution other than the normal distribution. The mean of the data is through a link function defined on a normal scale, such that regular linear regression can be applied. This is what the interest is

in this paper, to find relationships for the mean of the data. The procedure will be described in more detail for the case with methylation data, i.e. beta regression.

With a beta distribution defined as in Section 2.5, it implies that

$$y_{ij} \sim \text{Beta}(\mu_{ij}, \phi_i). \tag{2.8}$$

For a response that is beta distributed, the variables are restricted to values between 0 and 1. Because of these restrictions, the response variables can be seen as probabilities. To define linear relationships on the mean μ_{ij} based on some covariates directly, might lead to regression coefficients which for some covariate values give expectation values that does not fall in the range of $(0, 1)$. Therefore, we would either need restrictions on the regression coefficients, or we could transform the response variable to a normal scale$(-\infty, \infty)$ such that linear relationships can be found between the transformed mean and the covariates. The last approach is known as generalized linear models, which allows the mean of a response variable with an arbitrary distribution to vary linearly with some covariates through a link function.

For a beta distribution, the logit link function is commonly used. This means that if we are interested in doing inference on the mean μ_{ij}, we model the linear predictor η_{ij}, defined as

$$\eta_{ij} = \text{logit}(\mu_{ij}) = \log(\frac{\mu_{ij}}{1-\mu_{ij}}). \tag{2.9}$$

As an example, we consider a specific model evaluated in this paper for finding differences in mean between people that are classified as having Schizophrenia against Healthy ones,

$$\eta_{ij} = -1 + \alpha_i + \beta_i k_j. \tag{2.10}$$

Here k_j is a factor 0 or 1 deciding if person j is Healthy or has Schizophrenia, α_i is the mean of the persons that are classified as healthy at CpG site i, and β_i is the added effect in mean at this site of having Schizophrenia. -1 removes the intercept, which would be an effect calculated based on all the observations, which is not of interest for this evaluation. Regular linear regression tools can be applied to find the regression coefficients $\boldsymbol{\alpha}, \boldsymbol{\beta}$, where the resulting relationship between the mean μ_{ij} and the coefficients is

$$\mu_{ij} = \text{logit}^{-1}(\eta_{ij}) = \frac{\exp(\alpha_i + \beta_i k_j)}{\exp(\alpha_i + \beta_i k_j) + 1}. \tag{2.11}$$

The interpretation of the regression coefficients are that a unit increase affects the log odds of the mean in the beta distribution linearly. This is the approach we use in this paper, where the regression is done through the Bayesian framework(George E. P. Box 1992) with the INLA(Rue, Martino, and Chopin 2009) algorithm.

3 Exploratory analysis and T-test

In this Chapter, we explore the distribution of the methylation data. This includes exploration of the variation in the mean and the precision parameter of the distribution. Further, we explore the correlations along the chromosomes, and the strength of these correlations. Lastly, we perform a T-test to evaluate differences in mean between people having Schizophrenia and being Healthy.

3.1 Exploratory analysis of the Schizophrenia data set

Figure 3.1 displays the methylation data as time series dependent on the locations s_i along chromosome 1 at 250 subsequent locations. The mean for the different locations seem to be highly non-stationary, such that each location seem to have its own mean.

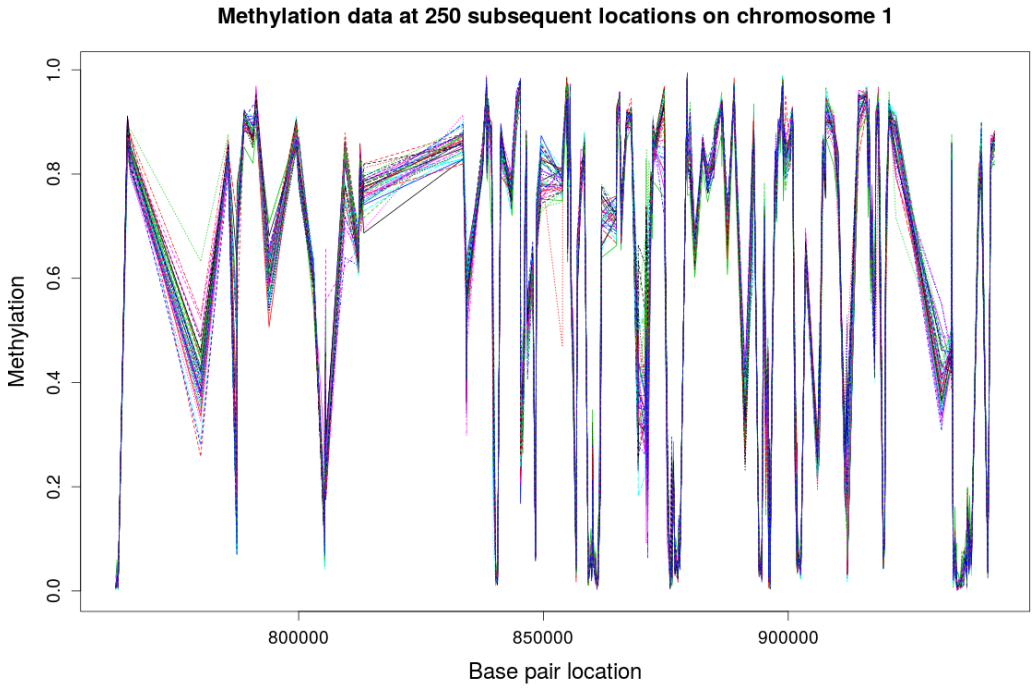


Figure 3.1: Matrix plot of methylation data at 250 subsequent locations at chromosome 1, for 40 randomly chosen people from the Schizophrenia data set.

In Figure 3.2, the empirical mean of each location is removed from the data set, and both the data and the mean are plotted against the CpG position i(identical distance

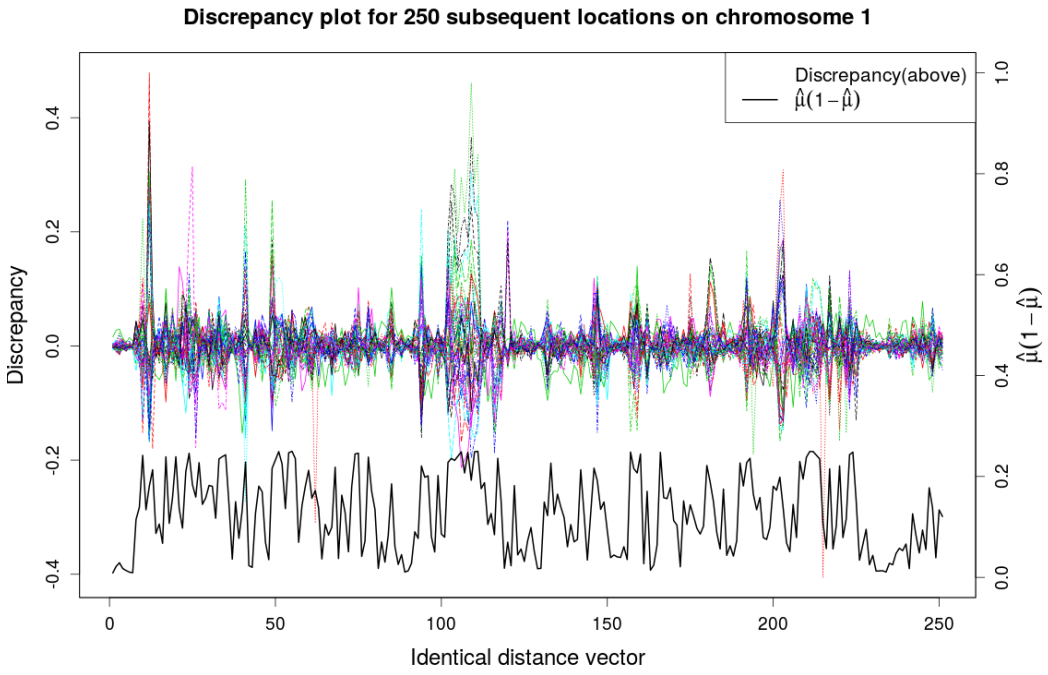


Figure 3.2: Discrepancy plot of methylation data at 250 subsequent locations at chromosome 1, for 40 randomly chosen people from the Schizophrenia data set. A function of the sample mean is plotted underneath, with axis(right hand side) chosen such that both plots are readable.

vector). The discrepancy data are therefore

$$\boldsymbol{y}_{D,i} = \boldsymbol{y}_i - \frac{1}{N}\sum_{j=1}^{N} y_{j,i} = \boldsymbol{y}_i - \hat{\mu}_i, \tag{3.1}$$

where $\boldsymbol{y}_i$ is a vector of methylation values of length N with one value for each person j and $\hat{\mu}_i$ is the empirical mean at location i. This is done, such that the discrepancy(how much the data varies) of the data can be easily evaluated at each location. The variation seems to be changing along the chromosome, and it seems to be largest in regions where the mean is around 0.5. This can be seen by the function of the sample mean, $\hat{\mu}(1-\hat{\mu})$, being above 0.2 for most of the regions where the data varies a lot. The beta distribution has these properties, which can be seen from Equation (2.7) in Section 2.5. The Equation (2.7) indicates that the variance alters with the mean of the location, where it is largest at $\mu = 0.5$. Since the amount of variation at locations with similar mean seem to vary, the assumption of a constant ϕ does not seem to be likely. Note that the analysis is done over a larger region then the one displayed in Figure 3.2.

To further investigate the claim that the data are beta distributed, we investigate the relationship between the sample mean and the sample variance of the data against these estimates of simulated beta distributed variables. This is done by calculating the sample mean and variance from the data and transforming these estimates to a_i and b_i estimates by formula

$$\begin{aligned} \hat{a}_i &= -\frac{\hat{\mu}_i(\hat{\mu}_i^2 - \hat{\mu}_i + \hat{\sigma}_i^2)}{\hat{\sigma}_i^2} \\ \hat{b}_i &= \frac{-(\hat{\mu}_i - 1)\hat{a}_i}{\hat{\mu}_i}, \end{aligned} \tag{3.2}$$

where

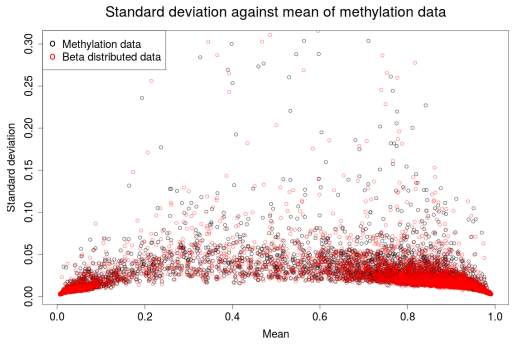
$$\begin{aligned} \hat{\mu}_i &= \frac{1}{N}\sum_{j=1}^{N} y_{ij} \\ \hat{\sigma}_i^2 &= \frac{1}{N-1}\sum_{j=1}^{N}(y_{ij} - \hat{\mu}_i)^2. \end{aligned} \tag{3.3}$$

Then we simulate N beta distributed variables from the empirical distribution $\text{Beta}(\hat{a}_i, \hat{b}_i)$, and calculate the sample mean and variance of the simulated variables. This is done to capture the natural variation of N beta distributed variables at each position i.

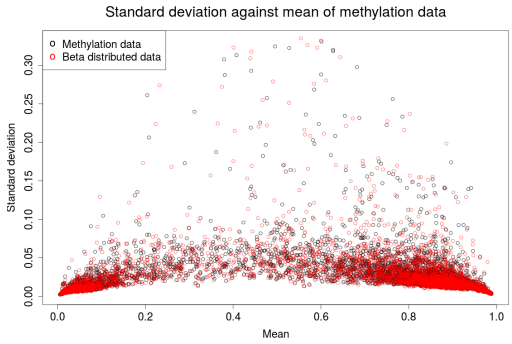
In Figure 3.3, the standard deviation is plotted against the sample mean of the methylation data and the beta distributed variables. This is done for 5000 subsequent locations from four different chromosomes. They seem to follow the same pattern, which reinforces the assumption that the methylation data are beta distributed.

The precision parameter for the beta observations does not seem to be constant along the chromosomes. In Figure 3.4, we have displayed the sample mean and the estimated standard deviation of methylation data from chromosome 1 against the sample mean and estimated standard deviation of beta distributed values with constant ϕ equal to $20, 50, 100$ and 200. The spread in the methylation data is not well accounted for by the different beta distributed values, which suggests that the precision parameter varies for the different locations in such a way that it can not be described by a constant ϕ.

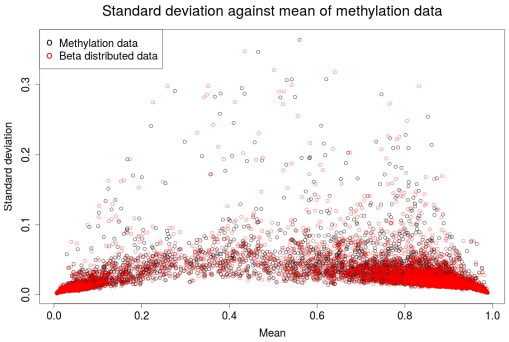
As has been mentioned earlier, we check the assumption of correlation in the data along the chromosomes. This is done through the sample auto-correlation function taken over the different chromosomes for each person. We need to evaluate the discrepancy data set, such that the time series can be seen as having stationary mean



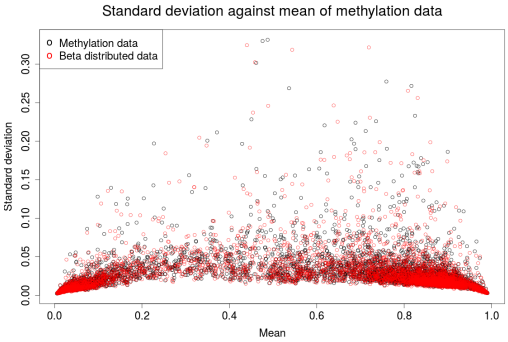
(a) 5000 subsequent locations at chromosome 1.



(b) 5000 subsequent locations at chromosome 2.



(c) 5000 subsequent locations at chromosome 5.



(d) 5000 subsequent locations at chromosome 11.

Figure 3.3: Empirical standard deviation against sample mean of methylation data from four different chromosomes and beta distributed samples.

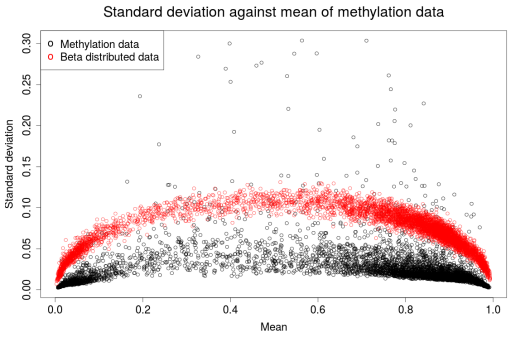
for each location along the chromosomes. The sample auto-correlation is defined as

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} = \frac{\sum_{i=1}^{n-k}(y_i - \tilde{y})(y_{i+k} - \tilde{y})}{\sum_{i=1}^{n}(y_i - \tilde{y})^2}, \tag{3.4}$$

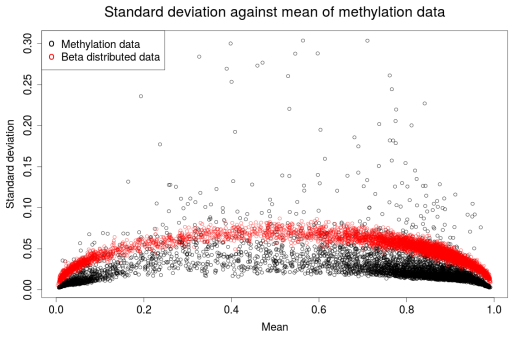
where $\hat{\rho}_k$ is the average lag k correlation, $\hat{\gamma}_k$ is the average lag k covariance which for $k = 0$ is the variance, $\tilde{y}$ is the average methylation along the chromosome and $i = 1,...,n$ are the subsequent positions under study. For time series without correlation, the auto-correlation is equal to 0 for $k \geq 1$. For large n, $\hat{\rho}_k$ is approximately normally distributed(Wei 2006) with mean ρ_k and variance approximated by Bartlett(Bartlett 1946)

$$\text{Var}(\hat{\rho}_k) \approxeq \frac{1}{n}(1 + 2\hat{\rho}_1^2 + 2\hat{\rho}_2^2 + \cdots + 2\hat{\rho}_m^2), \tag{3.5}$$

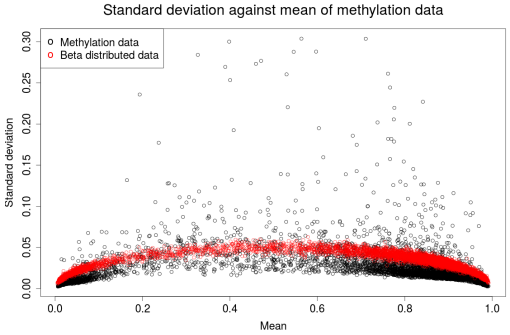
where m is the number of consecutive lag k correlations that are not equal to zero, $\hat{\rho}_k \neq 0$ for $k = 1,...,m$. This means that to test if we have a white noise process,



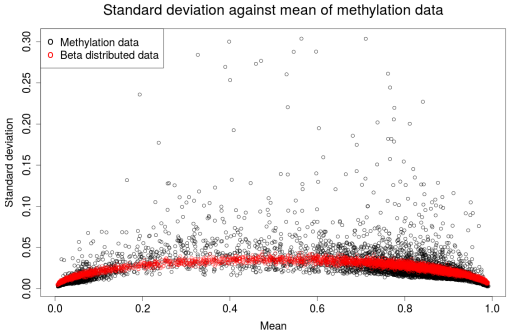
(a) 5000 subsequent locations at chromosome 1. The beta distributed values are simulated with constant $\phi = 20$.



(b) 5000 subsequent locations at chromosome 1. The beta distributed values are simulated with constant $\phi = 50$.



(c) 5000 subsequent locations at chromosome 1. The beta distributed values are simulated with constant $\phi = 100$.



(d) 5000 subsequent locations at chromosome 1. The beta distributed values are simulated with constant $\phi = 200$.

Figure 3.4: Empirical standard deviation against sample mean of methylation data from chromosome 1, with four different beta distributed samples with constant ϕ.

which is a process without correlations, we test each correlation of lag k against the hypothesis that $\rho_k = 0$ with standard error equal to $S_{\rho_k} = \sqrt{\frac{1}{n}}$.

In Figure 3.5, we have displayed the sample auto-correlation function of the three first lags for the 95 different people in the Schizophrenia data set, estimated by using the data from chromosome 1, where $n = 46866$. The plot indicates that there are correlations along the chromosome, and the hypothesis of a white noise process is rejected. The correlation also seems to vary in strength for the different persons. In the appendix A.1, these differences are shown through box-plots and compared to simulation of Gaussian random fields(GRFs)(Section 4.3). The conclusion is that the amount of correlation along a chromosome for different people seem to vary more than it would if the discrepancy data were realizations of a GRF with the same parameters.

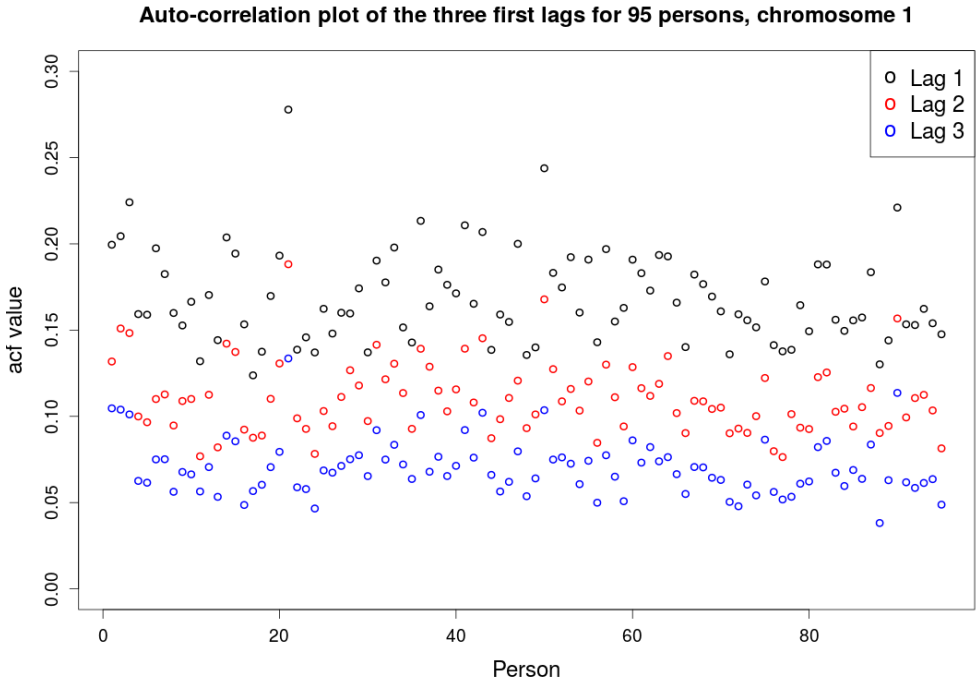


Figure 3.5: Auto-correlation of the three first lags for the 95 people in the Schizophrenia data set. The auto-correlation is calculated with data from chromosome 1.

With the explanatory analysis given, we have seen that the assumption of beta distributed data with non-stationary mean μ_i and precision parameter ϕ_i, seems to fit the data. From the acf plots, we have seen that there are some significant correlation along the chromosome, which might influence the evaluation of differently methylated regions and sites among persons classified as Healthy against Schizophrenia. Since the acf varies much between the different persons, the strength of the dependency structure might be different for each person.

3.2 Differently methylated regions; T-test

To get an overview of differences in mean μ between two groups of people, we specify a T-test. The test is thoroughly explained for the case study of the Schizophrenia data set, where each CpG cite i is tested independently. The two groups of people we are testing the differences in mean between, are:

$$\begin{aligned} y_{1,s_i}: &\quad \text{People classified with Schizophrenia} \\ y_{2,s_i}: &\quad \text{People classified as Healthy} \end{aligned}, \tag{3.6}$$

where s_i is the location evaluated and $\boldsymbol{y}_{1,s_i}$ and $\boldsymbol{y}_{2,s_i}$ are vectors of methylation values for location s_i, with length $N_1 = 62$ and $N_2 = 33$ being the number of people in group 1 and 2. The hypothesis to be tested is then stated as follows:

$$\begin{aligned} \mathrm{H}_0: &\quad \mu_1 = \mu_2 \quad \text{at location } s_i \in \text{HumanMethylation450k} \\ \mathrm{H}_1: &\quad \mu_1 \neq \mu_2 \quad \text{at location } s_i \in \text{HumanMethylation450k}. \end{aligned} \tag{3.7}$$

The null hypothesis is now tested with a two sided T-test, where the test variable is defined as

$$T_{s_i} = \frac{(\bar{y}_{2,s_i} - \bar{y}_{1,s_i}) - (\mu_2 - \mu_1)}{\mathrm{SE}}, \tag{3.8}$$

which under H_0 is equal to $T_{s_i} = \frac{\bar{y}_{2,s_i} - \bar{y}_{1,s_i}}{\mathrm{SE}}$ and has a student t-distribution with $(N_2 - 1) + (N_1 - 1) = 93$ degrees of freedom. The test variable is compared to the critical value for the t distribution with 93 degrees of freedom and significance level $\alpha = 0.05$, which is equal to 1.986(two sided). If $|T_{s_i}|$ exceeds 1.986, where $|\cdot|$ denotes the absolute value, we reject the null hypothesis for location s_i.

The standard error SE is calculated by formula

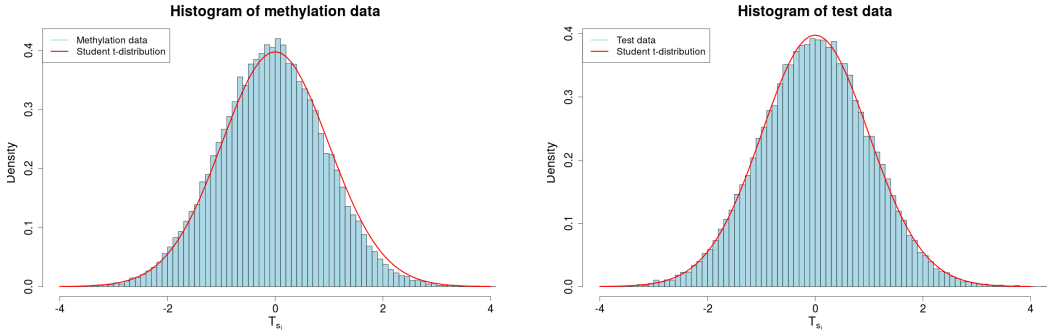
$$\mathrm{SE} = S_\mathrm{p} \sqrt{\frac{N_2 + N_1}{N_2 N_1}}, \tag{3.9}$$

where S_p is the pooled standard deviation, given by

$$S_\mathrm{p} = \sqrt{\frac{S_2^2(N_2 - 1) + S_1^2(N_1 - 1)}{(N_2 - 1) + (N_1 - 1)}}. \tag{3.10}$$

The pooled standard deviation works as a weighted average of the standard deviations to the two samples, since if one sample is much larger than the other, has more degrees of freedom, it should count for more of the variation. It is build on the assumption that the true standard deviation for each group is the same. By comparing the estimated standard deviation of each group with the standard deviation of simulated data samples of size N_1 and N_2 from two identical beta distributions, this assumption seems to be reasonable. To test the hypothesis for location s_i, we therefore need to estimate four quantities of the samples:

$$\begin{aligned} \bar{y}_{2,s_i} &= \frac{1}{N_2} \sum_{j=1}^{N_2} y_{j,2,s_i}, \quad \bar{y}_{1,s_i} = \frac{1}{N_1} \sum_{k=1}^{N_1} y_{k,1,s_i} \\ S_2^2 &= \frac{1}{(N_2 - 1)} \sum_{j=1}^{N_2} (y_{j,2,s_i} - \bar{y}_{2,s_i})^2, \quad S_1^2 = \frac{1}{(N_1 - 1)} \sum_{k=1}^{N_1} (y_{k,1,s_i} - \bar{y}_{1,s_i})^2. \end{aligned} \tag{3.11}$$



(a) Histogram of the T_i parameter with corresponding H_0 distribution.

(b) Histogram of T_i(test set) with corresponding H_0 distribution.

Figure 3.6

In Figure 3.6, we have displayed a histogram of all test variables created from applying the test to all 46866 locations at chromosome 1. The same is displayed for a test data set of same dimensions, containing beta distributed variables based on estimates of the mean and precision parameter of both samples(Schizophrenia and Healthy) seen as one. These variables are then divided into groups of size N_1 and N_2, and the same test as above is applied. This is done to see how the test performs for beta distributed variables from the same distribution.

Figure 3.6 shows that both histograms fit the H_0 distribution quite well, although the histogram of the methylation data seems to be a bit skewed. This might point towards some trend, that the methylation mean of the Schizophrenia group is a bit lower than the healthy group for most sites on chromosome 1. It might also be caused by some small deviations between the data and the assumed beta distribution.

In Table 3.1, the amount of significant test variables T_i are given from doing the test on all the chromosomes, individually. As we can see, each of the chromosomes have a significantly different mean at less locations than what we would expect from the type I error. This is further discussed in Chapter 8, and might point towards some underlying structure in the methylation data that is not yet quantified.

If we use the well used differently methylated positions finder algorithm $dmpFinder()$ (Hansen and Aryee 2015), we get the exact same CpG sites defined as differently methylated, meaning that their p-value is below 0.05. This is shown in Figure 3.7, where the test is done on a small part of chromosome 6, which is studied in the next chapters.

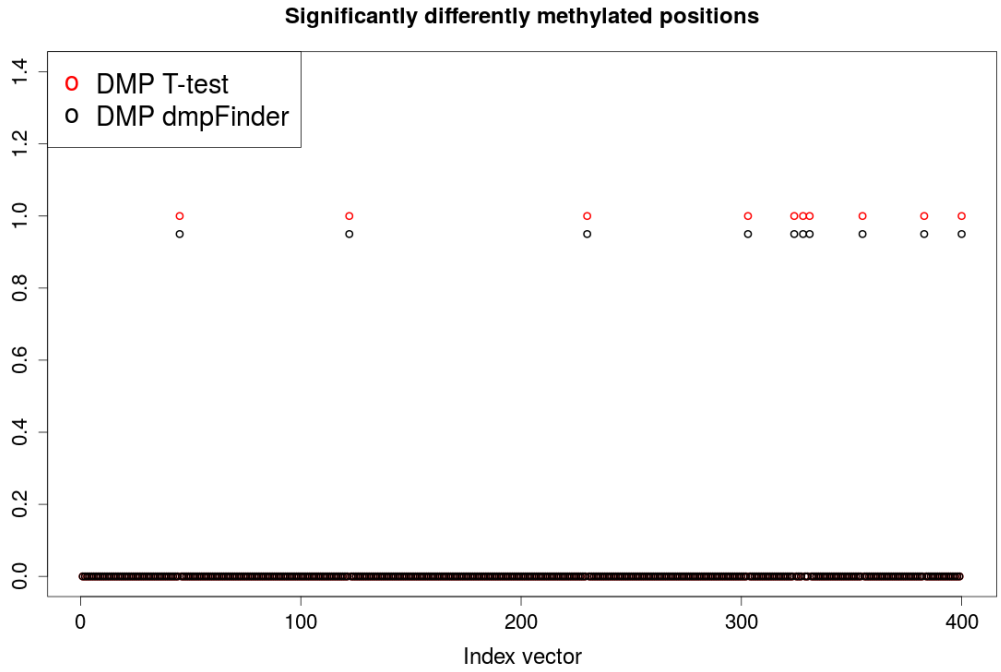


Figure 3.7: The points show which positions on 400 subsequent locations at chromosome six that are classified as differently methylated based on the T-test and the dmpFinder. DMP stands for differently methylated positions.

Table 3.1: Table over amount of significant test variables T_i at chromosome 1 – 22, compared with the amount from the test data.

Chromosome	Amount outside, meth. data	Amount outside, test data
1	0.043	0.049
2	0.042	0.051
3	0.042	0.051
4	0.039	0.050
5	0.042	0.050
6	0.041	0.050
7	0.042	0.049
8	0.041	0.050
9	0.048	0.046
10	0.041	0.051
11	0.041	0.048
12	0.040	0.051
13	0.045	0.049
14	0.044	0.052
15	0.043	0.046
16	0.043	0.049
17	0.043	0.049
18	0.041	0.046
19	0.045	0.049
20	0.043	0.048
21	0.043	0.051
22	0.048	0.051

4 Background; Latent Gaussian models, Bayesian inference and INLA

Because of the estimated correlations along the chromosomes, we want to investigate the possibility of spatial dependency in the data. In this and the following Chapters, we therefore introduce models with spatial dependency and some new background material. We also give a description of the methodology and algorithm used to do the inference.

4.1 Bayesian modeling and inference

Bayesian inference is a method of statistical inference in which Bayes theorem is used to update prior beliefs on some latent variable θ based on observable data $\boldsymbol{y}$. Consider having an assumed probability distribution for the observable variables dependent on the latent variable θ, defining the likelihood function

$$\pi(\boldsymbol{y}|\theta). \tag{4.1}$$

The parameter θ is an unknown quantity modeled through a suitable prior probability distribution $\pi(\theta)$, before observing any realizations $\boldsymbol{y}$ which can alter these beliefs. As mentioned earlier in the paper, we are interested in doing inference on the mean of the data, such that θ is in our case the mean, μ. The prior distribution can be informative or non-informative, reflecting the amount of information we have on the parameter prior to the observations. In the presence of a hierarchical structure or spatial (or temporal) dependence between the parameters, it would be more common to express the knowledge of θ through hyperparameters $\boldsymbol{\psi}$[1], such that the distribution becomes $\pi(\theta|\boldsymbol{\psi})$. This becomes clearer in the next section, describing the process of Bayesian inference with INLA. To find the posterior distribution of the parameter θ, Bayes theorem is used:

$$\pi(\theta|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\theta) \times \pi(\theta)}{\pi(\boldsymbol{y})}. \tag{4.2}$$

The posterior distribution $\pi(\theta|\boldsymbol{y})$ represents the uncertainty about the parameter of interest, θ, after having observed the data $\boldsymbol{y}$, which has some assumed relationship to θ through the likelihood function. Note that the marginal distribution of $\boldsymbol{y}$ in the denominator is considered as a normalizing constant since it is not dependent on θ, such that Bayes theorem is often reported as

$$\pi(\theta|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\theta) \times \pi(\theta). \tag{4.3}$$

[1] A hyperparameter is a parameter of a prior distribution; they are not parameters of the model for the underlying system under analysis.

The aim of Bayesian inference is therefore to obtain the marginal posterior distribution of each parameter in θ(only one in the example above), and the marginal posterior distribution of the hyperparameters $\boldsymbol{\psi}$ if any.

It is worth noticing that the interpretation of the parameters of interest in the Bayesian framework and the frequentist(classical) framework is different. In the Bayesian framework, the θ_is are characterized by a probability distribution. In the frequentist approach, θ_i is considered as a fixed unknown quantity and only its estimator $\hat{\theta}_i$, which is a function of the data, is a random variable(Blangiardo and Cameletti 2015). This gives rise to two different parameter estimation intervals; credibility regions and confidence intervals. The credibility region is used for Bayesian inference, and means that given the data, there is a 0.95 probability(for significance level $\alpha = 0.05$) that the true value θ_i falls within the region. The confidence interval, used for the frequentist approach, means that there is a 0.95 probability that the interval estimated contains the true value θ_i.

4.2 Latent Gaussian models and Bayesian inference with INLA

The following introduction is based on (Blangiardo and Cameletti 2015) and the same notation is used. Latent Gaussian models(Rue, Martino, and Chopin 2009) can be represented by a hierarchical structure containing three stages. The first stage is to define the conditionally independent likelihood function based on the assumed distribution for the observed data $\boldsymbol{y}$. A very general approach is to specify a distribution for y_i in terms of the mean μ_i, defined as a function of an additive linear predictor η_i through a link function $g(\cdot)$, such that $g(\mu_i) = \eta_i$. This is a result of generalized linear models, which is described in more detail for the case of beta regression in Section 2.6. The additive linear predictor is defined as

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m k_{mi} + \sum_{l=1}^{L} f_l(z_{li}), \tag{4.4}$$

where β_0 is s scalar representing the intercept, $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)$ quantify the linear effect of some covariates $\boldsymbol{k}$ on the response and $\boldsymbol{f} = (f_1(\cdot), ..., f_L(\cdot))$ is a collection of functions defined over a set of covariates $\boldsymbol{z} = (z_1, ..., z_L)$. These functions can assume different forms, such as nonlinear effects of covariates, time trends and seasonal effects or temporal or spatial random effects. By denoting all the latent (nonobservable) components of interest for the inference as $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$, and all hyperparameters as $\boldsymbol{\psi}$, the likelihood function can be specified as

$$\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{N} \pi(y_i|\theta_i, \boldsymbol{\psi}). \tag{4.5}$$

The next step is to assume a multivariate normal prior distribution on the inference parameters $\boldsymbol{\theta}$ with mean $\mathbf{0}$ and precision matrix dependent on the hyperparameters $\boldsymbol{\psi}$, $\boldsymbol{Q}(\boldsymbol{\psi})$. This gives the density function

$$\pi(\boldsymbol{\theta}|\boldsymbol{\psi}) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{Q}(\boldsymbol{\psi})|^{\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{Q}(\boldsymbol{\psi})\boldsymbol{\theta}), \tag{4.6}$$

where $|\cdot|$ denotes the matrix determinant and $'$ is the transpose operation. The components of the Gaussian field $\boldsymbol{\theta}$ are assumed conditionally independent, which makes the precision matrix $\boldsymbol{Q}(\boldsymbol{\psi})$ very sparse. This specification is known as Gaussian Markov random fields (Rue and Held 2005a), and gives rise to computational benefits by being able to use numerical methods especially developed for sparse matrices.

The last stage is to define prior distributions on the hyperparameters $\boldsymbol{\psi}$, $\pi(\boldsymbol{\psi})$. With this being specified, we can find the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ with Bayes formula, giving

$$\begin{aligned}
\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\psi}) \times \pi(\boldsymbol{\theta}|\boldsymbol{\psi}) \times \pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi}) \\
&\propto \pi(\boldsymbol{\psi}) \times \pi(\boldsymbol{\theta}|\boldsymbol{\psi}) \times \prod_{i=1}^{N} \pi(y_i|\theta_i, \boldsymbol{\psi}) \\
&\propto \pi(\boldsymbol{\psi}) \times |\boldsymbol{Q}(\boldsymbol{\psi})|^{\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{Q}(\boldsymbol{\psi})\boldsymbol{\theta} + \sum_{i=1}^{n} \log(\pi(y_i|\theta_i, \boldsymbol{\psi}))).
\end{aligned} \tag{4.7}$$

The objective with Bayesian inference is to obtain the marginal posterior distributions for each element in $\boldsymbol{\theta}$ and in $\boldsymbol{\psi}$,

$$\pi(\theta_i|\boldsymbol{y}) = \int \pi(\theta_i, \boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi} = \int \pi(\theta_i|\boldsymbol{\psi}, \boldsymbol{y}) \pi(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi} \tag{4.8}$$

and

$$\pi(\psi_k|\boldsymbol{y}) = \int \pi(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi}_{-k}. \tag{4.9}$$

We therefore need to compute $\pi(\boldsymbol{\psi}|\boldsymbol{y})$, from which all the relevant marginals $\pi(\psi_k|\boldsymbol{y})$ can be obtained, and $\pi(\theta_i|\boldsymbol{\psi}, \boldsymbol{y})$, which is needed to obtain the parameter marginals posteriors $\pi(\theta_i|\boldsymbol{y})$.

The INLA approach uses the assumptions of the model to produce numerical approximations to the posteriors of interest based on a Laplace approximation method. This method is explained thoroughly with examples in (Blangiardo and Cameletti 2015) and (Rue, Martino, and Chopin 2009), and will only be given briefly in the following paragraphs.

Consider the integral

$$\int f(x)dx = \int \exp(\log f(x))dx, \tag{4.10}$$

where $f(x)$ is the density function of a random variable X. By representing $\log(f(x))$ by means of a Taylor expansion evaluated at $x = x^*$, where x^* is set equal to the mode $x^* = \text{argmax}_x \log(f(x))$, we can approximate the integral with the following Taylor expansion substitution:

$$\log(f(x)) \approx \log(f(x^*)) + \frac{(x-x^*)^2}{2} \frac{\delta^2 \log(f(x))}{\delta x^2}\bigg|_{x=x^*}. \tag{4.11}$$

Here we have used the fact that $\frac{\delta \log(f(x))}{\delta x}\big|_{x=x^*} = 0$, since the density $f(x)$ is evaluated at the mode x^*. The integral of interest is then equal to

$$\int f(x)dx \approx \exp(\log(f(x^*))) \int \exp\left(\frac{(x-x^*)^2}{2} \frac{\delta^2 \log(f(x))}{\delta x^2}\bigg|_{x=x^*}\right)dx, \tag{4.12}$$

where the integrand can be associated with the density of a Normal distribution. This can be seen by replacing $-1\Big/\frac{\delta^2 \log(f(x))}{\delta x^2}\big|_{x=x^*}$ with σ^{2^*}, resulting in

$$\int f(x)dx \approx \exp(\log(f(x^*))) \int \exp\left(-\frac{(x-x^*)^2}{2\sigma^{2^*}}\right)dx, \tag{4.13}$$

where the integrand can be seen as the kernel of a Normal distribution with mean equal to x^* and variance equal to σ^{2^*}. The integral evaluated at the interval (a, b) can therefore be approximated by

$$\int_a^b f(x)dx \approx f(x^*)\sqrt{2\pi\sigma^{2^*}}(\Phi(b) - \Phi(a)), \tag{4.14}$$

where $\Phi(\cdot)$ denotes the cumulative density function of the Normal(x^*, σ^{2^*}) distribution.

With the approximation method given, we can compute the posterior of the hyperparameters as

$$\begin{aligned}
\pi(\boldsymbol{\psi}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{\theta}, \boldsymbol{\psi}|\boldsymbol{y})}{\pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{y})} \\
&= \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}, \boldsymbol{\psi})}{\pi(\boldsymbol{y})} \frac{1}{\pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{y})} \\
&\propto \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{\pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{y})} \\
&\approx \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\psi})\pi(\boldsymbol{\theta}|\boldsymbol{\psi})\pi(\boldsymbol{\psi})}{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{y})}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\psi})} =: \tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y}),
\end{aligned} \tag{4.15}$$

where $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{y})$ is the Gaussian approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{y})$ and $\boldsymbol{\theta}^*(\boldsymbol{\psi})$ is the mode for a given $\boldsymbol{\psi}$. The Gaussian approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{\psi}, \boldsymbol{y})$ turns out to be accurate, since

the true probability density appears to be almost Gaussian as it is prior distributed as a GMRF, $\boldsymbol{y}$ is generally not informative and the observation distribution is usually well-behaved(Blangiardo and Cameletti 2015, p. 110).

To compute $\pi(\theta_i|\boldsymbol{\psi},\boldsymbol{y})$ is slightly more complex, since there are usually more elements in $\boldsymbol{\theta}$ than in $\boldsymbol{\psi}$ such that the computation is more expensive. Several methods for approximating $\pi(\theta_i|\boldsymbol{\psi},\boldsymbol{y})$ are developed and explained in (Blangiardo and Cameletti 2015, pp. 109–112) and (Rue, Martino, and Chopin 2009). They consist of rewriting the vector of parameters as $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$ and use Laplace approximation to obtain

$$\begin{aligned}\pi(\theta_i|\boldsymbol{\psi},\boldsymbol{y}) &= \frac{\pi((\theta_i,\boldsymbol{\theta}_{-i})|\boldsymbol{\psi},\boldsymbol{y})}{\pi(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})} \\ &\propto \frac{\pi(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y})}{\pi(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})} \\ &\approx \left.\frac{\pi(\boldsymbol{\theta},\boldsymbol{\psi}|\boldsymbol{y})}{\tilde{\pi}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})}\right|_{\boldsymbol{\theta}_{-i}=\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\psi})} =: \tilde{\pi}(\theta_i|\boldsymbol{\psi},\boldsymbol{y}),\end{aligned} \tag{4.16}$$

where $\tilde{\pi}(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})$ is the Laplace Gaussian approximation of $\pi(\boldsymbol{\theta}_{-i}|\theta_i,\boldsymbol{\psi},\boldsymbol{y})$ and $\boldsymbol{\theta}^*_{-i}(\theta_i,\boldsymbol{\psi})$ is its mode.

Once we have $\tilde{\pi}(\theta_i|\boldsymbol{\psi},\boldsymbol{y})$ and $\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y})$, the marginal posterior distributions $\pi(\theta_i|\boldsymbol{y})$ (4.8) are then approximated by

$$\tilde{\pi}(\theta_i|\boldsymbol{y}) \approx \int \tilde{\pi}(\theta_i|\boldsymbol{\psi},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\psi}|\boldsymbol{y}d\boldsymbol{\psi}, \tag{4.17}$$

where the integral can be solved numerically through a finite weighted sum:

$$\tilde{\pi}(\theta_i|\boldsymbol{y}) \approx \sum_j \tilde{\pi}(\theta_i|\boldsymbol{\psi}^{(j)},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\psi}^{(j)}|\boldsymbol{y}\Delta_j \tag{4.18}$$

for some relevant integration points $\{\boldsymbol{\psi}^{(j)}\}$ with corresponding set of weights $\{\Delta_j\}$. As mentioned earlier, a more comprehensive explanation can be found in (Rue, Martino, and Chopin 2009) and (Blangiardo and Cameletti 2015).

4.3 Gaussian random fields

For many phenomenons, data seems to have some sort of dependence in space. This could be the distribution of a disease, the distribution of trees in the forest, weather phenomenons such as precipitation or as in this paper, the DNA methylation level along a chromosome. What these data sets have in common is that they are all collected from known locations in a domain, and can then be modeled as a realization from a stochastic process indexed by this domain.

One of the ways to model such a process is by using Gaussian Random Fields (GRF). These fields are used to model spatial processes that are continuous in space, and have the property that observations close to each other have more in common than observations far from each other. As shown in the explanatory analysis, the data seems to have correlation between the locations. A GRF might therefore be suitable for modeling the methylation data. To describe the Gaussian Random Fields, we base the explanation on one dimensional locations, since this is the relevant case for the methylation data.

If $\{\xi(s) : s \in D \subset R^1\}$ defines a GRF, it follows that

$$(\xi(s_1), ..., \xi(s_n)) \sim \mathrm{N}_n(\boldsymbol{\mu}, \Sigma),$$

where N_n is the n-variate normal distribution, $\boldsymbol{\mu}$ is the expectation vector and Σ is the covariance matrix, where $\Sigma_{ij} = \mathrm{Cov}(\xi(s_i), \xi(s_j))$ is the covariance between the GRF at location s_i and s_j in the one dimensional domain D. If the mean and the variance of the field are not depending on the location, and if the correlation between two points only depends on the distance between them, the GRF is said to be second-order stationary. This can be written as:

$$\begin{aligned} \mathrm{E}[\xi(s)] &= \mathrm{E}[\xi(s+t)] = \mu \\ \mathrm{Var}[\xi(s)] &= \mathrm{Var}[\xi(s+t)] = \sigma^2 \\ \mathrm{Corr}[\xi(s_i), \xi(s_j)] &= \mathrm{Corr}[\xi(s_i+t), \xi(s_j+t)] = \rho(s_j - s_i). \end{aligned} \tag{4.19}$$

In addition, if the correlation is only dependent on the absolute distance between the locations, $\rho(|s_j - s_i|)$, the field is said to be isotropic(Cressie 1993).

The covariance matrix Σ can be constructed from several different covariance functions, such as the exponential, the Gaussian and the Matérn covariance function. The different versions of covariance functions affect the properties of the spatial dependence, such that the right choice is important. We focus on the Matérn covariance function, given as

$$\mathrm{Cov}(\xi(s_i), \xi(s_j)) = \mathrm{Cov}(\xi_i, \xi_j) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}}(\kappa\|s_i - s_j\|)^{\lambda} K_{\lambda}(\kappa\|s_i - s_j\|), \tag{4.20}$$

since this is the most general of the three listed(exponential: $\lambda = 0.5$, Gaussian : $\lambda \to \infty$). Here K_{λ} is the modified Bessel function of the second kind and order $\lambda > 0$, which measures the degree of smoothness of the process and is usually kept fixed due to poor identifiability (Blangiardo and Cameletti 2015, p. 194). $\|s_i - s_j\|$ denotes the Euclidean distance between the two locations and σ^2 is the variance in the white noise process $\omega(s)$, identified in the Stochastic Partial Differential Equation (4.4). $\kappa > 0$ is a scale parameter, which has a more natural interpretation through the range parameter r, which is defined as the Euclidean distance between $\xi(s_i)$

and $\xi(s_j)$ for which the correlation is approximately 0.1. The link between these parameters is

$$r = \sqrt{8\lambda}/\kappa, \tag{4.21}$$

which is empirically derived(Lindgren, Rue, and Lindström 2011) for $\lambda \geq 0.5$.

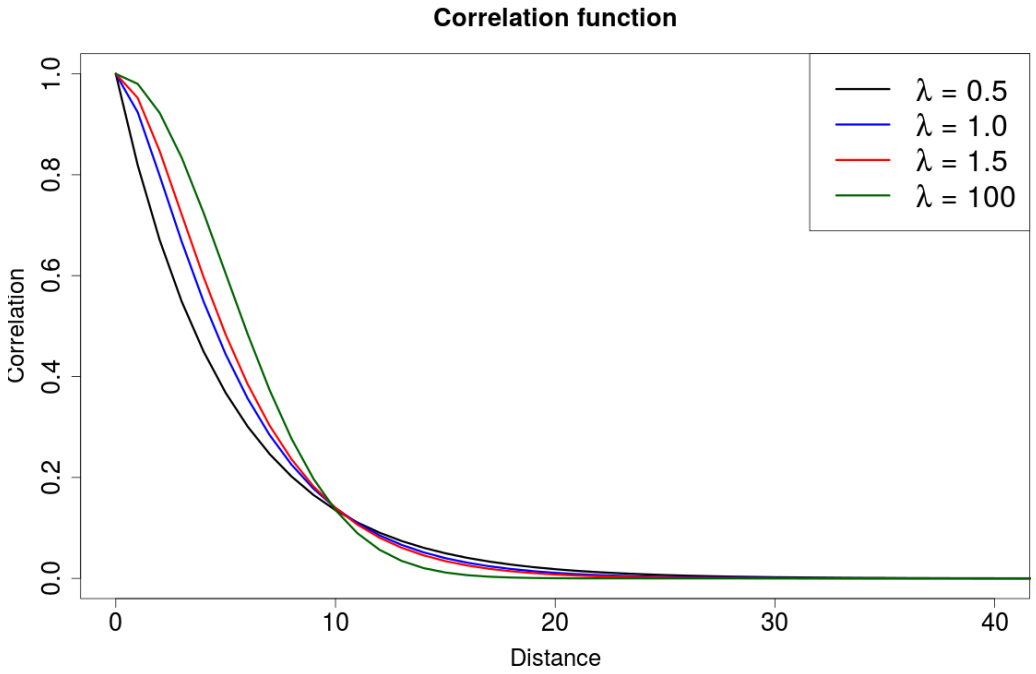
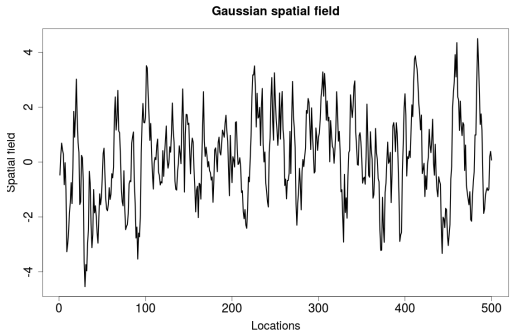


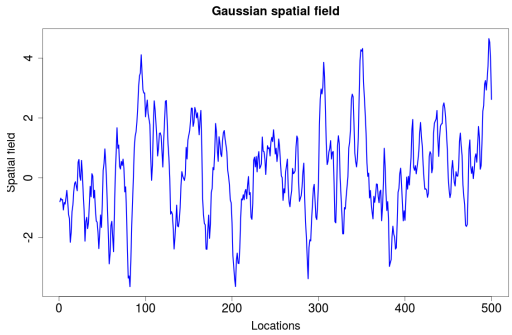
Figure 4.1: Correlation function with $\sigma^2 = 2$, range parameter $r = 10$ and λ equal to the values $0.5, 1.0, 1.5, 100$.

In Figure 4.1, we have displayed the correlation function for range r equal to 10 and different values for λ. The case with $\lambda = 100$ is essentially the Gaussian correlation function. For the different λs, the shape of the correlation function changes. For each of the correlation functions, a realization of the associated Gaussian field is given in Figure 4.2. From the Figures, we see that λ works as an extra smoothing parameter, which for higher values give smoother realizations of the GRF. The other smoothing parameter is the range, which for higher values will result in smoother realizations for fixed λs.

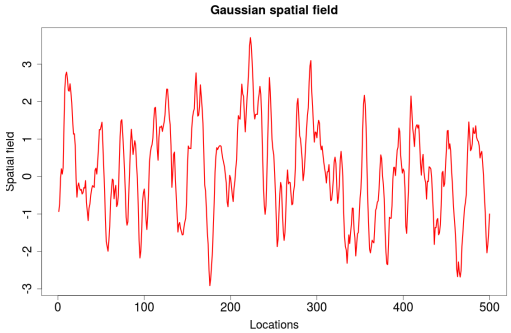
The reason for investigating the impact of λ is that we need to choose a value for this parameter when fitting a SPDE to the methylation data. Since λ can be seen as an extra smoothing parameter, it would seem that it is possible to obtain approximately equal correlation functions by altering the range parameter. This means that the choice of λ might not highly affect the approach. However, for



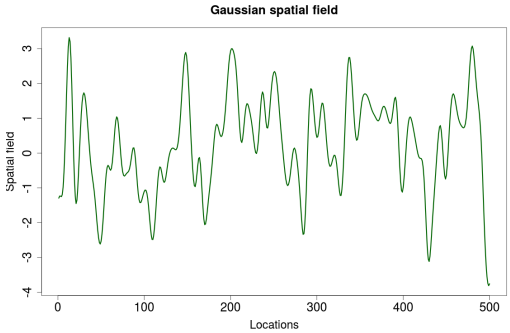
(a) A realization of a Gaussian random field with $\mu = 0$ and correlation function given in 4.1.



(b) A realization of a Gaussian random field with $\mu = 0$ and correlation function given in 4.1.



(c) A realization of a Gaussian random field with $\mu = 0$ and correlation function given in 4.1.



(d) A realization of a Gaussian random field with $\mu = 0$ and correlation function given in 4.1.

Figure 4.2: Realizations of Gaussian random fields with $\mu = 0$ and associated correlation function given by colors in Figure 4.1.

different values of λ, the relationship between the SPDE parameters κ, τ, σ^2 and the range r is affected. This might cause some deviation between the posterior estimates and the true values.

4.4 Stationary dependency structure with INLA - SPDE

With a latent Gaussian model defined as in (4.4) where we assume a spatial dependency in the linear predictor η(as given in (5.1)), we need some sort of tool to be able to model the dependency structure. To do this, Lindgren *et al.* (Lindgren, Rue, and Lindström 2011) proposed the Stochastic Partial Differential Equation(SPDE) approach, which consists in representing a continuous spatial process (i.e. a Gaussian Field(GF)) using a discretely indexed spatial random process (i.e. Gaussian Markov Random Field(GMRF(Rue and Held 2005b))). The starting point is the stochastic

partial differential equation

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}(\tau\xi(s)) = \omega(s), \tag{4.22}$$

where $s \in \mathbb{R}^d$ is the locations, Δ is the Laplacian operator($\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$), α controls the smoothness, $\kappa > 0$ is the scale parameter, τ controls the variance and $\omega(s)$ is a Gaussian white noise process. In our case, s is one dimensional and represents the locations along the DNA where observations of methylation have been done.

The exact and stationary solution to the equation (4.22) is the stationary GRF $\xi(s)$, which describes a smoothed version of the Gaussian white noise process on the right hand side. The solution has a Matérn covariance function given by Equation 4.20. The continuous solution $\xi(s)$ is approximated using a discrete representation in the form of a Markov structure, obtained by the finite element method through a basis representation defined on a triangulation of the domain:

$$\xi(s) = \sum_{g=1}^{G} \phi_g(s)\tilde{\xi}_g. \tag{4.23}$$

G is the total number of nodes of the triangulation, $\{\phi_g(s)\}$ is the set of basis functions and $\{\tilde{\xi}_g\}$ are zero mean Gaussian distributed weights with precision matrix $\boldsymbol{Q}(\kappa, \tau)$. The weights are chosen to approximate the solution in the mesh nodes, and with the basis functions we can transform the approximation of the field from the mesh nodes to a location s_i of interest by Equation (4.23). The precision matrix for the weights $\tilde{\xi}_g$ is defined as

$$\boldsymbol{Q}(\kappa, \tau) = \tau^2(\kappa^4\boldsymbol{D} + 2\kappa^2\boldsymbol{C} + \boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{C}), \tag{4.24}$$

where $\boldsymbol{D}$ is a diagonal matrix with entries $D_{ii} = \int \phi_i(\boldsymbol{s})d\boldsymbol{s}$ and $\boldsymbol{C}$ a sparse matrix with elements $C_{ij} = \int \nabla\phi_i(\boldsymbol{s})\nabla\phi_j(\boldsymbol{s})d\boldsymbol{s}$($\nabla$ denotes the gradient) for $i = 1, ..., G$ and $j = 1, ..., G$. To obtain the resulting precision matrix, Neumann boundary conditions have been used(Lindgren, Rue, and Lindström 2011). For the one dimensional problem considered in this paper, the number of nodes are equal to the number of locations considered, such that the dimension of $\boldsymbol{Q}$ is not reduced by the basis representation.

With the sparse precision matrix $\boldsymbol{Q}(\kappa, \tau)$ and Equation (4.23), we get an approximation to the solution of the SPDE 4.22 in the form of a Gaussian Markov Random Field(GMRF)(Rue and Held 2005b). This means that instead of having a dense covariance matrix $\Sigma = \boldsymbol{Q}^{-1}$ which is needed to represent a GRF, a sparse precision matrix is approximated. This results in reduced computational cost for matrix operations(Lindgren and Rue 2013).

4.5 The deviance information criterion

The deviance information criterion(DIC), proposed by Spiegelhalter *et al.* (Spiegelhalter et al. 2002), is the most commonly used measure of model fit based on the deviance for Bayesian models. It is a generalization of the Akaike information criterion(AIC), developed especially for Bayesian model comparison. The DIC is based on two components, one for quantifying the model fit and the other for evaluating the complexity of the model. The first component is measured through the posterior expectation of the deviance $D(\boldsymbol{\theta}) = -2\log(p(\boldsymbol{y}|\boldsymbol{\theta}))$, where $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood function. The second component is measured through the effective number of parameters, which is defined as:

$$p_D = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(D(\boldsymbol{\theta})) - D(\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta})) = \bar{D} - D(\bar{\boldsymbol{\theta}}). \tag{4.25}$$

The effective number of parameters is therefore defined as the posterior mean of the deviance minus the deviance of the posterior mean of the parameters. Meng and Rubin (MENG and RUBIN 1992) shows that such a difference is the key quantity in estimating the degrees of freedom of a test, and is a good measure for the complexity of a model. The DIC is

$$\mathrm{DIC} = \bar{D} + p_D, \tag{4.26}$$

where, as with the AIC, models with smaller DIC are better supported by the data. $\bar{D}$ measures how well the model fits the data, while p_D penalizes the complexity of the model. Note that INLA, instead of evaluating the deviance at the posterior mean of all parameters, uses the posterior mode of the hyperparameters $\boldsymbol{\psi}$. The reason is that the posterior marginals for some hyperparameters(especially the precision) might be skewed, such that the posterior mean might not be a good representation of the distribution's expectation (Blangiardo and Cameletti 2015), such that the mode is preferred.

5 Models and necessary SPDE parameter estimations

The latent Gaussian models for the linear predictor η_{ij} we consider in this paper, are

$$\begin{aligned} &1. \quad \eta_{ij} = \alpha_i + \beta_i k_j \\ &2. \quad \eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) \\ &3. \quad \eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij} \\ &4. \quad \eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}. \end{aligned} \tag{5.1}$$

α_i is the mean of the group of healthy people, β_i is an added effect for the group of people with Schizophrenia, k_j is a factor indicating if person j is in group 1 or 2, $\xi_j(s_i)$ is a SPDE realization explaining the spatial dependency between locations and ϵ_{ij} is an independent, identically distributed random effect. For each of these predictors, we try two different likelihood functions:

$$\begin{aligned} &\pi(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\phi}) = \prod_{i=1,j=1}^{n,N} \pi(y_{ij}|\mu_{ij}, \phi_i) = \prod_{i=1,j=1}^{n,N} \text{Beta}(\mu_{ij}, \phi_i), \\ &\text{with} \quad 1. \quad \phi_i = t_i \exp(\theta), \ (t_i \text{ equal } \hat{\phi}_i, \text{ and } \theta \text{ fixed as } 0) \\ &\qquad\quad 2. \quad \phi_i = \phi (\text{equal for all locations}). \end{aligned} \tag{5.2}$$

With these linear predictors and likelihoods, we want to investigate which specifications that give the best fit to the data. We also want to look into how utilizing the spatial dependency changes the results concerning differently methylated CpG sites. In Chapter 6, the different models are fitted to the case study of the Schizophrenia data set, while we in Chapter 7 evaluate some simulations.

Within the Bayesian framework we need to specify prior distributions for the four latent Gaussian models considered. For α_i and β_i, we choose non-informative priors equal to Normal$(0, 10^6)$. This is the default prior for the regression parameters in INLA, and such a vague prior distribution is typically chosen in the absence of information prior to the investigation. For the linear predictor 1, this results in a multivariate Normal distribution with mean $\mathbf{0}$ and precision matrix $\boldsymbol{Q}$ with elements only at the diagonal.

For the linear predictor 2, we need to specify priors for $\xi_j(s_i)$ and the SPDE parameters κ and τ. As explained in the Section 4.4, $\xi_j(\boldsymbol{s})$ is a Gaussian Markov random field with mean $\mathbf{0}$ and precision matrix specified by the parameters κ and τ, $\boldsymbol{Q}(\kappa, \tau)$. In this case, $\boldsymbol{Q}$ do not only have elements at the diagonal, such that interactions between sites can be taken into account. Prior specifications for κ and τ are found by doing some exploratory analysis, explained in Section 5.1. With

this analysis we choose starting values for the fitting process, τ_0 and κ_0, and in agreement with the notation used in INLA, we specify a prior distribution on the log transform of the parameters. This results in $(log(\tau), log(\kappa)) \sim \Big(\text{Normal}(\log(\tau_0), 0.1), \text{Normal}(\log(\kappa_0), 0.1)\Big)$, where the precision is set to 0.1. 0.1 is used as the precision to ensure a wide search area for the most fitting values, and to lower the influence of the prior selection.

For the linear predictor 3, we have added an independent, identically distributed random effect. For this effect, we assume a prior distribution equal to $\text{Normal}(0, \sigma^2)$, where we are interested in the precision parameter $\frac{1}{\sigma^2}$ as well. The prior distribution for the hyperparameter $\frac{1}{\sigma^2}$, is $\text{Gamma}(1, 0.00005)$. This means that we consider ϵ_{ij} as a random variable from the same distribution for all locations and persons.

In the fourth linear predictor, we have all the terms combined. This is the most complex model we evaluate, and we use the same priors as described above.

For the likelihood function, we consider two different approaches for the precision parameter ϕ. This is done to investigate what assumptions that seem to fit the data the best. As described in the exploratory analysis, the precision parameter ϕ seems to be varying for each position i. To consider the possibility of a different ϕ for each location, a new property for the beta family in INLA is used. Instead of assuming a prior on ϕ, we assume a prior on θ, where $\phi = t_i \exp(\theta)$. With the scales t_i we scale the distribution, and by considering θ as fixed, we can look into the assumption of a non-stationary, known ϕ. By specifying $\theta = 0$, and the scales t_i equal to the estimated ϕ_is,

$$\hat{\phi}_i = \frac{\hat{\mu}_i(1-\hat{\mu}_i) - \hat{\sigma}_i^2}{\hat{\sigma}_i^2}, \tag{5.3}$$

we treat the precision at each location as known.

In the second likelihood model, we treat the precision parameter ϕ as a random variable that is equal for each location. This is done, since it is possible to achieve a data set where the empirical ϕ_i estimates are varying in a way that seems to be best described by a non-stationary ϕ, with a stationary ϕ. As shown in the Figure 7.1, we see that the data set created with a constant ϕ and errors in the mean of the data created by a SPDE effect and an independently, identically distributed random effect, does seem to have a location dependent precision. We therefore want to investigate the possibility of describing the methylation data set with a constant, stationary ϕ.

A likelihood specified by scales t_i equal to $\hat{\phi}_i$ and treating θ as a random variable to be estimated, has also been investigated. However, a problem occurs when evaluating the different linear predictors. Since the SPDE effect and the iid random

effect seem to change the behavior of the spread of the data differently at different positions i, the scales for the distribution of ϕ did not seem to be fitting for all the linear predictors, and problems with the fitting process occurred. Therefore, we chose to only consider the two likelihoods specified in 5.2.

Even though the acf plot in Appendix A.1 indicates that each person might have a dependency structure with different parameters τ and κ, we choose to look into models where the realizations $\xi_j(\boldsymbol{s})$ come from the same SPDE. This is done to simplify the models and simulations considered. We also choose $\lambda = 1$, and use this as the extra smoothing parameter. Throughout the rest of the paper, we denote σ_0^2 as the variance of the white noise process $\omega(s)$, which is a SPDE parameter found in Equation (4.22). This is to avoid confusion with the variance in the independently, identically random effect distribution, Normal$(0, \sigma^2)$. We also denote the linear predictor 1 as the simple model, the linear predictor 2 as the SPDE model, the linear predictor 3 as the iid error model, and the linear predictor 4 as the SPDE and iid error model. Note that this is only a naming convention to improve the readability.

5.1 Estimation of SPDE parameters

As explained in the previous Section, we need some prior specifications for the SPDE hyperparameters τ and κ. We therefore do some more exploratory analysis on the Schizophrenia data set, to obtain estimates for τ_0 and κ_0. Since these parameters have complicated definitions, we estimate the range parameter r, which is defined as the distance at which the correlation is approximately 0.1. The link between r and κ is found in Equation (4.21). When κ_0 is chosen, we vary τ_0, which controls the variance, until our samples are reasonable compared to the methylation data. As described in (Blangiardo and Cameletti 2015), the SPDE approach is prior sensitive, such that the prior specifications might affect the posterior results. In Chapter 7, we have tested the prior sensitivity for some simulations.

To estimate r, we focus on the lag 1 correlation in the data set. Since we have a lot of locations with different distance to the nearest neighbor, this should be enough correlations to be able to obtain an acceptable estimate of the range. We therefore calculate the mean correlation lag 1 between all CpG sites, based on all the people. As mentioned earlier, we look into models with $\xi_j(\boldsymbol{s})$ with the same SPDE parameters, such that we use the mean correlation to find an estimate of the center of the prior distributions for τ and κ. By having a large but reasonable precision parameter for the prior distributions, the posterior distributions of the parameters are able to vary towards the more fitting values based on the data.

In Figure 5.1, the mean correlation lag 1 between CpG sites on chromosome 1 that have a distance between them less than or equal to the Distance, are displayed. As we can see from the blue numbers, at least the first half has a lot of correlations

to take an average over. The first time the correlation is close to 0.1(0.0985) are for Distance $r = 650$. In Table 5.1, the r estimate based on data from each chromosome are given. The average estimate is equal to $\bar{r} = \frac{15200}{22} \approx 700$, which we use as an estimate for the priors. The range estimate varies a bit between the chromosomes, but 700 seems as an estimate we could use for all chromosomes. Since we are estimating by using the range at which the correlation drops below 0.1, some of the variation in the table can be accounted for by the fact that the correlation is not equal for each of the estimates, but approximately equal to 0.1. Although this is the case, we have no way of knowing if the range estimate should be equal at each chromosome.

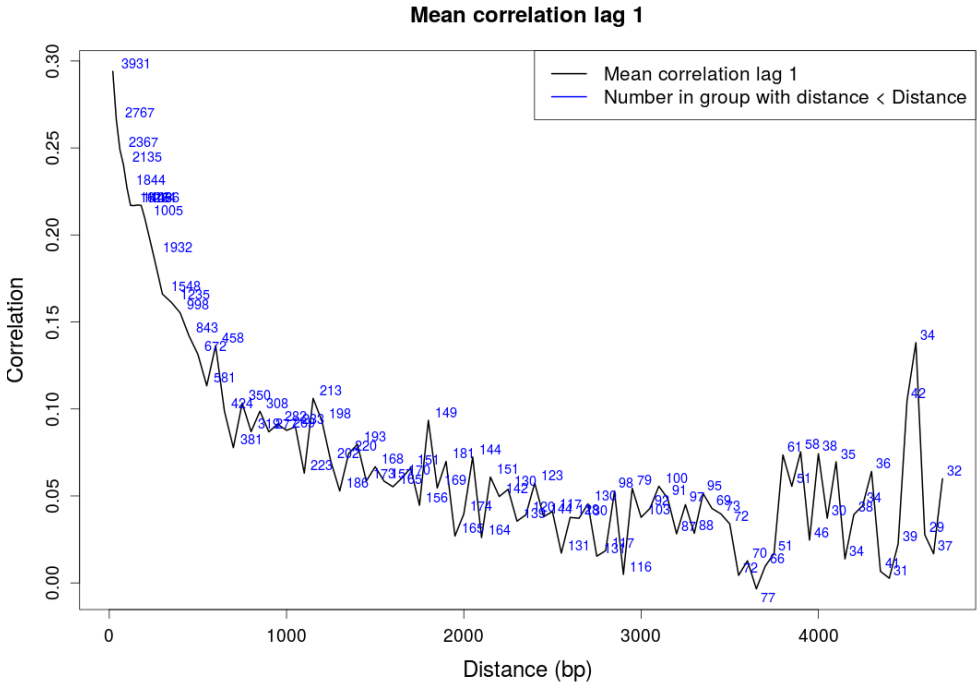


Figure 5.1: Mean correlation lag 1 of methylation data from chromosome 1. The blue numbers are the amount of correlations the average is taken over, between CpG sites with a distance between them less than the distance displayed, and bigger than the previous distance. The distance vector increases with 50 base pairs for each point.

Table 5.1: Table of estimated range parameter r with data from the different chromosomes.

Chromosome	Range estimate r
1	650
2	700
3	600
4	750
5	750
6	700
7	650
8	750
9	650
10	700
11	800
12	650
13	650
14	800
15	550
16	700
17	750
18	600
19	700
20	700
21	700
22	700

6 Results; case study

In this Chapter, we provide the results obtained using Bayesian regression with INLA on the Schizophrenia data set, for models and specifications described in Chapter 5.

Because of the computational cost of fitting a model with spatial dependency, we focus on a randomly chosen area of subsequent locations at chromosome 6, from CpG site 201 to 600. This results in $n = 400$ sites, and with $N = 95$ persons there are 38000 locations to be considered. In Figure 6.1, we have plotted the methylation data and the CpG islands found in this region.

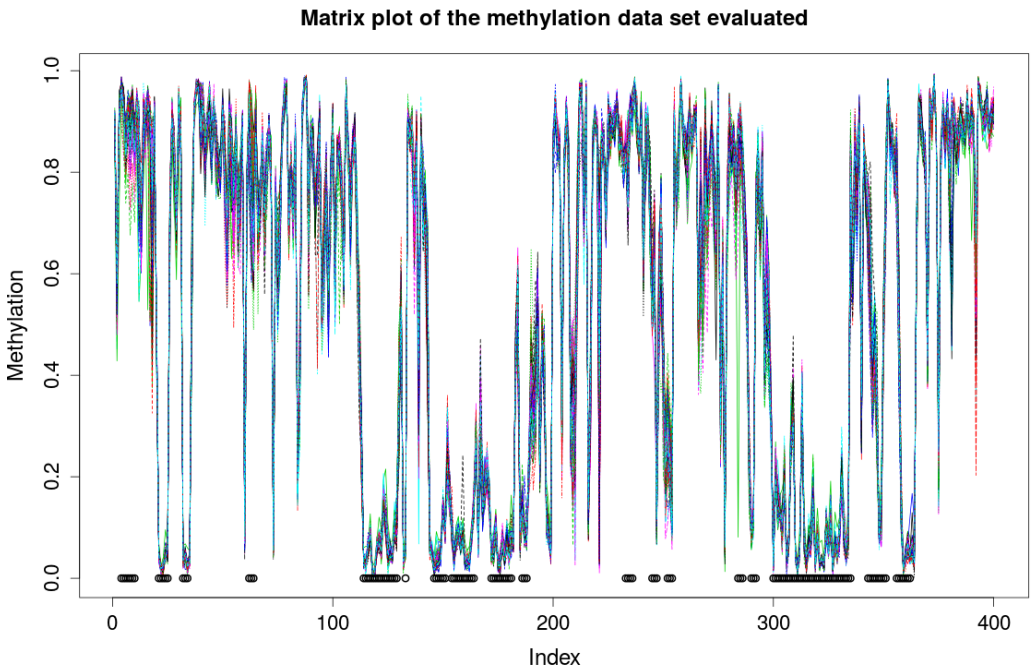


Figure 6.1: Matrix plot of the methylation data for all the 95 persons from chromosome 6. The black dots are sites that are part of a CpG island.

Table 6.1: Table of the linear predictors considered and the resulting deviance information criterion with likelihood function 1.

Lin. Pred.	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$	$\eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij}$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$
DIC	−181911.0	−183501.0	−181910.0	−183475.0

Table 6.2: Table of credibility intervals for the hyperparameters τ, κ, σ_0, r and $\frac{1}{\sigma^2}$, considered with likelihood function 1.

	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$	$\eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij}$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$
τ		[20598, 33419]		[15083, 20164]
κ		[0.000272, 0.000413]		[0.000365, 0.000473]
σ_0		[0.038, 0.095]		[0.056, 0.096]
r		[6841, 10418]		[5982, 7742]
$\frac{1}{\sigma^2}$			[4377, 82747]	[14357, 118960]

6.1 Schizophrenia data set

The first likelihood we consider is

$$\prod_{i=1,j=1}^{n,N} \text{Beta}(\mu_{ij}, \phi_i), \tag{6.1}$$

with $\phi_i = t_i \exp(\theta)$, with scales $t_i = \hat{\phi}_i$ and $\theta = 0$. We therefore treat the precision parameter as location dependent and known. In Table 6.1, we have displayed the deviance information criterion obtained by considering the four different models of the latent Gaussian field, found in (5.1).

The linear predictor that results in the best fit is $\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$, with a DIC equal to −183501. Even though the DIC has a tendency of favoring more complex models(Spiegelhalter et al. 2014), a decrease of approximately 1600 should be enough to indicate a better fit. By adding the random effect ϵ_{ij} to the linear predictor, we see that the resulting DIC indicates worse fit compared to those linear predictors without this term.

In Table 6.2, we have given the estimated credibility intervals of the hyperparameters obtained with the different linear predictors. If we consider the precision parameter for the iid random effect ϵ_{ij}, we see that it has a very wide credibility region in both of the models the term is included. This, together with the DIC, indicates that these models fit the data poorly, because of the uncertainty in the hyperparameters and the worse DIC. When looking at the credibility region for the

range parameter in the SPDE model, we see that it is much higher than the range estimated from Section 5.1. This might indicate that a non stationary approach for the SPDE could be reasonable, especially if we evaluate a larger area of the chromosome. This is discussed in more detail in Chapter 8.

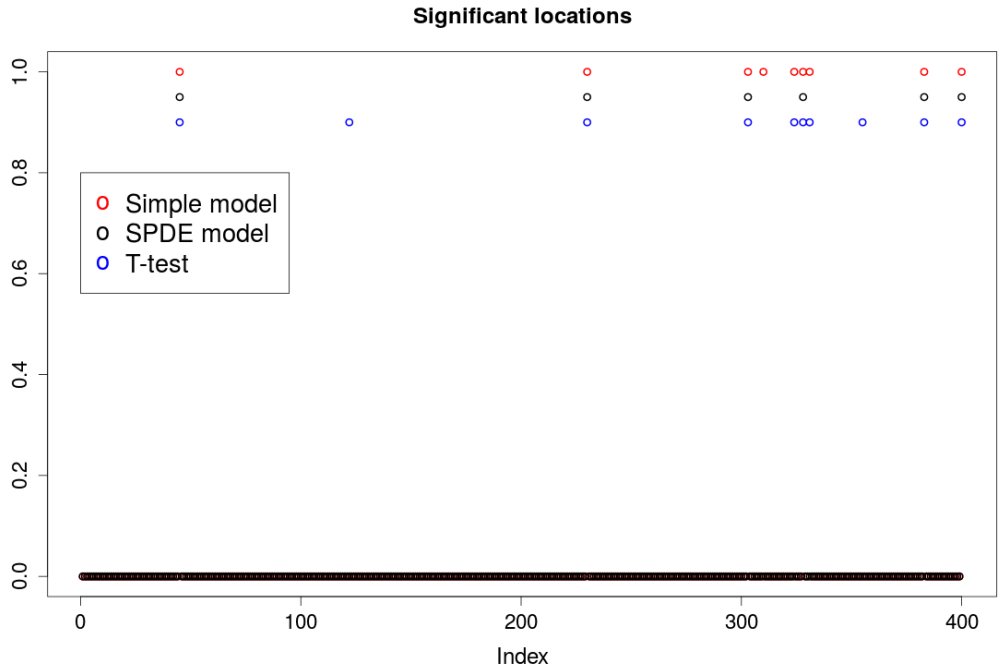
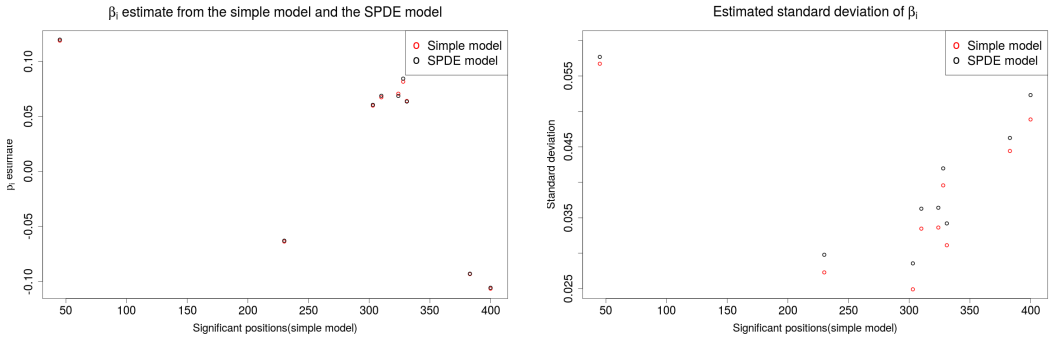


Figure 6.2: Differently methylated positions between Schizophrenia and Healthy patients found with the simple model and the SPDE model with likelihood function 1, and the T-test. Dots that are not equal to 0 indicate differently methylated positions.

In Figure 6.2, we have displayed the significantly different methylated positions found by considering the simple model and the SPDE model with likelihood function 1, and the ones obtained from the T-test. The significant effects β_i is defined as those effects that do not include 0 in the 95% credibility interval around its mean. With the simple model, we get almost the same positions defined as differently methylated as with the T-test. The differences can be explained by the fact that in the simple model, the beta distribution is taken into consideration and used to optimize the β_i values and the standard deviation estimates. This is not considered in the T-test.

When comparing the SPDE model with the simple model, we get fewer positions categorized as differently methylated, especially in the CpG island around position



(a) The estimated added effect β_i for each position that is classified as differently methylated for the simple model.

(b) The standard deviation of β_i at positions that are classified as differently methylated for the simple model.

Figure 6.3

320, which is the largest island displayed in Figure 6.1. As displayed in 6.3, this

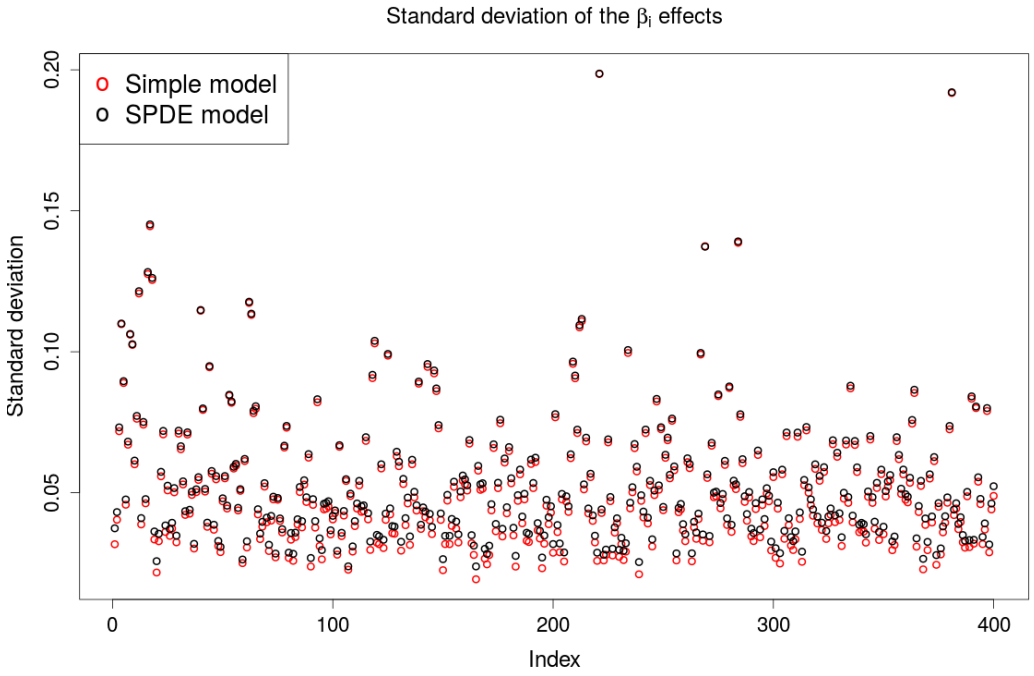


Figure 6.4: Standard deviation for the β_i effect from evaluating the simple model and the SPDE model with likelihood function 1.

seems be caused by the standard deviation of the β_i estimates which increases when the spatial dependency is taken into account and the precision ϕ_i is fixed. With the SPDE model, which are to describe some of the variation with spatially dependent

errors, the fixating of the precision parameter ϕ results in an over estimation of the uncertainty to the estimated effects β_i. We see that the estimated β_i values are slightly different, but the standard deviation is very different and consistently higher than for the simple model. This means that by fixating the precision and modeling with a SPDE structure, the amount of differently methylated positions will be reduced, because of the increase in uncertainty for the other estimates in the latent Gaussian model. This might lead to rejection of positions that are differently methylated, since the standard deviation is consistently over estimated, as shown in Figure 6.4. This lead to a lower type I error than specified, with only 6 out of 400 positions being classified as differently methylated on a 0.05 significance test.

In Figure 6.5, we have displayed the fitted $\xi_j(s_i)$ realizations around the CpG island found around position $i = 320$. We can see that the differently methylated sites within the island seem to be highly dependent on one person. It is therefore likely that the these sites should not be classified as differently methylated, and the SPDE model has made reductions of differently methylated sites in this area. But since the reason for this reduction is that the estimated standard deviation of each β_i estimate is raised, the conclusion might be drawn on wrong assumptions.

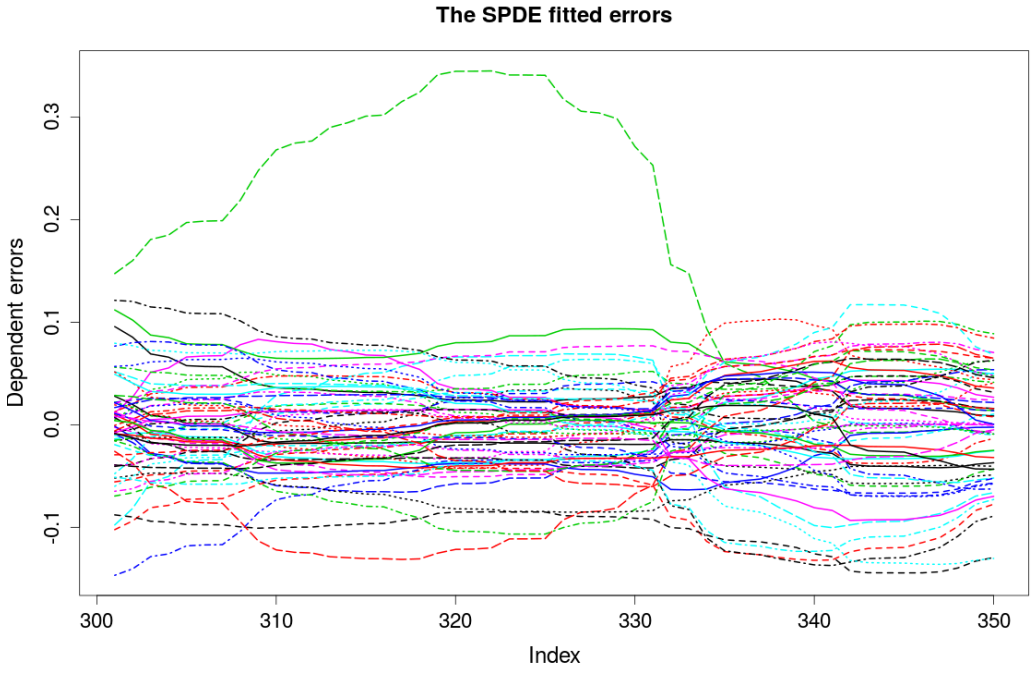


Figure 6.5: SPDE fitted errors. Area around the CpG island where the two significant locations are dropped.

Table 6.3: Table of the linear predictors considered and the resulting deviance information criterion with likelihood function 2.

Lin. Pred.	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$	$\eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij}$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$
DIC	−160120.0	−172089.3	−181496.6	−212238.0

The second likelihood function we consider, is

$$\prod_{i=1,j=1}^{n,N} \text{Beta}(\mu_{ij}, \phi), \tag{6.2}$$

with ϕ being a location independent, random variable. In Table 6.3, we have displayed the deviance information criterion obtained by evaluating the four different linear predictors.

When considering likelihood function 2, we do not get a good result for the DIC of the simple model $\eta_{ij} = \alpha_i + \beta_i k_j$, compared to the one obtained with likelihood function 1. With likelihood function 2 it is equal to −160120, and with likelihood function 1 it is −181911. This is caused by the fact that a location dependent variation in the spread of the data fits better than the assumption of a constant spread, such that a model which do not change the behavior of the variation benefits from a likelihood that does. With the SPDE model, we get a much better fit with a reduction in the DIC of 12000. Here we have taken the spatial dependency into account, and some of the variation is described by spatially dependent effects. Compared to the DIC value obtained by the SPDE model evaluated with likelihood 1, we see that it is much higher, indicating a worse fit. This is most likely caused by the fact that the location dependent variation is not yet well accounted for in the SPDE model with likelihood function 2.

With the iid error model, we get an interesting result concerning the DIC value. Compared to the simple model, we get a decrease of approximately 21000. This is almost equal to what we get when we evaluate the iid error model with likelihood function 1. This is caused by the fact that describing some of the variation with iid random effects in the linear predictor, results in describing some of the variation that is dependent on the mean, or in other words, the location. This is further described in the Chapter 7, and results in a much better fit than linear predictor 1 and 2, with likelihood function 2.

The last linear predictor under consideration is the most complex one, namely $\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$. With this SPDE and iid error model, we see that the resulting DIC indicates a much better fit than for all the other predictors considered, for both likelihoods. This model assumes that some of the variation is described by

Table 6.4: Table of credibility intervals for the hyperparameters ϕ, τ, κ, σ_0, r and $\frac{1}{\sigma^2}$, considered with likelihood function 2.

	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$	$\eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij}$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$
ϕ	[113.7, 117.0]	[191.9, 200.2]	[339.1, 344.9]	[965.1, 1065.3]
τ		[5041, 5744]		[6544, 8184]
κ		[0.000465, 0.000520]		[0.000461, 0.000564]
σ_0		[0.178, 0.227]		[0.115, 0.176]
r		[5438, 6084]		[5019, 6139]
$\frac{1}{\sigma^2}$			[34.56, 35.52]	[21.68, 22.56]

spatially dependent errors, some is described by the iid random effects in the linear predictor and the rest by a constant precision parameter ϕ. Note that some of the variation in the data is also described by estimating a different mean to the different groups.

In Table 6.4, we have given the credibility intervals for the hyperparameters of the different latent Gaussian models with likelihood function 2. From the simple model to the SPDE model, we see that some of the variation is explained by the spatially dependent effects, and the result is that the estimated credibility interval for the precision parameter ϕ is shifted to [191.9, 200.2] from [113.7, 117.0]. This is caused by the precision parameter ϕ being able to change, such that some of the variation can be absorbed by the spatially dependent process instead of only influence the other parameter estimations. The SPDE realizations also explain some of the location dependent variation, which fits the data better than a constant ϕ.

If we compare the resulting credibility interval for the range parameter r and σ_0 achieved with the SPDE model considered with likelihood 1 and 2, we see that the range estimate is considerably lower and the standard deviation σ_0 larger with likelihood 2. For σ_0, this can be intuitively explained by the fact that the fitting process for likelihood function 1 has a restriction on the precision parameter ϕ for each location, such that the amount of variation in the SPDE model that fits the data the best, is reduced. Because none of the variation can be absorbed by the spatially dependent errors, the resulting σ_0 estimate is lower than for considering a likelihood where the precision parameter has the possibility to change. However, the SPDE model with likelihood function 1 has a considerably lower DIC value than the one obtained with likelihood function 2. This is caused by the location dependent variation of the data being better described by the SPDE model and likelihood function 1.

With the iid error model, the variability caused by the precision parameter ϕ in the beta distribution is greatly reduced, and the credibility interval is shifted upwards to [339.1, 344.9]. The estimated credibility interval for the precision parameter $\frac{1}{\sigma^2}$

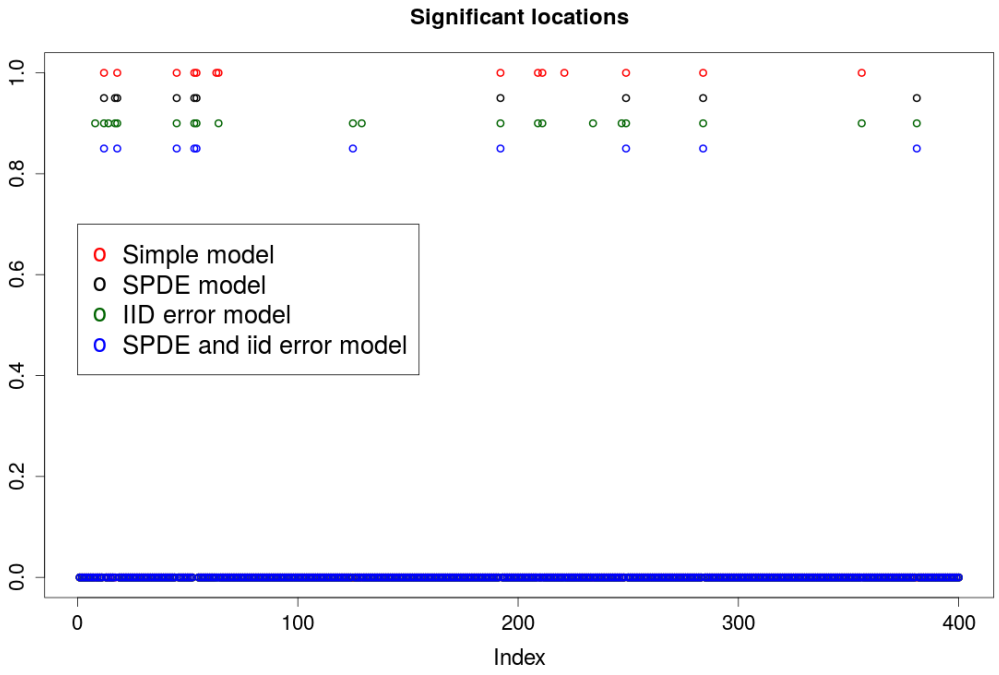
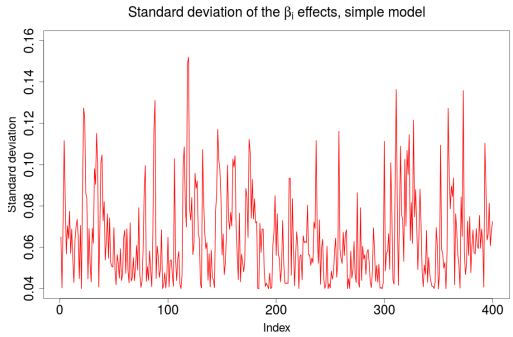


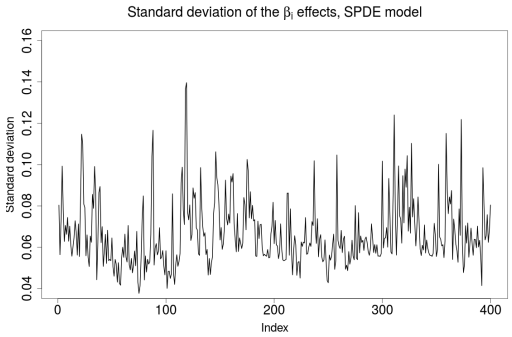
Figure 6.6: Differently methylated positions between Schizophrenia and Healthy patients found with the different linear predictors and likelihood function 2. Dots that are not equal to 0 indicate differently methylated position.

of the Normal error distribution is $[34.56, 35.52]$, indicating a standard deviation estimate of $[0.168, 0.170]$. When considering the SPDE and iid error model, we see that the credibility interval for $\frac{1}{\sigma^2}$ is lowered to $[21.68, 22.56]$, resulting in a standard deviation estimate $[0.211, 0.215]$. This means that more of the variation is being explained by the independently, identically distributed random effects when a spatial structure is considered as well. For the spatial dependent process, the credibility interval for the standard deviation σ_0 estimate is reduced to $[0.115, 0.176]$ and the credibility interval for the range is wider than when considered without the iid random effects. This is caused by the fact that having more parameters to be estimated that can interact with each other usually increases the uncertainty of the estimates. With the SPDE and iid error model, the credibility interval for the precision parameter ϕ is increased in both uncertainty and value to $[965.1, 1065.3]$, which is as expected when adding two random effects to the linear predictor that can explain some of the underlying variation in the data.

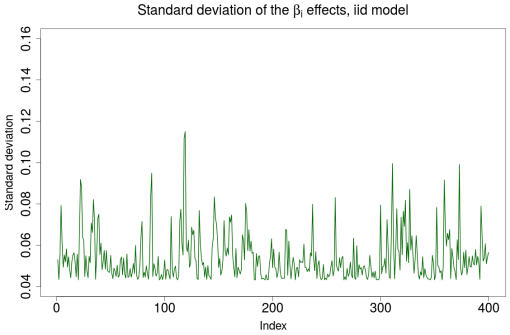
In Figure 6.6, we have displayed the differently methylated positions found with the linear predictors and likelihood function 2. Some of the positions found are equal to those in Figure 6.2, but most of them are different. This is mainly caused



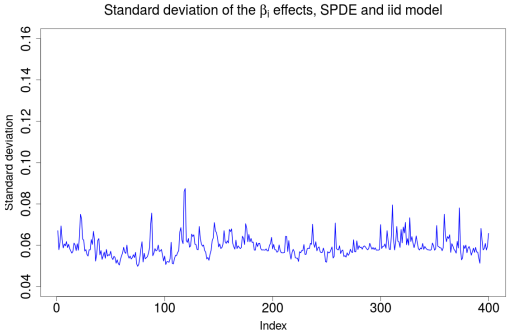
(a) Standard deviation of the β_i effects in the simple model.



(b) Standard deviation of the β_i effects in the SPDE model.



(c) Standard deviation of the β_i effects in the iid error model.



(d) Standard deviation of the β_i effects in the SPDE and iid error model.

Figure 6.7: The standard deviation of β_i in the different models considered for likelihood function 2.

by the fact that the likelihood function is changed. If we compare the number of significant positions obtained by the iid error model and the ones obtained with the SPDE and iid error model, we see that there is a reduction in differently methylated locations, and that this occurs usually around places where there are many differently methylated positions found by the iid error model. To explain this, we consider the following example.

If we have a data set without spatially dependent errors, the differently methylated positions found would be independent of the neighboring positions. However, for a data set containing spatial dependency such as this, we see from Figure 6.5 that a significantly different methylated position affects and is affected by neighboring positions as well, such that they often come in clusters. By taking the spatial dependency into account, we can utilize the information that lies in the correlated data to obtain a better fit and be able to reduce the amount of such clusters found.

The reason for mainly considering the last two models is that for likelihood function 2, these models fit the data best. With the iid random effect we describe some of the location dependent variation, while with the last model, we include the spatial dependency as well. It is worth noticing that none of the models considered here treats the locations inside the island around point 320 as differently methylated, which from Figure 6.5 seems to be a wrong conclusion.

In Figure 6.7, we have displayed the standard deviation of the β_i estimates with the linear predictors and likelihood function 2. By considering the SPDE model against the simple model, we see that the variation in the standard deviation is less varying for the SPDE model. Here we have taken the spatial dependency into account and some of the location dependent variation. The same is true by considering the iid random effect model against the simple model, where we account for even more of the location dependent variation. By looking at the most complex model, the combination of the SPDE and iid random effect model, we see that it produces the least varying standard deviation of the β_i estimates. We see that by accounting for the spatial dependency and some of the location dependent variation, we have quantified most of the site specific variation caused by these quantities, such that smoother standard deviation estimate of the β_i estimates can be found. The different models influences the β_i estimates as well, as displayed in Figure 6.8.

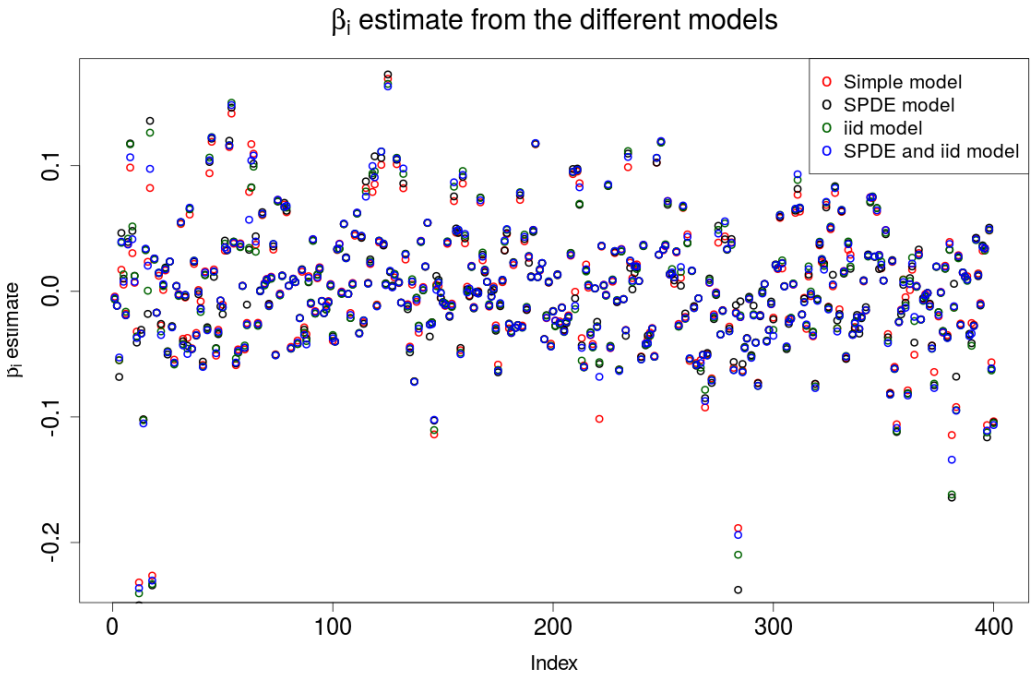


Figure 6.8: β_i estimate from the different models.

7 Simulation study

7.1 Motivation and Creation

To further explore the statistical methodology related to the results obtained in the previous Chapter, we consider some simulations. With these simulations, we want to look into:

1. What happens if we fit a latent Gaussian model with a spatial dependency effect to a data set without spatial dependency?

2. How do the credibility intervals obtained for the hyperparameteres look like when fitted to a data set created with a SPDE structure based on the same parameters τ and κ?

3. How does prior specifications for the hyperparameters of the resulting model obtained in the previous chapter, $\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$ with likelihood function 2, affect the posterior estimates of these parameters?

4. How does utilizing the spatial dependency in a data set affect the β_i parameter estimations and the differently methylated positions found?

For question 1 and 2, we are only considering one simulated data set for each, and for question 3 and 4, we are considering four data sets. These limitation of the simulation study is caused by the computational cost of the fitting process.

With the first simulated data set, we want to look into question 1. Therefore, this data set is simulated by having no spatial dependency. Here we only consider one of the likelihoods, namely the first one where we treat ϕ_i as known. We also only consider model one and two for the latent Gaussian field, since we are interested in the effect of fitting a SPDE structure to a data set without spatial dependency. To simulate the data, we use the estimated mean $\hat{\mu}_i$ from each location in the real data set we are evaluating, such that it is closely related to the real methylation data. For each person at each location, we then calculate a_i and b_i with formula

$$\begin{aligned} a_i &= \mu_i \phi_i \\ b_i &= \phi_i - \phi_i \mu_i, \end{aligned} \tag{7.1}$$

where a_i and b_i are the parameters to be used to draw a random beta distributed variable y_{ij} for each person j at location i. ϕ_i is here drawn from a gamma$(2, 0.007)$ distribution, which results in ϕ_is on the same scale as the ones in the real data set. The results of the simulation is evaluated in 7.2.1.

With the second simulated data set, we look into question 2. To simulate the data, we use the following procedure:

$$\begin{aligned} \eta_{ij} &= \text{logit}(\hat{\mu}_i) + \xi_j(s_i) \\ \mu_{ij} &= \text{logit}^{-1}(\eta_{ij}) \\ a_{ij} &= \mu_{ij}\phi_i \\ b_{ij} &= \phi_i - \phi_i\mu_{ij} \\ y_{ij} &= \text{rbeta}(a_{ij}, b_{ij}), \end{aligned} \tag{7.2}$$

where $\xi_j(s_i)$ is a realization of a SPDE with parameters $\tau = 10730$ and $\kappa = 0.0007$($r = 4000$ and $\sigma_0 = 0.07$), and ϕ_i is randomly chosen by a uniform($min = 200, max = 400$) distribution. The resulting data set with y_{ij} as the methylation value for person j at location i, has a variation that is caused by both ϕ_i and the SPDE realizations. For this simulated data set, we fit the latent Gaussian model $\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$ with likelihood function 1, with a slightly different approach. Instead of using the estimated ϕ_is as the scales t_i, we use the true ϕ_is, and we let θ be a random variable to be estimated with prior distribution logGamma$(1, 0.1)$. This means that the distribution of θ is scaled with the real weights, and we expect the posterior estimate of θ to be around 0. This is an approach that is not possible to do for the Schizophrenia data set, since we do not know the real scaling of the distribution of ϕ. However, by looking into this model and simulation, we get an idea of how the credibility regions for the posterior estimates of the hyperparameters varies when the simulations are done by considering realizations of a SPDE model based on the same κ and τ parameters.

With the last four simulated data sets, we look into question 3 and 4. Here, we want to further investigate the parameter estimations and prior specifications for the best fitting model, $\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$ with likelihood function 2. We also compare the differently methylated positions found by the SPDE and iid error model with the T-test. The simulation of the data is done by a similar procedure as above:

$$\begin{aligned} \eta_{ij} &= \text{logit}(\hat{\mu}_i) + \xi_j(s_i) + \epsilon_{ij} \\ \mu_{ij} &= \text{logit}^{-1}(\eta_{ij}) \\ a_{ij} &= \mu_{ij}\phi \\ b_{ij} &= \phi - \phi\mu_{ij} \\ y_{ij} &= \text{rbeta}(a_{ij}, b_{ij}), \end{aligned} \tag{7.3}$$

where $\epsilon_{ij} \sim$Normal$(0, \sigma^2)$ and the precision parameter ϕ is equal for each location. The four simulations we consider are based on the parameters $r = 5700$, $\sigma_0 = 0.15$, $\frac{1}{\sigma^2} = 22$ and $\phi = 1000$. These are all values inside the credibility regions of the estimated hyperparameters from the Schizophrenia data set considered in Chapter 6.

For the first of these four simulated data sets, we run the fitting process with default priors for ϕ and $\frac{1}{\sigma^2}$, being Gamma($shape = 1, rate = 0.1$) and Gamma($shape = 1, rate = 5 \cdot 10^{-5}$). For the SPDE parameters τ and κ we use the priors specified by $\log(\tau) \sim$logNormal($\log(\tau_0), 1$) and $\log(\kappa) \sim$logNormal($\log(\kappa_0), 1$) where τ_0 and κ_0 are the correct values 7135 and 0.0005. For the second simulated data set, we change the prior specification on $\frac{1}{\sigma^2}$ to Gamma($shape = 1, rate = 0.1$), which varies closer to the real value 22. For the third simulated data set, we check the impact of changing the priors of the SPDE parameters to $\log(\tau) \sim$logNormal($\log(1878), 0.1$) and $\log(\kappa) \sim$logNormal($\log(0.004), 0.1$). Here we have used the same prior specification for $\frac{1}{\sigma^2}$ as in the second simulation. For the fourth simulated data set, we have used the same specification as for simulation 2, but with an initialization of the precision parameter ϕ equal to 1000.

The dimension of all the simulated data sets are equal to the Schizophrenia data set, which results in $n = 400$ locations and $N = 95$ people, with $N_1 = 62$ classified as having Schizophrenia and $N_2 = 33$ being Healthy. We have added some small values to the mean of the Schizophrenia group at three locations, such that the groups have a differently methylated site at position $i = 100, 250, 300$. The values that are added are on logit scale equal to $(-0.10, 0.3, 0.5)$.

7.2 Results

7.2.1 No spatial dependency

As we can see from table 7.1, the resulting DIC value does not indicate a better fit with the more complicated model containing the spatial dependency, although the values are close to equal. But as we can see from the posterior hyperparameter credibility intervals for τ and κ, given in Table 7.2, they are very wide. This means that there is a high uncertainty in the estimates that are used in the resulting fit. Based on the worse DIC value and the wide credibility regions, we would prefer the simpler model over the SPDE model.

Table 7.1: Table of the linear predictors considered and the resulting deviance information criterion.

Lin. Pred.	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$
DIC	−200072.1	−200072.3

With both the likelihood functions we consider for the real data set, we get a lower DIC when taking the spatial dependency into account in the latent Gaussian

Table 7.2: Table of credibility intervals for the hyperparameters τ and κ.

	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$
τ	$[1237, 305466]$
κ	$[0.000601, 0.128125]$

model. The credibility regions for the posterior hyperparameter estimates are also much narrower, which indicates less uncertain estimates than for a data set without spatial dependency. This indicates a better fit to the data, and the assumption of spatial dependency in the Schizophrenia data set are strengthened.

7.2.2 Spatial dependency and non stationary ϕ

In Table 7.3, we have displayed the resulting credibility intervals of the hyperparameters $\exp(\theta)$, τ, κ, r and σ_0. The true parameters for the data set are $\theta = 0$, $\tau = 10730$, $\kappa = 0.0007$ $r = 4000$ and $\sigma_0 = 0.07$, such that all the true parameters are within the 95% credibility regions.

Table 7.3: Table of credibility intervals for the hyperparameters $\exp(\theta)$, τ, κ, r and σ_0.

	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$
$\exp(\theta)$	$[0.990, 1.017]$
τ	$[9368, 10844]$
κ	$[0.000667, 0.000758]$
r	$[3731, 4241]$
σ_0	$[0.065, 0.085]$

The credibility intervals of the hyperparameters for the simulation considered are narrower than the ones obtained for the Schizophrenia data set. This indicates that the Schizophrenia data set might be better described by a latent Gaussian model which allows the SPDE parameters for each person to be different. This is further discussed in Chapter 8.

7.2.3 Simulations of the complex model

The last simulations we want to look into are of $\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$ with a constant precision parameter ϕ. Here we want to investigate how not accounting for the spatial dependency and the location dependent variation influences the β_i estimates and their standard deviation. As described in Chapter 7, we have added some values to the mean at three locations for the group of simulated Schizophrenia patients. These are found at position $100, 250, 300$, and the values added are on logit scale equal to $(-0.10, 0.3, 0.5)$. With the different linear predictors, we want to look into how the prior specifications alters the posterior parameter estimations, and we want to compare the amount of differently methylated positions for the SPDE and iid error model and the T-test.

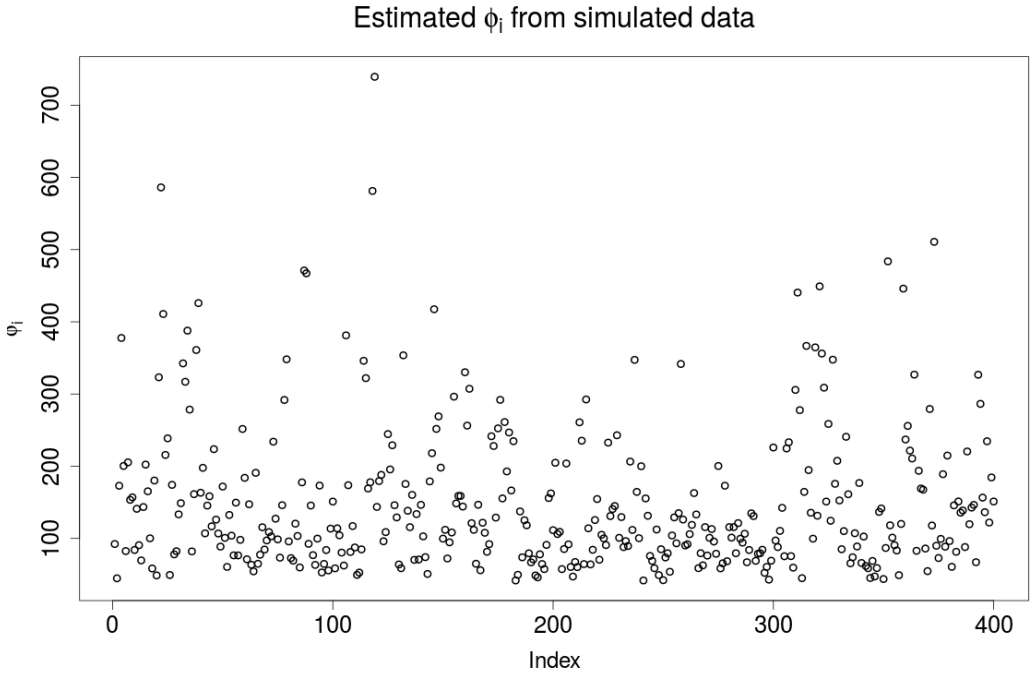


Figure 7.1: $\hat{\phi}_i$ calculated from the first simulated data set.

In Figure 7.1, we have displayed the resulting estimated ϕ_i, $\hat{\phi}_i$, for each location i for simulation 1. The precision parameters seem to be location dependent, caused by the realizations of $\xi_j(s_i)$ and ϵ_{ij}. The realizations of the SPDE model over the irregular grid, results in different amount of variation dependent on the density of positions. For the ϵ_{ij} which are random effects of the same distribution, the constant spread variable σ^2 does not imply a constant spread for the beta distribution. Since the variance in the beta distribution is dependent on the mean, a constant spread in the linear predictor η_{ij} will result in a variation in spread for the data dependent on

the mean, or in other words, the locations.

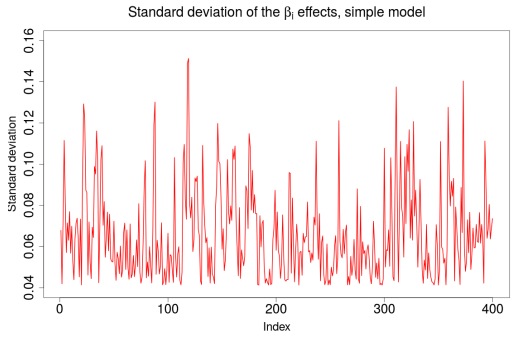
Table 7.4: Table of the linear predictors considered and the resulting deviance information criterion with likelihood function 1, for the simulated data sets 1, 2, 3 and 4.

Lin. Pred.	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$	$\eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij}$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$
DIC 1	−156585.4	−166504.9	−201831.1	−205118.8
DIC 2	−157095.5	−166843.9	−200851.4	−207353.1
DIC 3	−157440.5	−166726.6	−203934.3	−204909.9
DIC 4	−156642.9	−166444.1	−202240.6	−204996.3

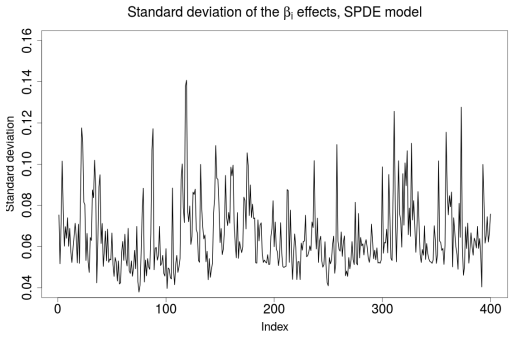
As we can see from Table 7.4, the DIC decreases from the linear predictor 1 to the linear predictor 4 for all simulations. This is as expected from the creation of the data and follows the same pattern as for the Schizophrenia data. From the simple model to the SPDE model, we get a decrease around 10000 for each of the simulations, which is slightly lower than for the real data. This might point towards a slightly larger dependency in the Schizophrenia data than the one used for the simulations. The decrease from the simple model to iid error model is much higher than for the Schizophrenia data. The decrease is around 45000, which is more than twice the decrease obtained in the Schizophrenia data. With the SPDE and iid error model, we get a decrease around 48000 – 50000 compared to the simple model, which is slightly lower than the decrease in the Schizophrenia data(52000).

In Table 7.6, we have given the credibility intervals of the hyperparameters for all the simulations. In simulation 1, 2 and 4, we see that the credibility intervals for τ and κ with model 4 contains the true parameters, except for τ in simulation 4, which is slightly underestimated. These simulations had prior distribution centered around the correct values. For simulation 3, we see that the new prior specification does seem to have a influence on the resulting fit, where the credibility interval for τ and r are slightly lower than the correct values, and the credibility interval for κ is slightly larger. With the different prior specification on $\frac{1}{\sigma^2}$ and the initialization of ϕ, clear differences can not be spotted for the simulations considered.

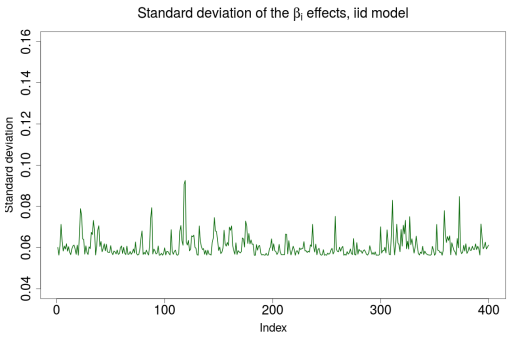
A clear difference between the case study and the simulations is the fitting of the linear predictor 3. In the simulations, we see that the iid random effect process absorbs much more of the variation than for the Schizophrenia data. This results in a more uncertain precision parameter estimate, but a much higher DIC value. The estimated credibility region for the precision parameter $\frac{1}{\sigma^2}$ is below the true value, indicating that more of the variation is described by the independently, identically distributed random effect than what the simulations where created with. Therefore, the different hyperparameters seem to be interfering with each other, such that the



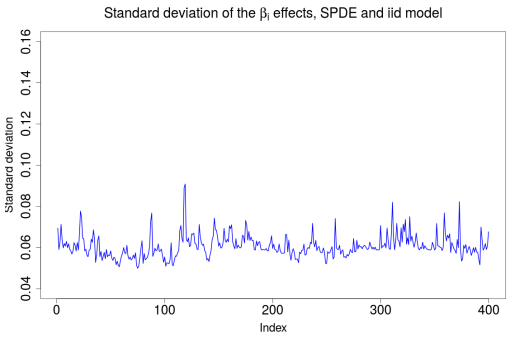
(a) Standard deviation of the β_i effects in the simple model.



(b) Standard deviation of the β_i effects in the SPDE model.



(c) Standard deviation of the β_i effects in the iid error model.



(d) Standard deviation of the β_i effects in the SPDE and iid error model.

Figure 7.2: The standard deviation of β_i with the different models considered on simulation 4. The other simulations give similar results.

identifiability of the parameters is reduced. With the linear predictor 4, we see that most of the hyperparameters for the SPDE model is inside the 95% credibility interval, but the precision parameters ϕ and $\frac{1}{\sigma^2}$ are slightly under and over estimated, respectively.

In Figure 7.2, we see the estimated standard deviation of the β_i estimates for simulation 2. The other simulations resulted in approximately equal figures, with only some small variations. The Figure is very similar to the one displayed for the results of the Schizophrenia data, 6.7, except for the linear predictor 3. The difference is caused by the under estimation of the precision parameter $\frac{1}{\sigma^2}$, resulting in a smoother estimation of β_i's standard deviation. With the most complex model, $\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$, the figures are almost identical.

In Figure 7.3, we have displayed the significantly different positions found with the SPDE and iid error model and the T-test. With three positions having different

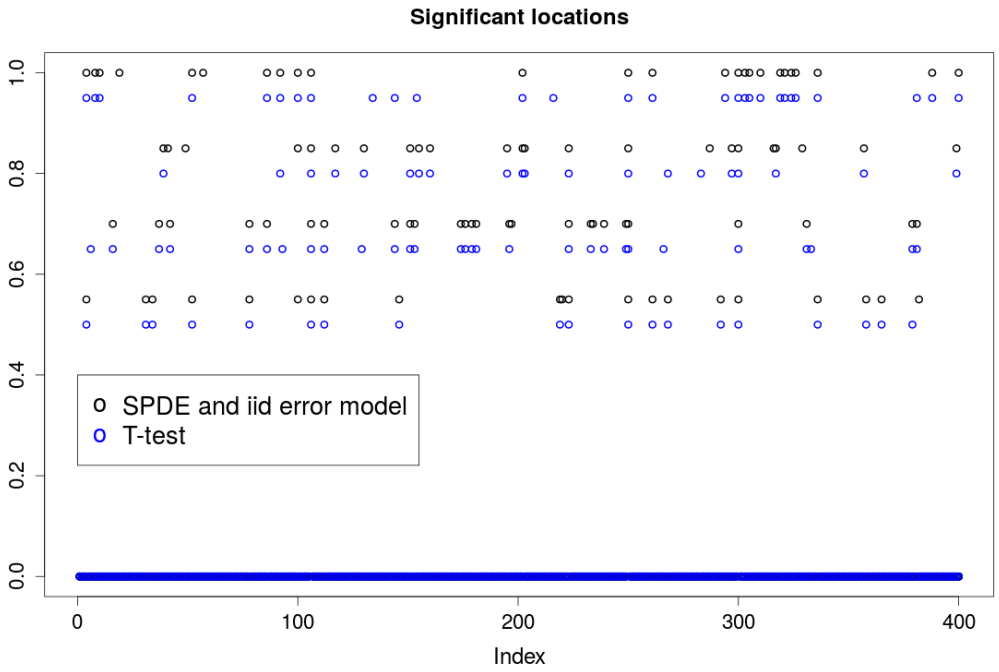


Figure 7.3: Significantly differently methylated positions found in the simulated data set 1, 2, 3 and 4, with the SPDE and iid error model and the T-test. Dots that are not equal to 0 indicate differently methylated positions.

Table 7.5: Amount of differently methylated positions found with the two methods.

	Simulation 1	Simulation 2	Simulation 3	Simulation 4
SPDE and iid error model	25	23	26	21
Amount	0.0625	0.0575	0.065	0.0525
Correct DMP found	3	3	2	3
T-test	28	20	29	19
Amount	0.07	0.05	0.0725	0.0475
Correct DMP found	2	3	2	2

mean, we would expect $3 + 0.05 \cdot 397 \approx 23$ positions to be differently methylated on a significance level 0.05, if the different tests were to find these three positions. This results in an amount equal to 0.0575.

In Table 7.5, we have displayed the amount of differently methylated positions found with the SPDE and iid error model and the T-test, evaluated on the simulated data. They are all reasonable amounts of differently methylated positions, when 400 locations are considered(see appendix A.2), and 3 positions are set to be different.

With the SPDE and iid random effect model, we find all three positions that were set to be differently methylated, except for position 100 in simulation 3. With the T-test, we only find the position 100 as differently methylated in simulation 2. Therefore, to utilize the spatial dependency seems to be able to find small differences better. Although the latent Gaussian model which takes the spatial dependency into account performs better than the T-test, four simulations are to few to draw any conclusion.

Table 7.6: Table of credibility intervals for the hyperparameters ϕ, τ, κ, σ_0, r and $\frac{1}{\sigma^2}$, for each of the simulations 1, 2, 3 and 4, respectively.

	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$	$\eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij}$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$
ϕ	[105.0, 108.0]	[162.5, 168.7]	[762.8, 991.2]	[708.1, 833.0]
τ		[4118, 6190]		[6624, 7610]
κ		[0.000502, 0.000664]		[0.000463, 0.000527]
σ_0		[0.129, 0.257]		[0.133, 0.173]
r		[4258, 5638]		[5370, 6100]
$\frac{1}{\sigma^2}$			[15.12, 15.98]	[22.21, 23.50]

	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$	$\eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij}$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$
ϕ	[106.3, 109.3]	[162.2, 167.9]	[739.2, 1449.6]	[799.8, 934.0]
τ		[4740, 5809]		[5936, 7178]
κ		[0.000545, 0.000647]		[0.000489, 0.000577]
σ_0		[0.141, 0.206]		[0.128, 0.183]
r		[4375, 5190]		[4898, 5787]
$\frac{1}{\sigma^2}$			[14.36, 16.18]	[22.36, 23.78]

	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$	$\eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij}$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$
ϕ	[107.4, 110.5]	[162.5, 168.3]	[671.4, 809.5]	[712.8, 830.7]
τ		[5405, 6160]		[6145, 6951]
κ		[0.000515, 0.000577]		[0.000515, 0.000575]
σ_0		[0.149, 0.191]		[0.133, 0.168]
r		[4898, 5496]		[4916, 5492]
$\frac{1}{\sigma^2}$			[15.72, 16.61]	[23.92, 25.44]

	$\eta_{ij} = \alpha_i + \beta_i k_j$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i)$	$\eta_{ij} = \alpha_i + \beta_i k_j + \epsilon_{ij}$	$\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$
ϕ	[105.0, 108.0]	[163.1, 168.1]	[676.2, 777.6]	[734.5, 864.7]
τ		[4969, 5657]		[5863, 6948]
κ		[0.000554, 0.000622]		[0.000484, 0.000562]
σ_0		[0.151, 0.193]		[0.136, 0.187]
r		[4550, 5105]		[5033, 5841]
$\frac{1}{\sigma^2}$			[15.68, 16.42]	[23.50, 24.90]

8 Discussion/Conclusion

In this paper, we have investigated DNA methylation data, where the main focus has been on a data set containing 62 persons classified as having Schizophrenia, and 33 Healthy ones. Through an exploratory analysis of chromosome 1 to 22, we have seen that the data is likely to be beta distributed, with a location dependent mean and varying precision parameter ϕ. With further analysis of correlations and sample auto-correlations, we have seen that the methylation data seems to be influenced by spatial dependency.

To investigate the spatial dependency in the data, several models and likelihoods have been studied in this paper. The main focus has been on finding differences in mean between groups of people and to try to quantify the spatial dependency structure. As we have seen in Chapter 6, models which accounts for spatial dependency result in a lower DIC value. This indicates a better fit to the data. With the simulation studied in 7.2.1, we have seen that a data set without spatial dependency does not take advantage concerning the DIC value by including spatial dependency effects in the latent Gaussian model. The resulting credibility intervals for the hyperparameters τ and κ are very wide, which also indicates that such a model does not fit the data well. For the real data set, the DIC indicates a better fit for both likelihood functions and all models considered, where the spatial dependency has been taken into account. The credibility regions obtained from the different models are also much shorter than the ones obtained by the simulation without dependency. Although only one simulation of this case is done, the results obtained, especially for the hyperparameters, do not fit the results obtained by studying the Schizophrenia data set. This is a further evidence of spatial dependency in the methylation data.

With the first likelihood function considered, we see that the SPDE model decreases the DIC value, which indicates a better fit. This means that having fitted realizations $\xi_j(s_i)$ of a SPDE build on parameters τ and κ in the mean, improves the log likelihood despite of the increase of effective parameters that need to be estimated. However, by treating the precision parameter ϕ as fixed, the spatial dependency process is not able to absorb any of the variation in the data, resulting in an overestimation of the variance of the other estimated parameters, such as β_i. This results in conditionally fewer positions treated as differently methylated. This also seem to result in an underestimation of the σ_0 hyperparameter and a overestimation of the range r, compared to the more fitting model, which is the linear predictor 4 with likelihood function 2.

The distribution of the test variables in the T-test is a bit skewed compared to the Student t-distribution with degrees of freedom equal to 93. This seems to

be a pattern that repeats for different chromosomes. The amount of significantly different methylated positions, displayed in Table 3.1, are also consistently lower than 0.05 for all the chromosomes, which indicates that the amount is lower than what we expect for the type I error. When considering scales such as chromosome 1($n = 46866$), we would expect that the amount of differently methylated cites would be approximately 0.05, even with spatial dependency along the chromosome. For the simulated data sets studied in 7.2.3, the amount of differently methylated positions varies within natural boundaries for the $n = 400$ positions considered, but for the Schizophrenia data set, the obtained amount is lower. This might point towards some bias in the T-test for the methylation data, or some underlying structure in the data that is not yet quantified.

As seen from the results of the Schizophrenia data and the simulations, there is possible that the data can be described by a latent Gaussian model such as $\eta_{ij} = \alpha_i + \beta_i k_j + \xi_j(s_i) + \epsilon_{ij}$, and a location independent precision parameter ϕ. However, there are some differences between the simulated and the Schizophrenia data set. For the simulations, the amount of differently methylated positions found with both the T-test and the linear predictor 4 varies within natural boundaries, but for the Schizophrenia data set, it is lower for both tests. The fit of linear predictor 3 also gives different results between the simulations and the Schizophrenia data set. This might again point towards some underlying structure in the Schizophrenia data, that the models tested are not able to quantify. A study of other regions would be necessary to obtain more knowledge of this possible underlying structure.

With INLA and the SPDE approach, we are able to model the spatial dependency for subsequent parts of the human genome. Because of the interactions between effects considered in the latent Gaussian model, the process of fitting is complex and expensive. We are therefore not able to look into much larger regions, such as whole chromosomes. However, as seen for the simulations considered, we see that by utilizing the spatial dependency in the data, we obtain better estimates of differently methylated positions than with the T-test. Therefore, to find differently methylated positions, we could use a hierarchical approach. By first investigating the data with an algorithm such as the dmpFinder or a T-test, we could find regions that seem promising of containing differently methylated sites. By then using an approach such as the one considered in this paper, we could investigate the region by taking advantage of the spatial dependency.

As seen from the simulations considered in 7.2.3, the resulting fit seem to be dependent on the prior specifications. Especially the centralization of the distribution seems to affect the result, as seen for the SPDE model. We have also seen that the hyperparameters in the simulation seem to interfere with each other, reducing the identifiability of the parameters in the simulations. To get more knowledge of how prior specifications influences the parameters of the model considered, further

studies are required.

By investigation of the sample auto-correlation for the different people, it would seem that their spatial dependency is different and therefore might be best described by independent τ and κ parameters. If we compare the resulting credibility regions for the hyperparameters in the simulation studied in 7.2.2 and the credibility regions obtained by the SPDE and iid error model with likelihood function 2 found in Table 6.4, we see that the credibility regions for the simulation is narrower, especially for σ_0 and the range r. Although different likelihood functions were considered for the simulated data set and the Schizophrenia data, and the prior specifications for the SPDE model parameters were different, it would be interesting to look into a model with specifications of individual SPDE structures. This can be justified by looking at the credibility interval in the simulations done in 7.2.3 as well, which also are a bit narrower on average then the ones obtained for the Schizophrenia data set. If this would be investigated, we might need to increase the number of sites since the amount of hyperparameters to be estimated would increase drastically, to reduce the effect of the Neumann boundary conditions on the hyperparameters.

The estimations might be affected by the fact that the methylation data is processed(Lowe and Rakyan 2013). This means that in addition to some normalizing procedure, missing values are imputed. This might affect the data, and since the INLA algorithm, and in general the Bayesian framework, is able to deal with missing data(Blangiardo and Cameletti 2015), raw data might be better to use. We could also use the fact that the data are spatially dependent to help with the imputation of missing values. This could be an interesting approach for further research.

For further research, it would be interesting to look into a non stationary approach for the SPDE model. We have seen that the estimated range parameter over the different chromosomes was approximately equal to 700, but the one obtained with the model that gave the best fit, had a credibility interval ranging from $[5019, 6139]$. This indicates that the range parameter might be changing for different regions. For a region of 400 CpG sites, it is possible that a stationary approach is the one that obtains the best fit, but it would be interesting to look into the possibility of a non-stationary approach as well. To optimize with a likelihood function defined as having a precision parameter ϕ for each location, that are to be estimated, could also be interesting. This would result in many hyperparameters to be estimated, such that it is possible that the amount of data(more people) would need to be larger to produce certain estimates.

Since the DNA lies in clusters coiled around each other, there is a possibility that the distance between some locations in 3D are closer than the distance measured in 1D. This means that the methylation process might affect "spaces" in the cells, such that correlations in a three dimensional space around each location should be

measured. The SPDE approach allows for this kind of dependency, such that this might also be an interesting approach for further research.

A Appendix

A.1 Auto-correlation comparison between methylation data and simulated data

As we can see in Figure A.1, the auto-correlation lag 1 seems to vary a lot among the people at the different chromosomes. As an example, the correlation varies from 0.12 to 0.28 at chromosome 1. This auto-correlation is calculated from $n = 46866$ distinct locations, such that the spread is probably caused by being realizations of different GRFs. To comparison, we have displayed the same plot for simulated data based on the same locations from chromosome $2-22$(chromosome 1 was to large to simulated data using inla.qsample()) in Figure A.2. The data are simulated using $\kappa = 0.004$, $\tau = 1877.814$ and $\lambda = 1$. This is equivalent to a range parameter $r = 700$ and $\sigma = 0.07$.

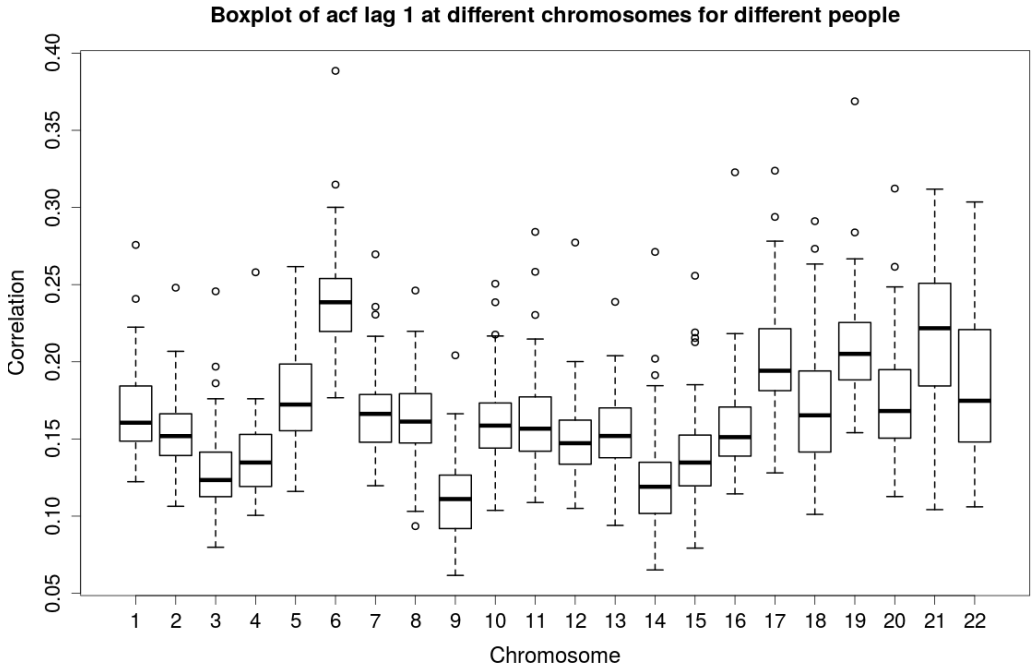


Figure A.1: Boxplot of the auto-correlation function at lag 1 at different chromosomes for different people. Evaluated for the Schizophrenia data set, with $N = 95$ people.

Figure A.1 and A.2 shows that the spread in auto correlation lag 1 for the Schizophre-

nia data set is larger than for the simulated data. This suggests that the differences between the people might be best described by having realizations from different Gaussian processes. The auto-correlation is on an average scale larger for the simulated data set than for the Schizophrenia data set. On chromosome 6, we see that the simulation data and the Schizophrenia data have a larger correlation than for the other chromosomes. This indicates that the CpG sites are denser on this chromosome than the other, such that the relevance of spatial dependencies are stronger at this chromosome. Other values for the simulations have been tested as well, giving the same results concerning the spread of the auto-correlation at each chromosome.

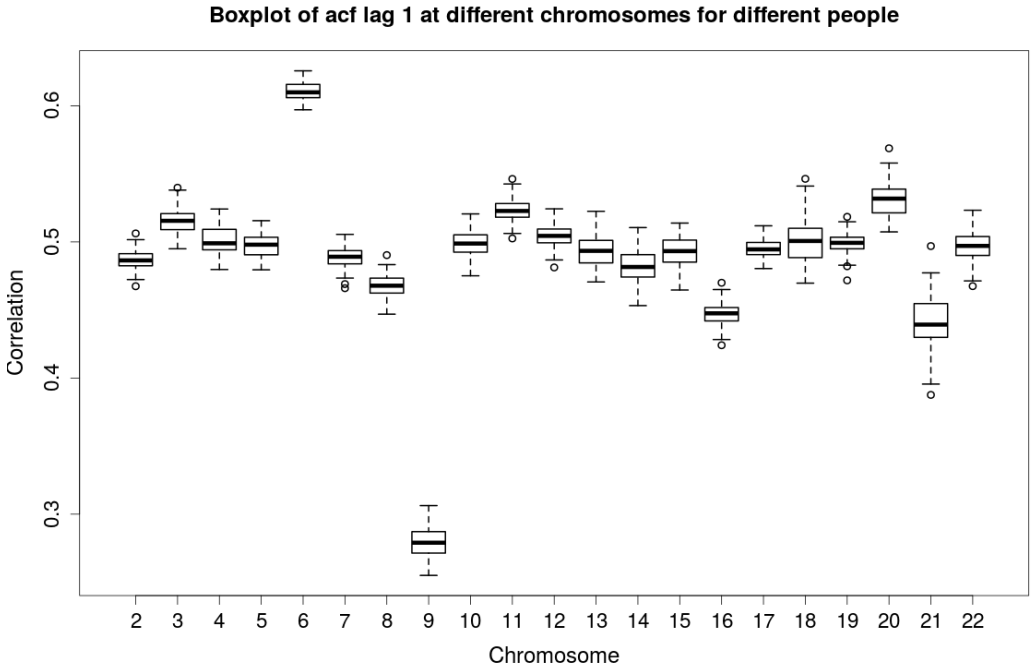


Figure A.2: Boxplot of the auto-correlation function at lag 1 at different chromosomes for different people. Evaluated for a simulated data set, with $N = 95$ people and $\kappa = 0.012$ and $\tau = 1343.212$.

A.2 Natural variation in the Type I error($n = 400$)

When we are considering differently methylated positions in a data set based on 400 locations, we can calculate the natural variation in the type I error by the following procedure.

The probability of obtaining a false positive is equal to 0.05, when considering

a test with significance level $\alpha = 0.05$. To obtain the natural variation, we consider the Bernoulli distribution with probability $p = 0.05$. The standard deviation for this distribution is then equal to $\sqrt{p(1-p)}$, and we can find a 95% prediction interval for the type I error by the following formula:

$$\begin{aligned} &0.05 \pm 1.96\sqrt{\frac{1}{n}0.05(1-0.05)} \\ &= [0.0286, 0.0714]. \end{aligned} \tag{A.1}$$

Bibliography

Aryee, M. J., A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry (2014).
"Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays". In: *Bioinformatics* 30.10, 1363–1369. issn: 1460-2059. doi: 10.1093/bioinformatics/btu049 (cit. on pp. 1, 7).

Assenov, Y., F. Müller, P. Lutsik, J. Walter, T. Lengauer, and C. Bock (2014).
"Comprehensive analysis of DNA methylation data with RnBeads".
In: *Nat Meth* 11.11, 1138–1140. issn: 1548-7105. doi: 10.1038/nmeth.3115 (cit. on pp. 7, 9).

Barlow, D. P. (2011).
"Genomic Imprinting: A Mammalian Epigenetic Discovery Model".
In: *Annual Review of Genetics* 45.1, 379–403. issn: 1545-2948.
doi: 10.1146/annurev-genet-110410-132459 (cit. on p. 6).

Bartlett, M. S. (1946). "On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series".
In: *Supplement to the Journal of the Royal Statistical Society* 8.1, p. 27.
issn: 1466-6162. doi: 10.2307/2983611 (cit. on p. 16).

Baylin, S. B. and P. A. Jones (2011). "A decade of exploring the cancer epigenome — biological and translational implications". In: *Nat Rev Cancer* 11.10, 726–734.
issn: 1474-1768. doi: 10.1038/nrc3130 (cit. on p. 6).

Bibikova, M., B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, and et al. (2011).
"High density DNA methylation array with single CpG site resolution".
In: *Genomics* 98.4, 288–295. issn: 0888-7543.
doi: 10.1016/j.ygeno.2011.07.007 (cit. on pp. 1, 7).

Blangiardo, M. and M. Cameletti (2015).
Spatial and Spatio-temporal Bayesian Models with R-INLA. John Wiley & Sons (cit. on pp. 24, 25, 27, 28, 32, 35, 61).

Cressie, N. A. C. (1993). "Statistics for Spatial Data".
In: *Wiley Series in Probability and Statistics*. issn: 1940-6347.
doi: 10.1002/9781119115151 (cit. on pp. 2, 28).

Feinberg, A. P. (2007).
"Phenotypic plasticity and the epigenetics of human disease".
In: *Nature* 447.7143, 433–440. issn: 1476-4687. doi: 10.1038/nature05919 (cit. on pp. 1, 6).

Feinberg, A. P. and B. Tycko (2004). "The history of cancer epigenetics".
In: *Nat Rev Cancer* 4.2, 143–153. issn: 1474-1768. doi: 10.1038/nrc1279 (cit. on p. 6).

Gardiner-Garden, M. and M. Frommer (1987).
"CpG Islands in vertebrate genomes".
In: *Journal of Molecular Biology* 196.2, 261–282. ISSN: 0022-2836.
DOI: 10.1016/0022-2836(87)90689-9 (cit. on p. 1).
George E. P. Box, G. C. T. (1992). *Bayesian Inference in Statistical Analysis*.
John Wiley and Sons, Inc. (cit. on pp. 2, 11).
Hansen, K. D. and M. J. Aryee (2015).
"The minfi User's Guide Analyzing Illumina 450k Methylation Arrays". In:
(cit. on pp. 1, 20).
Hilbe, J. M. (1994). "Generalized Linear Models".
In: *The American Statistician* 48.3, 255–265. ISSN: 1537-2731.
DOI: 10.1080/00031305.1994.10476073 (cit. on p. 10).
Horvath, S. (2013). "DNA methylation age of human tissues and cell types".
In: *Genome Biology* 14.10, R115. ISSN: 1465-6906.
DOI: 10.1186/gb-2013-14-10-r115 (cit. on p. 10).
Human Genome Sequencing Consortium, I. (2004).
"Finishing the euchromatic sequence of the human genome".
In: *Nature* 431.7011, 931–945. ISSN: 1476-4679. DOI: 10.1038/nature03001
(cit. on p. 2).
Lindgren, F. and H. Rue (2013).
"Bayesian Spatial and Spatio-temporal ModeModel with R-INLA". In:
(cit. on p. 31).
Lindgren, F., H. Rue, and J. Lindström (2011).
"An explicit link between Gaussian fields and Gaussian Markov random fields:
the stochastic partial differential equation approach". In: *Journal of the Royal
Statistical Society: Series B (Statistical Methodology)* 73.4, 423–498.
ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2011.00777.x
(cit. on pp. 2, 29–31).
Lowe, R. and V. K. Rakyan (2013).
"Marmal-aid – a database for Infinium HumanMethylation450".
In: *BMC Bioinformatics* 14.1, p. 359. ISSN: 1471-2105.
DOI: 10.1186/1471-2105-14-359 (cit. on pp. 8, 61).
MENG, X.-L. and D. B. RUBIN (1992).
"Performing likelihood ratio tests with multiply-imputed data sets".
In: *Biometrika* 79.1, 103–111. ISSN: 1464-3510. DOI: 10.1093/biomet/79.1.103
(cit. on p. 32).
Martin, M. and Z. Herceg (2012).
"From hepatitis to hepatocellular carcinoma: a proposed model for cross-talk
between inflammation and epigenetic mechanisms". In: *Genome Med* 4.1, p. 8.
ISSN: 1756-994X. DOI: 10.1186/gm307 (cit. on p. 6).
McDermott, E., E. J. Ryan, M. Tosetto, D. Gibson, J. Burrage, D. Keegan,
K. Byrne, E. Crowe, G. Sexton, K. Malone, and et al. (2015).
"DNA Methylation Profiling in Inflammatory Bowel Disease Provides New

Insights into Disease Pathogenesis". In: *Journal of Crohn's and Colitis* 10.1, 77–86. ISSN: 1876-4479. DOI: 10.1093/ecco-jcc/jjv176 (cit. on p. 1).

Meissner, A. (2010). "Epigenetic modifications in pluripotent and differentiated cells". In: *Nature Biotechnology* 28.10, 1079–1088. ISSN: 1546-1696. DOI: 10.1038/nbt.1684 (cit. on p. 6).

Mirella Gonzalez-Zulueta Christina M. Bender, A. S. Y. and T. Nguyen (1995). "Methylation of the 5' CpG Island of the p16/CDKN2 Tumor Suppressor Gene in Normal and Transformed Human Tissues Correlates with Gene Silencing". In: *University of Southern California, School of Medicine, Las Angeles, California* (cit. on pp. 1, 6, 9).

Rue, H. and L. Held (2005a). *Gaussian Markov Random Fields: Theory and Applications*. Vol. 104. Monographs on Statistics and Applied Probability. London: Chapman & Hall (cit. on pp. 2, 25).

Rue, H. and L. Held (2005b). *Gaussian Markov Random Fields*. Chapman & Hall (cit. on pp. 30, 31).

Rue, H., S. Martino, and N. Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, 319–392. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2008.00700.x (cit. on pp. 2, 11, 24, 25, 27).

Rukova, B., R. Staneva, S. Hadjidekova, G. Stamenov, V. Milanova, and D. Toncheva (2014). "Genome-Wide Methylation Profiling of Schizophrenia". In: *Balkan Journal of Medical Genetics* 17.2. ISSN: 1311-0160. DOI: 10.2478/bjmg-2014-0070 (cit. on p. 6).

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). "Bayesian measures of model complexity and fit". In: *J Royal Statistical Soc B* 64.4, 583–639. ISSN: 1467-9868. DOI: 10.1111/1467-9868.00353 (cit. on p. 32).

— (2014). "The deviance information criterion: 12 years on". In: *J. R. Stat. Soc. B* 76.3, 485–493. ISSN: 1369-7412. DOI: 10.1111/rssb.12062 (cit. on p. 40).

Takai, D. and P. A. Jones (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22". In: *Proceedings of the National Academy of Sciences* 99.6, 3740–3745. ISSN: 1091-6490. DOI: 10.1073/pnas.052410099 (cit. on pp. 8, 9).

Wei, W. W. S. (2006). *Time series analysis; Univariate and Multivariate Methods*. Pearson Education (cit. on p. 16).