Bayesian Hidden Markov Models for finding DNA Copy Number Changes from SNP Genotyping Arrays

by

Matthew Kowgier

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Dalla Lana School of Public Health
University of Toronto

Copyright © 2012 by Matthew Kowgier
DNA copy number variations (CNVs), which involve the deletion or duplication of subchromosomal segments of the genome, have become a focus of genetics research. This dissertation develops Bayesian HMMs for finding CNVs from single nucleotide polymorphism (SNP) arrays.

A Bayesian framework to reconstruct the DNA copy number sequence from the observed sequence of SNP array measurements is proposed. A Markov chain Monte Carlo (MCMC) algorithm, with a forward-backward stochastic algorithm for sampling DNA copy number sequences, is developed for estimating model parameters. Numerous versions of Bayesian HMMs are explored, including a discrete-time model and different models for the instantaneous transition rates of change among copy number states of a continuous-time HMM. The most general model proposed makes no restrictions and assumes the rate of transition depends on the current state, whereas the nested model fixes some of these rates by assuming that the rate of transition is independent of the current state. Each model is assessed using a subset of the HapMap data. More general parameterizations of the transition intensity matrix of the continuous-time Markov process produced more accurate inference with respect to the length of CNV regions. The observed SNP array measurements are assumed to be stochastic with distribution determined by the underlying DNA copy number. Copy-number-specific distributions,
including a non-symmetric distribution for the 0-copy state (homozygous deletions) and mixture distributions for 2-copy state (normal), are developed and shown to be more appropriate than existing implementations which lead to biologically implausible results. Compared to existing HMMs for SNP array data, this approach is more flexible in that model parameters are estimated from the data rather than set to \textit{a priori} values. Measures of uncertainty, computed as simulation-based probabilities, can be determined for putative CNVs detected by the HMM. Finally, the dissertation concludes with a discussion of future work, with special attention given to model extensions for multiple sample analysis and family trio data.
Acknowledgements

First, I’d like to acknowledge my advisor, Rafal Kustra, for introducing me to the fields of data mining and genomics, introducing me to Emacs, and for providing guidance, computing resources and a stimulating environment for biostatistical research. His enthusiasm and passion for science, especially biostatistics, is truly inspiring. Despite delays and setbacks, he has supported me throughout my candidature. For this, I am grateful.

Thank you to Michael Escobar, a thesis committee member and mentor, for teaching me Bayesian methods and opening my mind to the Bayesian view of probabilities. It was a pleasure to work with you.

A debt of gratitude is also owed to Radford Neal, the other thesis committee member, who taught me how to think objectively and has given me the confidence to explore my own ideas. In addition to his valuable suggestions, Radford’s meticulous approach to research has been a highly positive influence on me.

I would also like to acknowledge all others involved in this research, including Laurent Briollais, Steven Gallinger and Celia Greenwood.

Special thanks to my CADQOL collaborators, George Tomlinson and Gary Naglie. Thank you both for providing me with so many opportunities to participate in clinical research, as well as giving me invaluable advice and perspective.

Thank you to the members of the Statistical Genomics Laboratory. Without the day-to-day companionship and support from them, this experience would not have been as enjoyable. They include Zhengfei Chen, Billy Chang, Nicholas Mitsakakis, Adam Zagdanski, Xiaofei Shi, Niloofar Arshadi and Konstantin Shestopaloff.

Some of my fondest memories during my doctoral studies came at Varsity Arena where I played for the SGS Cougars, 2010-11 intramural champions. Thank you to all my teammates for the fun times, especially Brandon McFarlane and Alexei Halpin.

I also gratefully acknowledge the scholarship from the Natural Science and Engineering Research Council of Canada.
Life has thrown me many curve balls over the years but throughout them all my constant has been the support and love of my parents, Maria and Edward Kowgier. Thank you for giving me this life, standing by me and always believing in me, even when I gave you no reason to. Also a big thank you to my brother, Mark, and my sister-in-law, Jen. I have learned much from you both and thank you for your support.

A man’s character is his fate. (Heraclitus)
Contents

1 Introduction
   1.1 Biological background ........................................... 1
      1.1.1 Genetic information ......................................... 1
      1.1.2 Single nucleotide polymorphisms ......................... 2
      1.1.3 Copy number variations .................................... 3
      1.1.4 Array comparative genomic hybridization .................. 6
      1.1.5 SNP genotyping arrays ...................................... 6
   1.2 Copy number inference ........................................... 8
      1.2.1 Copy number analysis with SNP genotyping arrays ........ 9
      1.2.2 Preprocessing ............................................... 10
      1.2.3 Other preprocessing algorithms ............................ 14
   1.3 Thesis aims ..................................................... 16

2 Review of Hidden Markov Models for Copy Number Analysis 18
   2.1 Introduction ..................................................... 18
   2.2 Hidden Markov models ........................................... 19
      2.2.1 Background ................................................. 19
      2.2.2 Fitting HMMs ............................................... 22
      2.2.3 The forward-backward algorithm ........................... 23
      2.2.4 The Viterbi algorithm ...................................... 27
2.3 Previous research .................................................. 28
2.4 Outline of the rest of the thesis ................................. 31

3 A Bayesian Discrete-Time HMM with Normal Emissions 33
3.1 Introduction ......................................................... 33
3.2 Two real CNV data sources ..................................... 35
  3.2.1 Observed data from SNP genotyping arrays .......... 35
  3.2.2 Titration data .................................................... 36
  3.2.3 HapMap data ..................................................... 37
  3.2.4 Analysis of samples with known DNA copy number of the X chromosome ................................ 38
3.3 Bayesian Discrete-time HMMs ................................. 39
  3.3.1 Copy number model .............................................. 40
  3.3.2 Prior distributions .............................................. 42
  3.3.3 Model behaviour ................................................. 44
3.4 Parameter estimation ............................................. 47
  3.4.1 A Markov chain Monte Carlo algorithm .................. 47
  3.4.2 Sampling copy number sequences ......................... 48
  3.4.3 Sampling the model parameters ......................... 50
3.5 Bayesian copy number sequence estimation ................. 52
  3.5.1 Marginal distribution of $C_j$ ............................... 52
  3.5.2 Marginal distribution of $C$ ................................. 53
3.6 Choosing the number of states ............................... 55
3.7 Implementation of the algorithm .............................. 56
3.8 Analysis of HapMap data ...................................... 56
3.9 Discussion of Results .......................................... 59
4 A Continuous-time HMM

4.1 Introduction ................................................. 63
4.2 Continuous-time HMMs .......................... 64
  4.2.1 Copy number model ......................... 65
  4.2.2 Computing the transition probabilities .... 67
  4.2.3 Parsimonious submodel .................. 69
  4.2.4 Heterogeneous model for CNPs ......... 72
4.3 Bayesian continuous-time model .......... 72
  4.3.1 Priors ............................................. 72
  4.3.2 Sampling the model ............................ 75
4.4 Data analysis ........................................... 77
  4.4.1 Model 1 (BCHMM1) ....................... 77
  4.4.2 Model 2 (BCHMM2) ....................... 78
4.5 Chapter Summary and Discussion of Results .... 81

5 A Continuous-Time HMM with Copy-Number-Specific Emissions 83

5.1 Introduction .............................................. 83
5.2 Novel HMM ................................................. 83
  5.2.1 Emission probabilities, 2-copy state .... 85
  5.2.2 Emission probabilities, 0-copy state .... 85
  5.2.3 Priors ............................................. 87
  5.2.4 Sampling .......................................... 89
5.3 Data Analysis ............................................ 89
  5.3.1 Model 1 (BCHMMDmix) ................... 90
  5.3.2 Model 2 (BCHMMDmix2) ................. 93
  5.3.3 Model 3 (BCHMMDmix3) ................. 95
5.4 Chapter Summary and Discussion of Results .... 98
6 Bayesian Inference for Subsequences 100

6.1 Introduction ......................................................... 100
6.2 Inference on subsequences ........................................... 100
  6.2.1 Estimating posterior probabilities of CNVs ................. 101
  6.2.2 Approach to visualizing the MCMC results .............. 103
  6.2.3 Data analysis of a subsequence .............................. 104
  6.2.4 Further data analysis of identified CNVs ............... 106
6.3 Bayes factor for assessing copy number segments .......... 107

7 Discussion and Future Research 111

7.1 Summary ............................................................. 111
7.2 Future Work .......................................................... 116
  7.2.1 Multiple sample analysis ................................. 116
  7.2.2 Family trio data .............................................. 119

Appendix A 123
Appendix B 125
Appendix C 126
Bibliography 130
List of Tables

1.1 DNA sequences of 10 bases at the same locus with point mutations occurring at the sixth base. ................................................................. 2

3.1 Estimated mean and standard deviation for various levels of DNA copy number of the X chromosome. ......................................................... 41

3.2 Posterior summary of the parameters of the transition probability matrix for BDHMM. ................................................................. 61

3.3 Posterior summary of the parameters of the emission distribution for BDHMM. 62

4.1 Posterior summary of the parameters of BCHMM1, first two columns, and BCHMM2, last two columns. ................................................................. 82

5.1 Estimated copy number, per SNP, for BCHMMDmix2 (row) and BCHMMDmix3 (column). ................................................................. 97

5.2 Posterior summary of the parameters of BCHMMDmix2, first two columns, and BCHMMDmix3, last two columns. ................................................................. 99

6.1 Summary of the 16 detected CNVs for the analysis of sample NA18550 given in Chapter 5. Length is the length of the CNV in bps, defined as difference between the end position and start position. Percent overlap refers to the number of SNPs in the CNV that overlap with those reported in McCarroll et al. (2008). ................................................................. 110
List of Figures

1.1 Density estimates of the log of the probe-level data for 20 arrays. Each line corresponds to an array. Left plot, before calibration for allelic crosstalk. Right plot, after calibration for allelic crosstalk. ........................................ 11

1.2 Observed raw CNs as a function of the genomic position in mega base pairs (Mbs) for sample NA18550 on chromosome 10. ................................. 15

1.3 Average ROC curves for CRMA and PLASQ. The ROC curves are averages across 5605 SNPs on the X chromosome. ................................. 16

2.1 Graphical representation of the dependence structure of a HMM. .......... 20

3.1 Histogram of the lengths of the 1,320 CNPs reported in McCarroll et al. (2008). 32 CNPs with lengths greater than 200,000 bps were truncated from this histogram. ................................. 38

3.2 Raw CNs averaged across replicates for X chromosome titration data (3X, 4X, 5X), and raw CNs for one male (1X) and one female (2X) HapMap sample. The separate boxplots correspond to DNA samples with various numbers of copies of the X chromosome. The dotted lines correspond the theoretical log-ratio values: log(1/2), log(2/2), ..., log(5/2). ................................. 40

3.3 Observed raw CNs as a function of the genomic position in mega base pairs (Mbs) for sample NA18550 on the second arm of chromosome 10. ................................. 58
3.4 Subplot of 2000 SNPs. The top displays shows the observed raw CNs for
the first 2000 SNPs on the chromosome and the bottom display shows the
estimated copy number of these SNPs based on the BDHMM. . . . . . . 60

4.1 Graphical representation of the continuous-time HMM. . . . . . . . . . . 65

4.2 Whole chromosome display. The top displays shows the observed raw CNs
for sample NA18550 on the second arm of chromosome 10, the middle
and bottom displays show the estimated copy number of these SNPs for
BCHMM1 and BCHMM2, respectively. . . . . . . . . . . . . . . . . . . . . . . . 78

4.3 Subregion of 2000 SNPs. The top displays shows the observed raw CNs
for the first 2000 SNPs on the chromosome, the middle display shows
the estimated copy number of these SNPs for BDHMM, and the bottom
display shows the estimated copy number of these SNPs for BCHMM1. . 79

4.4 Trace plots of the $\nu$-parameters for BCHMM2. . . . . . . . . . . . . . . . . . 80

5.1 Estimated density using kernel density estimation for data in the 1- and
2-copy states grouped according to copy number values for a randomly
selected MCMC iteration for BCHMM1. . . . . . . . . . . . . . . . . . . . . . 84

5.2 Histogram of raw CNs for 29 samples sharing the same homozygous dele-
tion (0-copy state). The CNV contains 27 SNPs. Superimposed Normal
distribution had a mean of -1.9 and a standard deviation of 1. Superim-
posed extreme value distribution had a $\beta$ parameter of 5. . . . . . . . . . . 86

5.3 Whole chromosome display. The top displays shows the observed raw CNs
for sample NA18550 on the second arm of chromosome 10 and the bottom
display shows the estimated copy number of these SNPs for BCHMMMDmix. 90
5.4 Subregion of 2000 SNPs. The top display shows the observed raw CNs for the first 2000 SNPs on the chromosome, the middle display shows the estimated copy number of these SNPs for BCHMM, and the bottom display shows the estimated copy number of these SNPs for BCHMMDmix.

5.5 Subregion of 200 SNPs containing two SNPs that appear to have been misclassified by BCHMMDmix. Green points correspond to 3-copy state predictions, whereas red points correspond to 0-copy state predictions.

5.6 Posterior distributions for the expected interval length in base pairs. Prior densities are superimposed in blue. The prior density for the 3-copy state (lower right panel) was rescaled to fit inside the plotting region.

5.7 Posterior distributions for $\beta$. The prior density is superimposed in blue.

6.1 Subregion of 16 SNPs containing CNV 5. Left plot, the sampled segments times their MCMC frequency. Right plot, marginal probabilities at each SNP in the subregion.
Chapter 1

Introduction

In this chapter we provide an introduction to copy number analysis using SNP genotyping arrays. In the first section, some necessary biological background is provided. In the second section, we summarize the goals of copy number inference, review the use of SNP genotyping arrays for copy number inference, and discuss the preprocessing of probe-level data from SNP genotyping arrays.

1.1 Biological background

In this section we provide a basic summary of the underlying molecular genetics and biology relevant to the data studied in this thesis: SNP genotyping array data. For a more comprehensive overview, see Griffiths et al. (2004) and Strachan & Read (2004).

1.1.1 Genetic information

The genetic information of humans is located in chromosomes which are found, as two copies, in the nucleus of every cell (with the exception of mature red blood cells). The state of having two complete sets of chromosomes in a cell is referred to as diploid, each chromosome is one long molecule of deoxyribonucleic acid (DNA). The monomers
of DNA are nucleotides, each nucleotide consists of a phosphate group, a ribose sugar molecule, and one of four nitrogenous bases: adenine (A), guanine (G), thymine (T), or cytosine (C). Two complementary strands make up a DNA molecule and are connected by multiple hydrogen bonds between pairs of bases on opposing strands. In this set-up, thymine bonds with adenine, and cytosine with guanine. The genetic information is encoded in the sequence of bases making up each DNA fragment. The specific location of a sequence of DNA on a chromosome is called the locus (or loci for plural). Differences in genetic information between individuals is characterized by differences in their DNA. The DNA loci that harbor such differences are said to be polymorphic, and the variations that occur at a polymorphic locus are called alleles. For example, Table 1.1 shows four possible alleles for a DNA sequence of length 10 base pairs (written as 10 bp). Here, the alleles are characterized by point mutations at the sixth nucleotide of the sequence. In the next section we discuss an important type of point mutation: single nucleotide polymorphisms.

Table 1.1: DNA sequences of 10 bases at the same locus with point mutations occurring at the sixth base.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTGCAACCAT</td>
<td>A</td>
</tr>
<tr>
<td>ATTGCCCCCAT</td>
<td>C</td>
</tr>
<tr>
<td>ATTGCGCCAT</td>
<td>G</td>
</tr>
<tr>
<td>ATTGCTCCAT</td>
<td>T</td>
</tr>
</tbody>
</table>

1.1.2 Single nucleotide polymorphisms

The most common form of genetic variation are single nucleotide polymorphisms (SNPs, pronounced snips) which are point mutations that occur at a single base (i.e., A, T, C, or G) in the genome and at a frequency greater than 1% in a given population. SNPs
are biallelic, meaning that they have only two variants, or alleles, in the population. (The other two possible bases are either absent or have a very low frequency.) The more prevalent allele is called the major allele (or A allele) and the less prevalent allele is called the minor allele (or B allele). Therefore, each SNP is associated with three possible genotypes (e.g. AA, BB, or AB for an A/B SNP). For example, in Table 1.1, if we treat the A allele as the most prevalent DNA sequence, the C allele differs from the A allele by a single base location at which a SNP is said to have occurred.

Central to understanding genetic variation within and among populations is the problem of inferring haplotypes from genotypes of SNPs. Haplotypes are a set of linked alleles on the same chromosome which tend to be inherited together. Haplotypes arise due to the phenomenon of genetic linkage by which alleles that are close together on the same chromosome tend to be inherited together. Haplotypes come in pairs, one from your mother and one from your father. The pair of haplotypes represents a person’s genetic profile, sometimes referred to as the diplotype. Whereas genotypes are read across a pair of chromosomes, haplotypes are read down a pair of chromosomes. A person’s genetic profile is important because it can, for example, determine the person’s likelihood of developing a certain disease, or it can determine the person’s response to a drug. Researchers are interested in determining which haplotypes are present in the genome (only some of the possible haplotypes actually occur) and whether or not these haplotypes are associated with diseases of interest. This is the goal of the International HapMap Project which seeks to provide a catalogue of human genetic variation and has already discovered over one million SNPs (The International HapMap Consortium 2005). We will discuss and summarize SNP genotyping array data from the HapMap Project in Section 3.2.

1.1.3 Copy number variations

Recently researchers have begun to focus on genome structural variation in addition to what was previously believed to be the most common form of genetic variation, SNPs.
Genome structural variation includes insertions, deletions, translocations, and inversions of genomic material. These genomic rearrangements result in what are known as copy number variations (CNVs). CNVs are defined as subsequences, or subchromosomal regions, of the genome that are variable in DNA copy number, that is, they have more or less than the usual two copies of DNA. Aneuploidies were traditionally used to refer to whole-chromosome deletions (monosomy) and duplications (e.g. trisomy). Thus, CNVs are extensions of this definition for subchromosomal regions that are variable in copy number, and have only recently moved into the forefront of genetics research due to the advent of DNA microarrays to scan the genome. CNVs include amplifications (more than two copies of DNA), hemizygous deletions (one copy of DNA), and homozygous deletions (no DNA). With respect to size, CNVs were traditionally defined as a segments of DNA that are 1 kb or larger, whereas smaller variations (< 1 kb) are referred to as insertions or deletions (indels) (Feuk et al. 2006). However, with the advent of high resolution microarrays, this distinction is no longer important for the purposes of developing methods for inferring DNA copy number from high resolution microarrays. Microarrays and how they are used to estimate DNA copy number will be discussed in the next section.

DNA copy number changes may be of various source and origin. CNVs are often distinguished from copy number aberrations (CNAs). CNAs are non-inherited genomic changes occurring in somatic cells, whereas CNVs are genomic changes occurring in germline cells. CNVs may be either inherited (that is, the DNA changes are passed from parent to offspring) or caused by de novo mutation (that is, a mutation that neither parent possessed nor transmitted). The latter mutations may occur during either meiosis or mitosis. For example, during meiosis various processes involved in the rearrangement of chromosomes that leads to genetic diversity, such as genetic recombination, can cause segments of chromosomes to be deleted or duplicated. Similar events can occur during mitosis. The mutational mechanisms involved are not well understood (Scavetta & Tautz 2010), but several mechanisms have been proposed, including unequal recombination mechanisms,
homology-directed and non-homologous repair of double-strand breaks, and errors in replication. These mechanisms are covered in detail in Gu et al. (2008). It is also of interest to understand the processes that control the size of CNV regions – this is also an active area of research.

Inherited CNVs with frequencies larger than 1% in a given population are referred to as copy number polymorphisms (CNPs). In this thesis, we are concerned with developing methods for finding germline CNVs, either inherited or de novo, from normal tissue. Methods for finding CNAs from tumorous tissue would require modifications to the methods proposed in this thesis. A map of copy number variation of the human genome with the exact boundaries of rare variants and inherited CNVs is currently a work-in-progress, but so far significant contributions have been made by Redon et al. (2006), Zogopoulos et al. (2007), and McCarroll et al. (2008). The development of this map is critical for facilitating association studies of human disease involving DNA copy number regions, as well as learning about the distribution of the length of CNVs. As mentioned, current maps are incomplete, and the reported range and size of CNVs are inaccurate, reflecting the resolution of the platform used to detect the CNVs and the power and accuracy of the prediction algorithm used.

CNVs are of interest because they often encompass genes (Freeman et al. 2006), many of which are dosage sensitive. This means that CNVs may act to alter the transcriptional levels of these genes, and copy number variation of dosage sensitive genes can lead to a wide range of clinical phenotypes. Therefore, CNVs may be more important than SNPs in terms of disease susceptibility as well as drug response. Indeed, such DNA alterations constitute an important class of genetic mutations which are important genetic underpinnings of many diseases and other phenotypic information (McCarroll & Altshuler 2007, Feuk et al. 2006). For example, inherited CNVs have been shown to be involved in genetic susceptibility to common diseases such as glomerulonephritis (Aitman et al. 2006), HIV (Gonzalez et al. 2005) and Alzheimer’s disease (Rovelet-Lecrux et al. 2006).
1.1.4 Array comparative genomic hybridization

Traditionally, a cytogenetic method called comparative genomic hybridization was used to detect chromosomal aberrations along the genome in cancerous tissue. This method involves mixing labelled DNA from normal tissue and tumour tissue, then hybridizing the mix to normal metaphase chromosomes. Fluorescence intensity is then analyzed to infer the relative amount of DNA in the tumour sample. However, this method suffered from a severe drawback: the resolution was very limited, making genome-wide analysis impractical. To this end, recently researchers have started to hybridize the experimental DNA to microarrays which has greatly increased the resolution, giving rise to array comparative genomic hybridization (CGH). For a more detailed description of array CGH data, see Snijders et al. (2001).

One of the first documented uses of statistical models for estimating DNA copy number from array CGH data was by Fridlyand et al. (2004,). They developed a segmentation technique based on a discrete-time hidden Markov model (DHMM).

The next section discusses the array platform studied in this thesis: Single nucleotide polymorphism (SNP) genotyping arrays.

1.1.5 SNP genotyping arrays

SNP genotyping arrays (often referred to as SNP arrays) are a type of microarray that can be used to find chromosomal aberrations along the genome. They were originally developed for genome-wide genotyping, however they are also capable of providing information on structural variation such as chromosomal copy number. Indeed, in addition to containing polymorphic probes the latest generation of SNP arrays incorporates non-polymorphic probes that were specifically designed to interrogate regions that may contain DNA copy number variants (e.g., the Affymetrix SNP Array 6.0). These developments in array technology reflect recent research indicating that copy number variations (CNVs) are
more common than was previously thought (Redon et al. 2006, McCarroll et al. 2008). The basic setup of the SNP array is the same as the DNA microarray: The target DNA is digested using restriction enzymes that cleave the DNA at known locations, the resulting fragments of DNA are then amplified using polymerase chain reaction (PCR), and the amplified target DNA is labeled and hybridized to a microarray. For a more detailed summary see the seminal paper by Kennedy et al. (2003). Each SNP is represented on the array by a set of probes, which is called a probe set. Each probe in a probe set is a synthetic DNA sequence of 25 bases, called a 25-mer oligonucleotide. The oligonucleotides are differentiated by four design characteristics: (1) the binding SNP allele (A or B); (2) whether or not they perfectly match the target DNA, perfect match (PM) or mismatch (MM) probes; (3) the SNP location with respect to the probe (centered or offset); and (4) sense or antisense strand.

So, how do we use SNP arrays to measure the DNA copy number? Since the probes in a given probe set/SNP are designed to interrogate the target DNA surrounding that SNP, if the target DNA were to lie in a region of duplication (deletion), when we hybridize the target DNA to the microarray the probes that interrogate this region will show relatively stronger (weaker) hybridization affinities, resulting in relatively higher (lower) signals. Thus, sequences of probe sets showing relatively higher (lower) signals can be seen to correspond to copy number duplications (deletions).

One of the most widely used line of arrays are the Affymetrix GeneChip Human Mapping Arrays. The resolution for this line is a pair of chips covering approximately 500,000 (500K) SNPs. These chips are the Nsp and Sty chips, named after the restriction enzymes that are used to digest the DNA onto the chips. Recently, Affymetrix released a new array called the Affymetrix SNP Array 6.0 which contains over 1.8 million genomic markers (906,000 SNPs and 946,000 non-polymorphic markers). The non-polymorphic markers are designed to interrogate non-polymorphic regions that may harbor CNVs. Although they are not SNPs, in our statistical models we do not distinguish between
the polymorphic and non-polymorphic markers, hence for simplicity we may refer to all markers on the arrays as SNPs. These are truly high density arrays and we focus on applications using this platform.

Despite similarity of technology with array CGH, the use of and the development of new methodology for SNP arrays is needed for a few reasons. First, SNP arrays have a higher density than array CGH since SNP arrays use many more probes. Second, the endpoints from array CGH data are ratios of the test DNA sample to that of a reference DNA sample. This makes it an ideal platform for finding CNAs in tumour tissue. However, they are not as useful for finding CNVs in normal tissue since it is difficult to determine whether a detected CNV occurred in the test or reference sample. Thus, SNP arrays are more suitable when we are interested in, as we are in this thesis, finding germline CNVs – CNVs in normal cell lines, rather than tumour cell lines. Third, the probe design on the arrays is different. For example, SNP arrays use short probes that produce allele-specific measurements, whereas array CGH probes are longer, non-allele-specific, and tend to overlap with adjacent probes. These technical differences require different model specifications and, therefore, model fitting procedures. In addition, SNP arrays are already used for genome-wide genotyping. Therefore, if good methods for copy number estimation are developed, they have the potential to yield more information than array CGH.

1.2 Copy number inference

In this section we review the main goals for copy number analysis, specifically when using the SNP genotyping array platform. We also explain how the probe-level data are normalized and summarized to produce single locus copy number estimates, which are used as the endpoint for analysis in the rest of the thesis.
1.2.1 Copy number analysis with SNP genotyping arrays

SNP genotyping arrays were originally developed for genome-wide genotyping, however the technology is also capable of producing predictions of DNA copy number. The objectives for analyzing SNP arrays can be summarized into three main categories. The first, single locus estimation, includes both single locus copy number estimation and genotype estimation. This step typically involves applying preprocessing methods to the data to normalize the probe-level data. As we have learned from DNA microarray data, the preprocessing step is crucial for downstream analysis. It involves removing various biases that exist in the data, including PCR fragment length, sequence effects such as GC content, and across array variation. Then, for each probe set/SNP, the normalized probe-level data are summarized using some approach, one of which we describe for copy number data in the next section. These summarized signals are referred to as the single locus estimates because they do not make use of information across loci. The second step uses the single locus estimates from the previous step to infer, for example, copy number gain or loss along the genome of a single sample. We may also be interested in finding regions with a loss-of-heterozygosity (LOH). (Although LOH is not a topic of this thesis, it is an important genomic event that often results in tumorigenesis. More details on LOH and how it can be detected from SNP arrays can be found in Beroukhim et al. (2006).) An example includes the application of a hidden Markov model (HMM) to the sequence of single locus copy numbers to infer the underlying DNA copy number for a given array. The third step, tries to draw inference across multiple samples. For example, we may be interested in finding common regions of CNVs, called copy number variables regions (CNVRs), or we may be interested in assessing differences in CNV frequencies between different subgroups (e.g. cases versus controls). In this thesis, we focus on issues surrounding the second and third steps. We use existing methods for the first step which we describe in the next section.
1.2.2 Preprocessing

The raw data that we get from SNP arrays are scanned images. These images are converted to continuous intensity values for each probe, where a probe corresponds to a spot on the image. For each SNP $i$, sample or chip $j$, and probe $k$, we denote each pair of probe-level measurements by $x_{ijk} = (x_{ijkA}, x_{ijkB})$, where $i = 1, \ldots, M$, $j = 1, \ldots, N$, $k = 1, \ldots, K_i$, $K_i$ is the number of probes for SNP $i$, and $x_{ijkA}$ represents the continuous intensity for allele $A$ and $x_{ijkB}$ represents the continuous intensity for allele $B$. There are several steps involved in preparing the probe-level data for copy number analysis, basically involving removing various known biological and technical biases that exist. Much work has already been done on this, so to preprocess the probe-level data we use an existing procedure called copy number estimation using robust multichip analysis (CRMA) (Bengtsson et al. 2008). We use this procedure in subsequent chapters to preprocess and then summarize the normalized data. Next, we give an overview of this procedure. More details can be found in the original article by Bengtsson et al. (2008).

Calibration for allelic crosstalk

The DNA sequence of the paired probes, $x_{ijkA}$ and $x_{ijkB}$, differ only by one nucleotide. (Remember that a probe corresponds to a 25-mer oligonucleotide.) As a result, there is often a correlation between the two allele signals, which is referred to as allelic crosstalk. To correct for crosstalk we use a sample-specific affine transformation. This correction typically removes most of the differences in the distributions of probe intensities between samples, as was shown in Bengtsson et al. (2008). Indeed, this is seen in Figure 1.1, which shows the density estimates of the probe-level data for 20 Affymetrix SNP arrays corresponding to 20 HapMap samples both before (left plot) and after (right plot) allelic crosstalk calibration. We see that the calibration was successful in removing much of the inter-array differences, so further normalization is not necessary.
However, in other cases inter-array differences may still persist even after correction for allelic crosstalk, perhaps due to different scanning settings, or different amounts of dyes. In such cases, we can further normalize the data using quantile normalization.

Figure 1.1: Density estimates of the log of the probe-level data for 20 arrays. Each line corresponds to an array. Left plot, before calibration for allelic crosstalk. Right plot, after calibration for allelic crosstalk.

Quantile normalization

Various methods have been proposed for normalization of noisy genomic data. We use the quantile normalization method of Bolstad et al. (2004). Quantile normalization normalizes the data by forcing the quantiles of each sample distribution to match a given target distribution. Specifically, if the target distribution is $F$ and the empirical cumulative distribution function (CDF) of a sample $j$, denoted by $x_j = \{x_{ijk}\}_{1 \leq i \leq M, 1 \leq k \leq K_j}$, is $G$, then the data are transformed using $\tilde{x}_j = F^{-1}(G(x_j))$. We note that quantile normalization may not always be appropriate, for example, when some of the samples under
study are expected to harbor a large number of aberrations. Also, if calibration for allelic crosstalk was successful, as in the previous example, quantile normalization may not be necessary.

In what follows, we denote the normalized probe signals as $\tilde{x}_{ijk}$.

**Probe-level summarization using robust multichip analysis**

After the normalization steps described above, we need to summarize the set of probe signals for each SNP $i$. Since we are interested in total DNA copy number, we use the non-polymorphic probe signals which are computed as a sum of the allele signals,

$$
\tilde{x}_{ijk} = \tilde{x}_{ijkA} + \tilde{x}_{ijkB}
$$

for all $i$, $j$, and $k$. Next, based on the work of Bengtsson et al. (2008), to summarize the probe information at each SNP we use a multi-chip linear model:

$$
\log_2(\tilde{x}_{ijk} + e) = \log_2 \psi_{ij} + \log_2 \zeta_{ik} + \epsilon_{ijk}.
$$

(1.1)

where $\{\psi_{ij}\}_j$ is the chip effect for sample $j = 1, \ldots, N$ at SNP $i$, $\{\zeta_{ik}\}_k$ is the probe effect for probe $k = 1, \ldots, K_i$ at SNP $i$, $\epsilon_{ijk}$ are the error terms with a mean of zero and a SNP-specific variance, and $e$ is a global constant (discussed below). For identifiability the probe effects are constrained such that $\prod_k \zeta_{ik} = 1$. To account for the noisy probe-level data, the model is fit using robust $M$ estimators via an iterative reweighted least squares algorithm with the **affyPLM** R package (Bolstad 2004). This method results in parameter estimates that are more robust to outlying probe-level values. In addition, the constant $e$ is added to the signal to ensure that the chip effect is well defined on the logarithmic scale when the signal is very close to zero due to either a copy number of zero or a noisy signal. Adding this constant has the effect of biasing the probe-level data as well as the resulting chip effects, but it lowers the likelihood of obtaining non-defined chip effects.

From the model equation 1.1, notice that the model is SNP-specific in that information between chromosomal positions is not utilized. The chip effect $\psi_{ij}$ summarizes the DNA
concentration present at SNP \( i \) for chip \( j \) accounting for the effect due to probes, \( \zeta_{ik} \).

For copy number analysis we are interested in the chip effect, \( \psi_{ij} \). We will use these chip effects to compute measures of the copy number along the genome. First though, we describe some additional known probe effects that need to be corrected for.

**Normalization for sequence effects**

The design of probe sequences can often influence the binding affinity between the probes and the target DNA. It is well known that the measurement intensity of a probe depends on its GC content as well as its PCR fragment length (Nannya et al. 2005). During a PCR cycle, which takes place after the target DNA has been digested and cleaved by a restriction enzyme, shorter DNA fragments are more likely to be fully replicated compared with longer ones, and thus, as pointed out by Bengtsson et al. (2008), longer extracted DNA fragments are associated with weaker signals. Similarly, probes with a higher GC content have a greater probe affinity since GC bonds are stronger due to having more hydrogen bonds between G and C nucleotides compared to, for example, A and T nucleotides. We correct for PCR fragment length and GC content by modeling the mean of the chip effects as a smooth function of the PCR fragment length and GC content. Readers are referred to Bengtsson et al. (2008) for full details. We denote the corrected chip effects, which account for the sequence effects, by \( \tilde{\psi}_{ij} \).

**Computing single locus raw copy numbers**

For data analysis we typically take the logarithm of the ratio of the chip effect to the estimated SNP-specific diploid reference signal. These measures are called the single locus copy numbers (CNs), or for short, raw CNs. For each SNP \( i \) and sample \( j \), the raw CNs are given by \( Y_{ij} \). Namely,

\[
Y_{ij} = \log_2(\tilde{\psi}_{ij}/\tilde{\psi}_{Ri}).
\]  

(1.2)
Thus, the raw CNs are continuous values, such that sufficiently large raw CNs indicate evidence of a copy number gain, while sufficiently small raw CNs indicate evidence of a copy number loss. The reference diploid signal is the mean diploid signal. In practice, we estimate it as the median across a large control set which is expected to be diploid for the majority of samples. For case-control studies we can use the set of controls to estimate the reference diploid signals. While, in the absence of a set of controls, we can use the HapMap data to estimate the reference diploid signals. In general, however, an in-house reference set is preferable.

Figure 1.2 shows an example of observed raw CNs on chromosome 10 for one HapMap sample, sample NA18550, plotted against the physical genomic location, in bps, of a reference genome. The physical locations of the reference genome used here are based on the coordinates of the SNP markers provided by NCBI, build 37 (hg19) assembly. We note here that in the remainder of the thesis the reported position of a SNP is its position in the reference genome, not its actual position in the target genome being analyzed. The array platform used here is the Affymetrix SNP array 6.0. In total, there are raw CNs for 93,428 genomic markers (consisting of both SNPs as well as some non-polymorphic markers) in this whole-chromosome plot. The points circled in red correspond to SNP markers that lie in regions of copy number polymorphisms as identified in a previous study by McCarroll et al. (2008). As we can see, the raw CNs of these markers exhibit either noticeably higher or lower signals than surrounding markers, however, there is still a substantial degree of noise present. Ultimately, we are interested in inferring these regions, which are sequences of SNPs having non-2-copy values, based on the sequence of observed raw CNs. We will discuss the HapMap data in more detail in Chapter 3.

1.2.3 Other preprocessing algorithms

Bengtsson et al. (2008) showed that their multichip procedure for computing raw CNs, which we described in the previous section, outperforms existing methods based on the
Figure 1.2: Observed raw CNs as a function of the genomic position in mega base pairs (Mbs) for sample NA18550 on chromosome 10.

classification of male and female samples on the nonpseudoautosomal portion of the X chromosome. (Males only have one copy of DNA on the nonpseudoautosomal portion of the X chromosome, while females have two.) Here we compare their procedure to another competitive method, probe-level allele-specific quantitation (PLASQ) (Laframboise et al. 2007), which Bengtsson et al. (2008) did not discuss. We used 60 HapMap samples (25 female, 35 male) with various ethnicities to conduct the analysis. The average (across SNPs) ROC curve is shown in Figure 1.3, where TPR is the true positive rate which for this example corresponds to correctly calling males males, and FPR is the false positive rate which for this example corresponds to incorrectly calling females males. We see that
CRMA and PLASQ are quite similar, although initially CRMA tends to rise at a slightly higher rate, implying that for conservative thresholds (low FPRs) it does a slightly better job at identifying males. On average, CRMA shows good accuracy in classifying males as males.

Figure 1.3: Average ROC curves for CRMA and PLASQ. The ROC curves are averages across 5605 SNPs on the X chromosome.

1.3 Thesis aims

This thesis investigates the use of HMMs for inferring DNA copy number from SNP genotyping arrays. The central aim of the thesis is to develop a more statistical framework that appropriately deals with uncertainty around model parameters from the array under analysis. An ideal technique would also determine measures of uncertainty for each detected region from the HMM. This is important because researchers often require the ability to rank detected regions so that they can focus further experimental efforts on the most promising regions.
Chapter 2 of this thesis provides a comprehensive review of HMMs including areas of application and algorithms for parameter estimation such as the forward-backward algorithm and the Viterbi algorithm. Limitations, previous implementations of HMMs for copy number data, and an outline of the structure of the rest of the thesis, are provided at the end of Chapter 2.
Chapter 2

Review of Hidden Markov Models for Copy Number Analysis

2.1 Introduction

Copy number variations (CNVs) and the problem of inferring them from SNP genotyping arrays were introduced in Chapter 1. Copy number data from normal cell lines is often characterized by long stretches of no variation, interspersed with typically small regions of CNVs (that is, subchromosomal regions with variable DNA copy number). Hidden Markov models (HMMs) are well-suited to modelling such sudden changes in the data, enabling them to make good predictions of copy number along the genome. For this reason, HMMs are a commonly used technique for the genome-wide detection of CNVs. Typically, HMMs are estimated in two steps:

1. Model parameters, including the parameters of the emission distribution and the transition probability matrix, are estimated using the Baum-Welch algorithm.

2. Conditional on the estimated model parameters, the most probable DNA copy number sequence is estimated using the Viterbi algorithm.
The Baum-Welch algorithm (Baum et al. 1970) is a special case of an EM algorithm (Dempster et al. 1977) that uses the forward-backward algorithm for HMMs, and the Viterbi algorithm (Viterbi 1967) is a dynamic programming algorithm which efficiently reconstructs the most probable sequence of hidden states underlying an observed sequence for a given set of model parameters.

This chapter first provides a detailed description of the HMM, then reviews the individual steps of the procedure for HMM estimation and highlights some of the shortcomings of this procedure. Finally, the chapter reviews some of the previous research in the literature on copy number analysis and outlines the rest of the thesis.

2.2 Hidden Markov models

This section first details the general structure of HMMs. The forward-backward algorithm and the Viterbi algorithm are then introduced as methods for model estimation.

2.2.1 Background

The HMM is a statistical model for sequential data consisting of a sequence of $M$ observable random variables, $Y = (Y_1, \ldots, Y_M)$, and a corresponding sequence of unobservable, or hidden, random variables, $C = (C_1, \ldots, C_M)$. The hidden variables are discrete multinomial variables such that $C_i \in (1, \ldots, K)$, $i = 1, \ldots, M$, where $K$ is the number of hidden states and $M$ is the number of observations. The observable random variables could be either continuous or discrete. In this work we will be focusing on observed data that are continuous. In the case of copy number data, $Y_i$ would correspond to the raw CN at SNP $i$ for a given sample of interest, and $C_i$ would correspond to the unobserved DNA copy number at SNP $i$, which we want to infer. In this thesis we will focus on single-sample methods, hence we drop the subscript used in Chapter 1 representing the sample $j$. 
The dependence structure of a HMM is shown in Figure 2.1 as a directed graphical model. Here, the circles represent random variables and the edges between variables represent probabilistic relationships between the variables, with the direction of the arrow indicating the direction of the dependence. This figure implies the following model assumptions.

1. **Markov assumption**: \( P(C_k = c_k | C_{k-1} = c_{k-1}, \ldots, C_1 = c_1) = P(C_k = c_k | C_{k-1} = c_{k-1}) \), or in words, the distribution of the current hidden state depends only on the previous hidden state.

2. **Conditional independence**: \( P(Y_k | C_1, \ldots, C_M, Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_M) = P(Y_k | C_k) \), or in words, conditional on the current hidden variable, the observed variable at position \( k \) is independent of all other observed variables and hidden variables.

It is easy to see that the standard mixture model is a special case of the HMM in which the hidden variables are independent of one another, which corresponds to removing the horizontal arrows in Figure 2.1.

![Figure 2.1: Graphical representation of the dependence structure of a HMM.](image)

Within this structural framework, the HMM is specified through three components. The first, the transition probability matrix, \( \Lambda \), is a \( K \times K \) matrix that describes the transitions of the Markov chain in terms of the probabilities of moving between the various states. Specifically, the elements of \( \Lambda \) are the transition probabilities which are defined by \( \lambda_{ij} = P(C_t = j | C_{t-1} = i) \), so that \( 0 \leq \lambda_{ij} \leq 1 \) and \( \sum_j \lambda_{ij} = 1 \) for \( i = 1, \ldots, K \).
Typically it is assumed that the transition probability matrix is the same between all of the adjacent observations. This is referred to as a homogeneous model. In some cases, however, transitions between states may depend on additional factors such as the distance between the observational units. In such cases heterogeneous models, in which we allow $\Lambda$ to depend on these factors, are more appropriate. The second component, the initial state distribution, $\pi = (\pi_1, \ldots, \pi_K)$, is the probability distribution of the hidden variable of the first observational unit which is special in that it has no parents or preceding observations. Specifically, $\pi_k = P(C_1 = k)$. For the third component we need to assume a form for the conditional distribution of the observed variables $f_k(Y_t) = P(Y_t|C_t = k)$, called the emission distribution of the observed data, or just the emission. Of course, the emission will depend on a set of parameters that govern this distribution when the data is from class $k$. A common choice for the emission distribution for continuous observed variables is a Normal distribution, in which case the emission model parameters are the class mean, $\mu_k$, and the class variance, $\sigma_k^2$.

The theory of HMMs was first introduced by Leonard E. Baum and his colleagues in a series of seminal papers published in the late 1960s (Baum & Petrie 1966, Baum & Egon 1967, Baum et al. 1970).Shortly after, HMMs were applied to speech recognition problems (Bahl et al. 1974, Jelinek 1976). This was one of the first major applications of HMMs. Essential to the subsequent proliferation of the use of HMMs was a paper by Rabiner (1989) in which solutions for solving important problems on using HMMs were provided, such as how to efficiently compute the probability of the observed sequence, and how to determine the most probable sequence of hidden states given the observed sequence. Since then, the use of HMMs has proliferated and they have been used in a wide range of applications, some of which include bioinformatics, weather monitoring, finance, and online handwriting recognition.
2.2.2 Fitting HMMs

The parameters of the HMM are estimated by maximizing the likelihood \( P(Y|\theta) \) (viewed as a function of \( \theta \)), where \( \theta = (\mu, \sigma^2, \Lambda, \pi) \) and \( \mu = (\mu_1, \ldots, \mu_K) \) and \( \sigma^2 = (\sigma_1^2, \ldots, \sigma_K^2) \).

The likelihood is obtained by marginalizing over the hidden variables \( C \):

\[
P(Y|\theta) = \sum_C P(Y, C|\theta). \tag{2.1}
\]

The difficulty with evaluating this equation is that the number of terms in the sum is \( K^M \). Thus, the number of terms grows exponentially with the length of the chain, making it infeasible to compute directly, even when \( K \) is small. Furthermore, due to the Markov property of the model, the joint probability \( P(Y, C) \) does not factorize over \( M \), so we cannot treat the terms in the summation independently. As a result, direct maximization of equation 2.1 with respect to \( \theta \) is difficult. Due to these difficulties the standard approach to maximize the likelihood is to use a special version of the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) which utilizes the forward-backward algorithm to efficiently compute some of the quantities in the E-step, resulting in substantial computational savings (shown below).

The algorithm starts by picking initial values for the parameters, denoted by \( \theta^{\text{old}} \), and then it iterates between two steps. The first, called the E-step, computes the expected value of the log of the complete-data likelihood, \( P(Y, C|\theta) \), with respect to the conditional distribution of the hidden variables given the data \( Y \) and the current parameter estimates, \( P(C|Y, \theta^{\text{old}}) \). (Complete data refers to the situation where the copy number of each SNP is observed or known.) We treat this expectation as a function of \( \theta \) and denote it as \( Q(\theta, \theta^{\text{old}}) \). The second step, called the M-step, updates the parameters \( \theta \) by maximizing the expectation \( Q(\theta, \theta^{\text{old}}) \) with respect to \( \theta \). These two steps are repeated until convergence of the algorithm. For the HMM, the expected complete-data log likelihood
is given by

\[
Q(\theta, \theta^{\text{old}}) = \mathbb{E}_{C|Y, \theta^{\text{old}}} [\log P(Y, C|\theta)] \\
= \mathbb{E}_{C|Y, \theta^{\text{old}}} \left[ \sum_{k=1}^{K} \mathbf{1}(C_1 = k) \log(\pi_k) + \sum_{m=2}^{M} \sum_{j=1}^{K} \sum_{k=1}^{K} \mathbf{1}(C_m = k, C_{m-1} = j) \log \lambda_{jk} + \right. \\
\left. \sum_{m=1}^{M} \sum_{k=1}^{K} \mathbf{1}(C_m = k) \log P(Y_m|\mu_k, \sigma_k^2) \right] \\
= \sum_{k=1}^{K} \gamma_1(k) \log(\pi_k) + \sum_{m=2}^{M} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi_m(j, k) \log \lambda_{jk} + \sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_m(k) \log P(Y_m|\mu_k, \sigma_k^2). 
\]

Here,

\[
\gamma_j(k) = P(C_j = k|Y, \theta^{\text{old}}), 
\]

the posterior probability that the latent variable \( C_j \) equals \( k \), and

\[
\xi_m(j, k) = P(C_{m-1} = j, C_m = k|Y, \theta^{\text{old}}), 
\]

the joint posterior probability that two successive latent variables satisfy \( C_m = k \) and \( C_{m-1} = j \). We will refer to the \( \gamma \)'s as the responsibilities. Clearly the maximizers of \( Q(\theta, \theta^{\text{old}}) \) will be functions of the responsibilities and the \( \xi \)s. The next section shows how the forward-backward algorithm is used to compute the \( \gamma \)s and \( \xi \)s efficiently.

### 2.2.3 The forward-backward algorithm

The forward-backward algorithm, which in the context of HMMs was first proposed by Baum et al. (1970) and subsequently further developed by Rabiner (1989), allows us to compute the quantities required for the E-step of the EM-algorithm in an efficient manner by using a set of recursive equations. In what follows, we outline the derivation of the algorithm.

From Bayes’ theorem, we know that

\[
\gamma_m(j) = P(C_m = j|Y, \theta) = \frac{P(Y|C_m = j, \theta)P(C_m = j|\theta)}{P(Y|\theta)}. 
\]
We let $\alpha_m(j)$ represent the joint probability of observing the data up to time $m$ and the state $j$ for the hidden variable at time $m$:

$$\alpha_m(j) = P(Y_1, \ldots, Y_m, C_m = j).$$ \hfill (2.5)$$

These are referred to as the forward variables. Note that for convenience we have suppressed the dependence on $\theta$. Similarly, we let $\beta_m(j)$ represent the conditional probability of observing the rest of the observation sequence beyond time $m$, conditional on the hidden variable being in state $j$ at time $m$:

$$\beta_m(j) = P(Y_{m+1}, \ldots, Y_M | C_m = j).$$ \hfill (2.6)$$

These are referred to as the backward variables. The set of forward and backward variables can be combined to get the marginal distributions of the hidden variables. Specifically, by using equation 2.4 and the conditional independence of $Y_{m+1:M}$ and $Y_{1:m}$ given $C_m$, we can express $\gamma_m(j)$ in terms of $\alpha$ and $\beta$:

$$\gamma_m(j) = \frac{P(Y_{m+1}, \ldots, Y_M | C_m = j) P(Y_1, \ldots, Y_m | C_m = j, Y_{m+1}, \ldots, Y_M) P(C_m = j)}{P(Y)}$$

$$= \frac{P(Y_{m+1}, \ldots, Y_M | C_m = j) P(Y_1, \ldots, Y_m | C_m = j) P(C_m = j)}{P(Y)}$$

$$= \frac{\beta_m(j) \alpha_m(j) P(Y)}{P(Y)} \hfill (2.7)$$

Again, we will make use of conditional independence properties to express $\alpha_m(j)$ in terms
of $\alpha_{m-1}(j)$. For $j = 1, \ldots, K$,

$$
\alpha_m(j) = P(Y_m | C_m = j)P(Y_1, \ldots, Y_{m-1} | C_m = j)P(C_m = j) \tag{2.8}
= P(Y_m | C_m = j)P(Y_1, \ldots, Y_{m-1}, C_m = j) \tag{2.9}
= P(Y_m | C_m = j) \sum_{i=1}^{K} P(Y_1, \ldots, Y_{m-1}, C_m = j, C_{m-1} = i) \tag{2.10}
= P(Y_m | C_m = j) \sum_{i=1}^{K} P(Y_1, \ldots, Y_{m-1}, C_m = j | C_{m-1} = i)P(C_{m-1} = i) \tag{2.11}
= P(Y_m | C_m = j) \sum_{i=1}^{K} P(Y_1, \ldots, Y_{m-1} | C_{m-1} = i)P(C_m = j | C_{m-1} = i)P(C_{m-1} = i) \tag{2.12}
= P(Y_m | C_m = j) \sum_{i=1}^{K} \alpha_{m-1}(i)P(C_m = j | C_{m-1} = i), \tag{2.13}
$$

and

$$
\alpha_1(j) = P(Y_1, C_1 = j) = P(C_1 = j)P(Y_1 | C_1 = j). \tag{2.15}
$$

Equation 2.8 follows from $Y_m$ being independent of $Y_1, \ldots, Y_{m-1}$ given $C_m$, and equation 2.12 follows from $Y_1, \ldots, Y_{m-1}$ being independent of $C_m$ given $C_{m-1}$. At each step in the recursion there are $K$ terms in the summation which is computed for each value of $j = 1, \ldots, K$. Therefore, the overall cost of the forward part of the algorithm is $O(K^2M)$; that is, it is linear in the length of the chain.

Similarly, we can derive recursive equations that express $\beta_m$ in terms of $\beta_{m+1}$. For
\( j = 1, \ldots, K, \)

\[
\beta_m(j) = P(Y_{m+1}, \ldots, Y_M|C_m = j) \\
= \sum_{i=1}^{K} P(Y_{m+1}, \ldots, Y_M, C_{m+1} = i|C_m = j) \\
= \sum_{i=1}^{K} P(Y_{m+1}, \ldots, Y_M|C_{m+1} = i, C_m = j) P(C_{m+1} = i|C_m = j) \\
= \sum_{i=1}^{K} P(Y_{m+2}, \ldots, Y_M|C_{m+1} = i) P(Y_{m+1}|C_{m+1} = i) P(C_{m+1} = i|C_m = j) \\
= \sum_{i=1}^{K} \beta_{m+1}(i) P(Y_{m+1}|C_{m+1} = i) P(C_{m+1} = i|C_m = j). 
\]

We can obtain a starting condition for the backward recursions by evaluating (2.7) with \( m = M, \) giving

\[
P(C_M = j|Y) = \frac{P(Y, C_M = j)\beta_M(j)}{P(Y)}. \tag{2.16}
\]

Thus, to satisfy this equation we should set \( \beta_M(j) = 1 \) for all \( j. \) Also, from this equation it is easy to see that the likelihood \( P(Y) \) is given by

\[
P(Y) = \sum_{i=1}^{K} \alpha_m(i)\beta_m(i) \tag{2.17}
\]

for any choice of \( m. \) The most convenient value to use is \( m = M, \) which gives

\[
P(Y) = \sum_{i=1}^{K} \alpha_M(i). \tag{2.18}
\]

Therefore, by using the recursive equations we have reduced the computational cost of computing the likelihood from \( O(K^M) \) to \( O(K^2M), \) an enormous reduction when \( K \) is large.

Finally, we consider computation of the quantity \( \xi_m(j, k). \) Once again we use Bayes’ theorem to write \( \xi_m(j, k) \) as

\[
\xi_m(j, k) = \frac{\alpha_{m-1}(j)P(Y_m|C_m = k)P(C_m = k|C_{m-1} = j)\beta_m(k)}{P(Y)}. \tag{2.19}
\]
Thus, we can also compute \( \xi_m(j, k) \) through the recursions for the \( \alpha \) and \( \beta \) variables.

### 2.2.4 The Viterbi algorithm

The Viterbi algorithm is an efficient algorithm for finding the most probable sequence of hidden states, denoted by \( C = (C_1, \ldots, C_M) \), given an observation sequence \( Y \) and model parameters \( \theta \). That is, the Viterbi algorithm solves

\[
\hat{C} = \arg \max_C p(C | Y, \theta).
\]  
(2.20)

Note that the solution is a function of the estimated model parameters, so the Viterbi solution can be denoted as \( \hat{C}(\theta) \). The algorithm, given below, was invented by Viterbi (1967) and later implemented for HMMs by Rabiner (1989). Following closely the notation in Rabiner (1989) we define

\[
\delta_t(i) = \max_{c_1, c_2, \ldots, c_{t-1}} P(C_1 = c_1, C_2 = c_2, \ldots, C_t = i, Y_1, Y_2, \ldots, Y_t | \theta),
\]  
(2.21)

which is the highest probability along a single path ending in state \( i \) at location \( t \). Then the algorithm for finding the most probable sequence of hidden states is as follows:

1. **Initialization:**

\[
\delta_1(i) = P(C_1 = i) P(Y_1 | C_1 = i), \quad 1 \leq i \leq K.
\]  
(2.22)

2. **Recursion:** for \( t = 2, \ldots, M \),

\[
\delta_t(j) = \max_{1 \leq i \leq K} (\delta_{t-1}(i) P(C_t = j | C_{t-1} = i)) P(Y_t | C_t = j), \quad 1 \leq j \leq K
\]  
(2.24)

\[
\psi_t(j) = \arg \max_{1 \leq i \leq K} (\delta_{t-1}(i) P(C_t = j | C_{t-1} = i)), \quad 1 \leq j \leq K.
\]  
(2.25)

3. **Termination:**

\[
P^\star = \max_{1 \leq i \leq K} \delta_M(i)
\]  
(2.26)

\[
q_M^\star = \arg \max_{1 \leq i \leq K} \delta_M(i)
\]  
(2.27)
4. Path backtracking:

\[ q_t^* = \psi_{t+1}(q_{t+1}^*), \; t = T - 1, T - 2, \ldots, 1. \] (2.28)

2.3 Previous research

As mentioned at the beginning of the chapter, after applying the existing normalization and summarization methods to the probe-level data (which are described in Chapter 1), DNA copy number estimation from SNP arrays, using HMMs, typically consists of two steps. First, the Baum-Welch algorithm is used to estimate both the emission model parameters and the parameters of transition probability matrix. Second, the Viterbi algorithm computes the most probable sequence of copy number states based on the data and the estimated model parameters, giving an estimated DNA copy number sequence. Within this framework, numerous authors have already proposed the use of HMMs for inferring DNA copy number from SNP arrays. They include Zhao et al. (2004) (dChip), Colella et al. (2007) (QuantiSNP), Wang et al. (2007) (PennCNV), Wang et al. (2009) and Korn et al. (2008) (Birdseye). All of these HMMs use the Viterbi algorithm to reconstruct the DNA copy number. With respect to model specification and estimation (first step), QuantiSNP assumes that the copy number states are equally probable, that is, for each array the frequency of each copy state is the same. This is not a realistic assumption. PennCNV generalizes the HMM of QuantiSNP by relaxing this assumption, however, both QuantiSNP and PennCNV fix parameters of the transition probability matrix to somewhat arbitrary values. The HMM of Korn et al. (2008) also fixes parameters of the transition probability matrix so that transition probabilities out of the 2-copy state are low, while transition probabilities within the same state or returning to the 2-copy state are relatively high (Korn et al. 2008). They also fix parameters of the emission distributions based on training data and extrapolation techniques. Therefore, with respect to the first step of HMM estimation, a common problem with these HMMs
is that they fix model parameters to arbitrary values. Fixing the HMM model parameters before using the Viterbi algorithm is a problem for two main reasons: (i) it does not properly account for model parameter uncertainty, potentially leading to inaccurate inferences; and (ii) model parameters may actually differ between samples and/or chromosomes and assuming that the same set of model parameters applies for all samples and/or chromosomes is not realistic. These problems were also pointed out by Rydén (2008) who also did a comparison of HMM parameter estimation using the EM algorithm and using MCMC sampling for various HMMs and datasets, including a continuous-time HMM for array CGH data. In this case, he found that the EM-algorithm had problems getting convergence for the parameters of the transition intensity matrix, whereas the fully Bayesian approach produced more sensible parameter estimates. Thus, there appear to be some challenges in properly estimating model parameters of HMMs in the context of copy number data.

While dChip, QuantiSNP, PennCNV and Birdseye all use transition probability matrices that account for the physical distance between adjacent SNPs, none of these models are developed based on, or motivated by, a real stochastic process (e.g., a Markov process). Wang et al. (2009) develop a more realistic model in which they use a continuous-time HMM (CHMM) for the copy number analysis of Illumina genotyping arrays. Their CHMM, which was developed concurrently with ours, assumes that the transition rates between copy states are independent of the originating copy state (that is, the parameters of the columns of the $Q$-matrix, which govern the rate of transition between the copy states, are all the same). However, this may not be true in practice as the probability of transitioning to the 1-copy state may be higher when starting from the 2-copy state than the 3-copy state, for example. Furthermore, Wang et al. (2009) propose a Majorize-Minorize (MM) algorithm for parameter estimation of their model parameters, however, when analyzing real data they fix model parameters, suggesting that the MM algorithm does not work well for real data.
While the Viterbi algorithm is one of the most useful dynamic programming algorithms, there are a few drawbacks associated with using the Viterbi algorithm for copy number estimation. The issues are associated with its deterministic nature which does not support proper statistical inference. First, the Viterbi algorithm does not fully account for the uncertainty in the model parameters. That is, its solution, $C(\hat{\theta})$, is conditional on a single point estimate $\hat{\theta}$ and it does not consider other reasonable values of $\theta$. A consequence of this is that it makes it difficult to quantify the strength of the evidence of a detected CNV. Not surprisingly, then, most existing copy number methods do not assess the statistical significance of detected regions, nor can they rank them. This is a significant drawback which we discuss below. Other authors have recognized this shortcoming and have proposed alternative approaches; see e.g. Churchill & Lazareva (1999). Furthermore, as pointed out by Churchill & Lazareva (1999), we may not only be interested in the most probable copy number sequence, but also wish to have access to alternative copy number sequences. For example, we may be interested in the top 10 most probable copy number sequences. The focus of this dissertation will be on developing more statistical approaches to applying HMMs for the analysis of SNP arrays.

After running the Viterbi algorithm, we are left with a sequence of DNA copy number values. A sequence of non-2-copy values is taken as corresponding to a CNV. Due to the fact that these HMMs are more deterministic in nature in that they use the Viterbi algorithm to reconstruct the DNA copy number sequence, they do not produce uncertainty measures for detected CNVs. That is, they do not produce measures that researchers can use to rank detected CNVs from a genome-wide analyses. As noted by Guha et al. (2008), this is a problem when reporting results since researchers may have enough resources to experimentally validate only a portion of the regions that are found from a genome-wide scan. There are only a few methods that address this problem, all of which were developed for array CGH data. They include Guha et al. (2008) and Lai et al. (2008), both of which were developed for the analysis of tumour samples using array

2.4 Outline of the rest of the thesis

Chapter 3 of this thesis explores the use of Bayesian HMMs, which place prior distributions on both the parameters of the emission distributions and the transition probability matrix. Thus, not only do Bayesian HMMs account for full model parameter uncertainty but they also allow one to incorporate a priori biological information via specification of prior distributions. The chapter proposes a novel implementation of the forward-backward stochastic algorithm, within a Markov chain Monte Carlo (MCMC) algorithm, for sampling the DNA copy number sequence. Compared to previous implementations of HMMs for copy number data, which use the Viterbi algorithm to maximize $P(C|Y, \hat{\theta})$, this approach has the advantage of integrating over the model parameters and focuses on $P(C|Y)$ instead of $P(C|Y, \hat{\theta})$. The chapter also discusses Bayesian methods for DNA copy number estimation based on the MCMC results. Bayesian continuous-time HMMs, which use a continuous-time Markov process to model the DNA copy number, are studied in Chapter 4 and a modified version of the Bayesian continuous-time HMM with copy-number-specific emission distributions is presented in Chapter 5. Each of these chapters considers different versions of Bayesian HMMs. To assess the performance of the different models, real Affymetrix SNP array data from the HapMap study will be used. Known inherited CNVs will be used as a performance benchmark.

Next, Chapter 6 discusses Bayesian uncertainty measures for subsequences, or DNA regions, that are suspected of containing a CNV.

Finally, Chapter 7 summarizes the key results and conclusions, and discusses areas of future work including extending the models for family trio data and multiple sample
analysis.
Chapter 3

A Bayesian Discrete-Time HMM with Normal Emissions for Copy Number Data

3.1 Introduction

As highlighted in Chapter 2, previous implementations of HMMs for copy number (CN) data from SNP genotyping arrays include Wang et al. (2009), Colella et al. (2007) (QuantiSNP), Wang et al. (2007) (PennCNV), and Korn et al. (2008) (Birdseye). All of these HMMs fix unknown parameters of the corresponding transition probability matrix, $T(d)$. For example, the HMM of Colella et al. (2007) assumes that the copy number states are equally probable, which is not a realistic assumption, and both Colella et al. (2007) and Wang et al. (2007) fix key parameters of $T(d)$ to somewhat arbitrary values. Korn et al. (2008) fix parameters of the transition probability matrix in their HMM so that “transitioning out of a state reflecting normal copy number is low, whereas transitioning within the same state or returning to a normal copy number is relatively high.” They also fix parameters of the emission distributions based on training data and extrapolation
techniques. On the other hand, Wang et al. (2009) propose a Majorize-Minorize (MM) algorithm for parameter estimation of their model parameters, however, when analyzing real data they fix model parameters, suggesting problematic implementation of these models with full parameter uncertainty to the noisy genomic data. Thus, one of the problems with current implementations of HMMs in the field is that they fix unknown model parameters. Although this allows for a simpler implementation of these complicated models, as stated by Rydén (2008), such an approach is not appropriate because, from a pragmatic point of view, HMM model parameters may differ between samples and/or chromosomes. (Typically analysis is carried out chromosome-by-chromosome.) Thus, assuming the same model parameters apply to all chromosomes and samples, within and between studies, is inappropriate, especially for parameters of the transition probability matrix which are dependent on the number of CNVs a given sample contains. Furthermore, from a more statistical point of view, fixing model parameters does not properly account for model parameter uncertainty and may result in invalid inferences due to underestimation of variances.

As indicated in Chapter 1, the central aim of the thesis is to develop a more statistical framework for the copy number analysis of SNP array data that in estimating DNA copy number appropriately deals with uncertainty around model parameters from the array under analysis. This chapter explores the use of HMMs for copy number analysis. First, SNP genotyping array data with known biological structure, specifically known DNA copy number, are explored. A standard HMM is then introduced as a Bayesian model that accounts for full parameter uncertainty, and a novel implementation of a forward-backward algorithm for sampling copy number sequences is proposed for this model. The remainder of the chapter discusses model behaviour, how to estimate DNA copy number from the posterior distribution, and concludes with an analysis of real Affymetrix SNP array 6.0 data. The data analysis shows that this implementation of the standard HMM model with full parameter uncertainty does not actually work well when analyzing real
SNP genotyping array data, which can be quite noisy, in the sense that the results are biologically implausible.

### 3.2 Two real CNV data sources

After providing a detailed overview of the data structure of copy number data from SNP arrays, two real data sources that are used in the rest of the thesis are introduced and a simple exploratory analysis of array samples with known DNA copy number is given. To prevent confusion, note that although these real data sources consist of multiple arrays or samples, the methods for copy number estimation that we develop later in the chapter apply only for the analysis of a single array, so that multiple arrays would be analyzed separately. The data used in the exploratory analysis is a special case in which X-chromosome DNA was artificially constructed and thus these arrays have known DNA copy number. In this case information is utilized from all the arrays in the study to estimate the distribution of the data having various levels of DNA copy number.

Of course, genomic studies almost always involve multiple samples, and even in single-sample analysis information from the other arrays is used to compute the reference signal at each SNP (refer to Section 1.2.2 for details). A new model for multiple sample analysis, which utilizes information both across SNPs and between arrays, is discussed in Chapter 7.

#### 3.2.1 Observed data from SNP genotyping arrays

The data arising from SNP genotyping arrays are a sequence of probe signal intensities which we normalize and process using the procedure described in Section 1.2.2. This results in a sequence of raw CNs $Y = (Y_1, \ldots, Y_M)$ along with their corresponding physical locations, denoted by $d_j$, along the genome in base pairs (bps), where $M$ is the number of SNPs on the chromosome we are analyzing. For now, we restrict attention to a single
sample and data from different chromosomes are analyzed independently of each other.

The copy number process records the number of copies of DNA at specific locations along the genome. We let $C = C(t)_{0 \leq t \leq T}$ denote the unobserved copy number process corresponding to the sample under analysis, where $T$ is the length of the chromosome in bps and $C(t) \in \{0, 1, \ldots, K\}$, where $K$ is the maximum number of copies of DNA we allow in the model. That is, the DNA copy number at a given location is a non-negative integer and we tend to observe long stretches of DNA having copy number 2 (the diploid state), interspersed with short to medium length sequences of DNA having non-2-copy number. Typically $K$ is between 3 and 5. For convenience, we will denote the copy number at the observed SNP locations $(C(d_1), \ldots, C(d_M))$ as $(C_1, \ldots, C_M)$.

Again, our goal is to infer $C$ based on the observed data $Y$ and, more specifically, we are interested in finding sequences of non-2-copy values which would correspond to CNVs.

### 3.2.2 Titration data

The X chromosome titration data set (3X, 4X and 5X) contains three artificially constructed DNA samples containing abnormal amplification of the whole X chromosome (aneuploidies). There are four replicates of each DNA sample. The aneuploidies are a X trisomy (presence of three copies of chromosome X); a X chromosome tetrasomy (presence of four copies of chromosome X); and an X chromosome pentasomy (presence of five copies of chromosome X). These data were downloaded from the Affymetrix data resource center. The Coriell Cell Repository numbers for these cell lines are NA04626 (3X), NA01416 (4X), and NA06061 (5X). This is a unique data set in that the DNA samples have known biological structure. In the next section, we will use this data to explore the distribution of copy number measurements for various levels of copy number.
3.2.3 HapMap data

The International HapMap Project (The International HapMap Consortium 2005) is an international research endeavour whose primary goal is to map the human genome, i.e. to develop a haplotype map (HapMap) of the human genome. As a part of the project, genomic data from 270 individuals of varying ethnic backgrounds was collected. This included copy number data from SNP genotyping arrays. We will use these data to exemplify our methods. McCarroll et al. (2008) report genomic coordinates for 1,320 copy number polymorphisms (CNPs, i.e., common CNVs) from the 270 HapMap samples. To find these CNPs they used a model that examined all sets of narrow windows along the genome and, within each window, looked for highly correlated regions of cross-sample intensity patterns (see McCarroll et al. (2008) for more details). They further validated these regions by analyzing data from an independent experiment. However, the set of 1320 CNPs may not be exhaustive and certainly some uncertainty may still exist in the endpoints of these intervals. In particular, this set is probably biased toward larger and more common CNPs, which are the easiest to find. Nonetheless, currently this is the best available resource for copy number information of the HapMap samples and, when analyzing the HapMap data, we will use these common CNVs to guide and evaluate our methods, keeping in mind that it is not a perfect gold-standard.

Furthermore, such data could also be used to inform about mean lengths of copy number deletions and of amplifications, which are both hyperparameters of the continuous-time Markov model that we propose in Chapter 4. Figure 3.1 shows the histogram of the inferred lengths for the 1320 CNPs reported in McCarroll et al. (2008). Based on this study the mean length of CNPs is 26,460 bps.
Figure 3.1: Histogram of the lengths of the 1,320 CNPs reported in McCarroll et al. (2008). 32 CNPs with lengths greater than 200,000 bps were truncated from this histogram.

3.2.4 Analysis of samples with known DNA copy number of the X chromosome

As an exploratory analysis of data with known biological structure, specifically, known DNA copy number of the X chromosome, we used the HapMap data for $c \in (1, 2)$ and the titration data for $c \in (3, 4, 5)$. For $c = 1$, X chromosome data on the non-pseudoautosomal portion of the X chromosome, which consists of 86,632 raw CNs from male samples was used. There were 142 male samples in total. For $c = 2$, X chromosome data from female samples was used. There were 128 female samples in total, but 4 were removed because they were suspected of having X chromosome deletions (Ting et al. 2006), leaving 124 in the analysis. Note that males have only one copy of the X chromosome while females have two copies. Thus, X chromosome data from male samples
corresponds to a 1-copy state. Table 3.1 show estimates of the mean and variability for each copy number state. For $c \in (1, 2)$, where we have multiple samples, we pool estimates of the mean and variance across the samples. As expected, as the DNA copy number increases so does the mean level of the raw CNs, going from -0.64 for the 1-copy state to 0.85 for the 5-copy state. However, we do observe higher variability for non-2-copy state observations. In particular, the 1-copy state observations exhibit twice the variability as the 2-copy state observations, whereas observations in the the amplified states exhibit no more than 50% more variability than the observations in the 2-copy state. Figure 3.2 shows boxplots of raw CNs for 87,200 SNPs (86,632 SNPs for C=1,2) on the X chromosome. Data in each box corresponds to a different DNA copy number level, ranging from one to five copies of DNA. There is a noticeable bias between the theoretical values (the log ratios) and those that we observed in the data. Also, there is a substantial degree of noise present with notable outliers. The mean outlier rates across samples were 0.04, 0.03, 0.02, 0.02 and 0.03 for the 1-, 2-, 3-, 4- and 5-copy states, respectively, where the outliers were identified as those observations that were either above or below the first or third quartiles by an amount of 1.5 times the interquartile range (IQR). Under a Normal distribution, we would expect an outlier rate of approximately 0.007. Thus, the distribution of copy number data from SNP arrays is noisier than what we would expect under a Normal distribution.

Estimates of the state means and variances for these data will be used to specify prior distributions, or model hyperparameter values, of the emission distributions of the models that are discussed in the next section.

3.3 Bayesian Discrete-time HMMs

HMMs as methods for inferring DNA copy number were reviewed in Chapter 2. However little attention has been given to the challenge of estimating HMM model parameters
Figure 3.2: Raw CNs averaged across replicates for X chromosome titration data (3X, 4X, 5X), and raw CNs for one male (1X) and one female (2X) HapMap sample. The separate boxplots correspond to DNA samples with various numbers of copies of the X chromosome. The dotted lines correspond the theoretical log-ratio values: \( \log(1/2), \log(2/2), \ldots, \log(5/2) \).

for SNP array data, as well as using Bayesian HMMs, which account for full parameter uncertainty. To gain an understanding of the challenges and extensions possible in the process of building a hidden Markov model, a standard Bayesian discrete-time HMM (BDHMM) for copy number data is first reviewed, including an investigation into the model behaviour in terms of the probability distribution of a copy number sequence.

3.3.1 Copy number model

For ease of exposition, it is assumed that the copy number process can take three possible values: 1 (haploid), 2 (diploid) or 3 (triploid). (This could easily be extended to include more states, such as 0- and 4-copy states.) We consider a discrete-time model which
Table 3.1: Estimated mean and standard deviation for various levels of DNA copy number of the X chromosome.

corresponds to a general transition probability matrix for transitions between the copy number states,

\[
\Lambda = \begin{pmatrix}
\lambda_{11} & \lambda_{12} & \lambda_{13} \\
\lambda_{21} & \lambda_{22} & \lambda_{23} \\
\lambda_{31} & \lambda_{32} & \lambda_{33}
\end{pmatrix},
\]

where \( \lambda_{jk} = P(C_t = k|C_{t-1} = j) \). The 9 parameters of the transition probability matrix are denoted as \( \Lambda = (\lambda_1, \lambda_2, \lambda_3) \) where \( \lambda_j = (\lambda_{j1}, \lambda_{j2}, \lambda_{j3}) \). Thus, we can write the (unobserved) likelihood of \( C \) as

\[
P(C|Y, \theta, \phi) \propto \lambda_1^{N_1} \lambda_2^{N_2} \lambda_3^{N_3}
\]  

(3.1)

where \( N_j = (N_{j1}, N_{j2}, N_{j3}) \), \( \lambda_j^{N_j} = \lambda_{j1}^{N_{j1}} \lambda_{j2}^{N_{j2}} \lambda_{j3}^{N_{j3}} \), and \( N_{jk} = \sum_{t=2}^M 1(C_t = k, C_{t-1} = j) \), the number of transitions from state \( j \) to state \( k \).

The raw CNs are assumed to be generated from a conditional Gaussian model, whose parameters depend on the underlying and hidden, CN state. These Gaussian distributions are usually called emission distributions. So, independently for all \( i \),

\[
[Y_i|\mu_c, \sigma_c^2, C_i = c] \sim N(\mu_c, \sigma_c^2).
\]  

(3.2)

This is the simplest specification and, as we saw in the previous section, outliers do show up in SNP genotyping array data and can present serious problems if not accounted for
correctly by the model. Recently, mixture distributions have been used to account for the presence of outliers in SNP genotyping array data. See, for example, Colella et al. (2007) who used a mixture of a Normal and a Uniform distribution.

\[ [Y_i | \mu_c, \sigma_c^2, C_i = c] \sim \pi_{out,c}/(Y_{max} - Y_{min}) + (1 - \pi_{out,c})N(y_i; \mu_c, \sigma_c^2) \].

The uniform distribution is meant to capture outlying data points. We adopt a model with class-specific outlier rates to allow for the possibility that these rates may be different between copy states. In addition, we denote the sequence of outlier states at the observed SNP locations by Out = (Out_1, \ldots, Out_M) where a value of 1 for Out_i indicates that SNP i is an outlier within copy state C_i, and a value of 0 for Out_i indicates that SNP i is a non-outlier within copy state C_i.

So the collection of all model parameters now becomes

\[ \theta = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, \Lambda, \pi_{out,1}, \pi_{out,2}, \pi_{out,3}) \]

One of the advantages of this model is that it allows us to use conjugate prior distributions which support closed-form derivations of posterior quantities of interest such as the Bayes factor and the posterior distribution over the space of possible DNA copy number sequences, which is the focus of Section 3.3.3. The use of the Bayes factor as an uncertainty measure for detected CNVs is discussed in Chapter 6.

Again, we emphasize that the model can easily be extended to include additional copy number states, such as the 0-copy state and the 4-copy state. While not a primary focus of this thesis, the problem of picking the number of copy states is discussed in Section 3.6.

### 3.3.2 Prior distributions

We use a fully Bayesian approach to estimate the model parameters and the copy number with the following prior distributions on the unknown parameters of the emission
Chapter 3. A Bayesian Discrete-Time HMM with Normal Emissions

\[ \mu_1 \sim N(m_1, \sigma_1^2/b_1), \]
\[ \mu_2 \sim N(m_2, \sigma_2^2/b_2), \]
\[ \mu_3 \sim N(m_3, \sigma_3^2/b_3), \]

and

\[ \frac{1}{\sigma_1^2} \sim \frac{1}{d_{0,1}s_{0,1}^2} \chi_{d_{0,1}}^2, \]
\[ \frac{1}{\sigma_2^2} \sim \frac{1}{d_{0,2}s_{0,2}^2} \chi_{d_{0,2}}^2, \]
\[ \frac{1}{\sigma_3^2} \sim \frac{1}{d_{0,3}s_{0,3}^2} \chi_{d_{0,3}}^2. \]

This is the scaled Inv-\(\chi^2\) specification for the measurement variance, \(\sigma_c^2\); see Gelman et al. (2003) for a definition. Here, \(d_{0,c}\) are the degrees of freedom for the \(\chi^2\)-distribution and \(s_{0,c}^2\) is the expected variance for the \(c\)-copy state. Notice that the prior variance of the class means is tied to the measurement variance of the data, \(\sigma_c^2\), where the dependence is scaled by the hyperparameter \(b_c\); thus, \(b_c\) can be interpreted as the number of prior observations for the class mean and we treat it as a hyperparameter. Furthermore, although their values are unknown, we do know that the class means are ordered such that \(\mu_1 < \mu_2 < \mu_3\), and this constraint is imposed by constraining the prior parameter space with the constraints \(1(\mu_1 < \mu_2)\) and \(1(\mu_3 > \mu_2)\). This is similar, but less stringent, than the constraints that were used by Guha et al. (2008), who further constrained the copy-neutral state (i.e. 2-copy state) to be within the interval from \(-\epsilon\) to \(\epsilon\), and losses and gain to be below and above \(-\epsilon\) and \(\epsilon\), respectively, where \(\epsilon\) is a user-specified constant.

We also use conjugate prior distributions for the parameters of the transition probability matrix,

\[ \Lambda \sim \text{ProductDirichlet} (\mathbf{A} = (a_1, a_2, a_3)), \]
where \( \mathbf{a}_j = (a_{j1}, a_{j2}, a_{j3}) \) for \( j = 1, 2, 3 \).

For the outlier probabilities we use the priors:

\[
\begin{align*}
\pi_{\text{out},1} &\sim \text{Beta}(1, 20), \\
\pi_{\text{out},2} &\sim \text{Beta}(1, 20), \\
\pi_{\text{out},3} &\sim \text{Beta}(1, 20),
\end{align*}
\]

which favours low outlier probabilities while still allowing the possibility of higher probabilities, if the data supports it.

This leaves us with a set of hyperparameters,

\[
\phi = (m_1, m_2, m_3, d_{0,1}, d_{0,2}, d_{0,3}, s_{0,1}^2, s_{0,2}^2, s_{0,3}^2, \mathbf{A}, b),
\]

to specify. The Dirichlet prior parameters \( \mathbf{A} \) can be interpreted as pseudocounts of the number of transitions between states. For example, \( a_{12} \) is the prior number of observations for transitions between the 1-copy state and the 2-copy state, and \( a_{13} \) is the prior number of observations for transitions between the 1-copy state and the 3-copy state. This can be seen from the conditional posterior distribution

\[
P(\Lambda|Y, C, \phi) \propto \lambda_1^{N_1+a_{11}-1} \lambda_2^{N_2+a_{12}-1} \lambda_3^{N_3+a_{13}-1} \tag{3.3}
\]

Since it is rare for a duplication to follow a deletion, we expect the probability \( \lambda_{12} \) to be close to 1, however, this may not be the case for all samples and we typically use a prior with \( \mathbf{a}_j = (1, 1, 1) \) for \( j = 1, 2, 3 \). When exemplifying the method on real data, the results from the analysis of the data with known DNA copy number, provided in Section 3.2.4, are used to specify the hyperparameters of the emission distribution.

Details about the sampling method used for this model are given in Section 3.4.3.

### 3.3.3 Model behaviour

To better understand the HMM the probability of the data conditional on the copy number sequence but marginal over the model parameters \( \theta \) can be examined. The
marginal distribution of the data can be obtained by integration:

\[
P(Y|C) = P(Y,C)/P(C) \\
= \int P(Y|C, \theta)P(C|\theta)P(\theta) \, d\theta/ \int P(C|\Lambda)P(\Lambda) \, d\Lambda \\
= \int_{\Lambda} \int_{\sigma^2} \prod_{j=1}^{M} f(y_j; c_j) \left( \prod_{k=1}^{3} N(\mu_k; m_k, \sigma_k^2/b_k) \right) \text{Inv}-\chi^2(\sigma_k^2, s_{0,k}, d_{o,k}) \times \\
x p(C|\Lambda) \text{ProdDir}(\Lambda) \, d\mu \, d\sigma^2 \, d\Lambda/ \int_{\Lambda} p(C|\Lambda) \text{ProdDir}(\Lambda) \, d\Lambda.
\]

In this treatment, for simplicity, the outlier variables are ignored. (Integration for the model with outliers is not feasible since the number of terms in the inner integrand grows exponentially with the class size.) To simplify the integration the terms in the integrand are rearranged so that the denominator and outer integrand end up cancelling out, leaving

\[
\int_{\sigma^2} \left\{ \int_{\mu} \prod_{j=1}^{M} f(y_j; c_j) \left( \prod_{k=1}^{3} N(\mu_k; m_k, \sigma_k^2/b_k) \right) \, d\mu \right\} \prod_{k=1}^{3} \text{Inv}-\chi^2(\sigma_k^2, s_{0,k}, d_{o,k}) \, d\sigma^2.
\]

Then, the inner integrand is

\[
\left( \prod_{j:C_j=1} f(y_j; c_j = 1)N(\mu_1; m_1, \sigma_1^2/b_1) \right) \left( \prod_{j:C_j=2} f(y_j; c_j = 2)N(\mu_2; m_2, \sigma_2^2/b_2) \right) \times \\
x \left( \prod_{j:C_j=3} f(y_j; c_j = 3)N(\mu_3; m_3, \sigma_3^2/b_3) \right)
\]

so each class can be integrated separately. For each class the integral is a function of the posterior Normal given \( \sigma_k^2 \), that is, \( \mu_k|\sigma_k^2 \sim N((b_km_k + \sum_{j:C_j=k} Y_j)/(b_k + n_k), (\sigma_k^2)/(b_k + n_k)) \), where \( n_k = \sum_{i=1}^{M} 1(C_i = k) \). Therefore, the inner integrand evaluates to

\[
\prod_{k=1}^{3} \left( \frac{1}{\sqrt{2\pi}} \right)^{n_k} \frac{\sqrt{b_k}}{\sqrt{n_k + b_k}} (\sigma_k^2)^{-(n_k/2)} \exp \left\{ \frac{1}{\sigma_k^2} \left( s_k^2 + \frac{n_kb_k(\bar{y}_k - m_k)^2}{n_k + b_k} \right) \right\}.
\]

Then, the middle integrand, which is with respect to the variance parameters \( \sigma^2 \), becomes

\[
\prod_{k=1}^{3} \left( \frac{1}{\sqrt{2\pi}} \right)^{n_k} \frac{\sqrt{b_k}}{\sqrt{n_k + b_k}} (\sigma_k^2)^{-(n_k/2)} \exp \left\{ \frac{1}{\sigma_k^2} \left( s_k^2 + \frac{n_kb_k(\bar{y}_k - m_k)^2}{n_k + b_k} \right) \right\} \text{Inv}-\chi^2(\sigma_k^2; s_{0,k}, d_{o,k}),
\]
which is a function of the posterior scaled Inv-\(\chi^2\) distribution, Inv-\(\chi^2(\sigma_k^2; d_{0,k}s_{0,k}^2 + s_k^2 + \frac{b_n}{b_n + n_k}(\bar{y}_k - m_k)^2, n_k + d_{0,k})\). Thus, the integral evaluates to

\[
\prod_{k=1}^{3} \left( \frac{1}{\sqrt{2\pi}} \right)^{n_k} \left( \frac{\sqrt{b_k}}{\sqrt{n_k + b_k}} \right)^{d_{0,k}/2} \left( \frac{d_{0,k}/2}{\Gamma(d_{0,k}/2)} \right)^{s_{0,k}} \Gamma\left( \frac{n_k + d_{0,k}}{2} \right) \Gamma\left( \frac{n_k + d_{0,k}}{2} \right) \times \left[ \frac{1}{n_k + d_{0,k}} \left( d_{0,k}s_{0,k}^2 + s_k^2 + \frac{n_kb_k(\bar{y}_k - m_k)^2}{n_k + b_k} \right) \right]^{-\frac{(n_k + d_{0,k})}{2}}.
\]

So

\[
P(Y|C) = \prod_{k=1}^{3} \pi^{-n_k/2} \frac{\sqrt{b_k}}{\sqrt{n_k + b_k}} \left( d_{0,k}s_{0,k}^2 \right)^{d_{0,k}/2} \Gamma\left( \frac{n_k + d_{0,k}}{2} \right) \Gamma\left( \frac{d_{0,k}/2}{2} \right) \left[ d_{0,k}s_{0,k}^2 + s_k^2 + \frac{n_kb_k(\bar{y}_k - m_k)^2}{n_k + b_k} \right]^{-\frac{(n_k + d_{0,k})}{2}},
\]

Notice that \(P(Y|C)\) no longer depends on the model parameters but it does depend on the hyperparameters \(\phi\). Here, \(\Gamma(\cdot)\) is the Gamma function, \(\Gamma(x) = (x - 1)!\) for integer \(x\); \(n_j\) is the number of SNPs in the \(j\)-copy state; \(s_j^2\) is the within-class sample variance; and \(\bar{y}_j\) is the average of the raw CNs that belong to the \(j\)-copy state, the within-class sample mean.

So, what does this mean? Well, the conditional distribution of the copy number sequence \(C\) given the data is proportional to its prior times \(P(Y|C)\). Therefore, this tells us that the model attributes higher posterior probability to copy number sequences for which the within-class sample means are close to the corresponding prior class means, that is, \(\bar{y}_j\) is close to \(m_j\) for \(j = 1, 2, 3\). Also, the model attributes higher posterior probability to sequences that have small within-class sample variances, that is, small values of \(s_j^2\) for \(j = 1, 2, 3\). This provides theoretical justification for using a HMM to model DNA copy number data as well as other types of sequential data.

The probability of the data conditional on the copy number, \(P(Y|C)\), can also be used to choose the number of copy states. A simple procedure for doing so is proposed in Section 3.6.
3.4 Parameter estimation

So far in this chapter a Bayesian HMM was motivated and explored from a theoretical perspective for the copy number analysis of SNP array data. This section details the MCMC algorithm used to estimate the Bayesian HMM, including the Gibbs samplers for the model parameters, $\theta$, and a backward sampling algorithm for the DNA copy number $C$.

3.4.1 A Markov chain Monte Carlo algorithm

The posterior distribution of $(C, \theta)$, implicitly conditional on the fixed hyperparameters $\phi$, can be written as

$$P(C, \theta|Y) \propto P(Y|C, \theta)P(C|\theta)P(\theta) = \pi_{c_1} f(Y_1|c_1) \prod_{i=2}^{M} \left\{ f(y_i|c_i)\lambda_{c_{i-1},c_i} \right\} P(\theta)$$

To generate samples from this posterior distribution an MCMC algorithm can be used. Starting with initial values of the unknown model parameters $\theta^{(0)}$. Then for $t = 1, \ldots$, iteratively sample from the two conditional distributions:

1. Sample $C^{(t)} \sim P(C|Y, \theta^{(t-1)})$,

2. Sample $\theta^{(t)} \sim P(\theta|Y, C^{(t)})$.

Given a sufficient number of iterations, the samples collected by this algorithm will converge to samples from the desired posterior distribution.

We recognize that interest primarily lies in samples from $p(C|Y)$ which indicate the most likely reconstruction of $C$. Thus, another route is to sample directly from $P(C|Y)$, for example, by using equation 3.4 (assuming a uniform prior on $C$) and thus bypassing the need to sample the model parameters $\theta$. However, sampling from $P(C|Y)$ is more difficult because integrating over the model parameters destroys the Markov structure of
the model, and closed form solutions of $p(C|Y)$ are not available for more complicated Bayesian Markov models. Thus it is preferable to iteratively sample from the conditional distributions, $P(C|Y, \theta^{(t-1)})$ and $P(\theta|Y, C^{(t)})$, to arrive at $p(C|Y)$ via Monte Carlo integration. The following sections describe the specific approach used to sample $\theta$ and $C$, and to reconstruct $C$.

### 3.4.2 Sampling copy number sequences

The standard approach to obtain samples from $P(C|Y, \theta)$ is the direct Gibbs sampling algorithm (Robert et al. 1993). This is based on the equation

$$P(C_t = c|c_{-t}, Y, \theta) \propto \lambda_{c_{t-1},c} f(y_t|c) \lambda_{c,c_{t+1}}$$  \hspace{1cm} (3.5)$$

for $t = 1, \ldots, M$ and where $c_{-t} = (c_1, \ldots, c_{t-1}, c_{t+1}, \ldots, c_M)$. However, as pointed out by Scott (2002), direct Gibbs is not the best algorithm in terms of the mixing of the Markov chain. Alternatively, a stochastic version of the forward-backward (FB) algorithm can be used. This approach, although more complicated than Gibbs sampling, is more efficient. It is discussed in detail in both Scott (2002) and Churchill & Lazareva (1999). To understand how stochastic FB works, we reexpress the conditional posterior of $C$ as

$$P(C|Y, \theta) = P(C_M|Y, \theta) P(C_{M-1}|C_M, Y, \theta) \cdots P(C_1|C_2, \ldots C_M, Y, \theta)$$  \hspace{1cm} (3.6)$$

$$= P(C_M|Y, \theta) \prod_{t=1}^{M-1} P(C_{M-t}|C_{M-t+1:M}, \theta, Y),$$
where \( C_{[M-t+1:M]} = (C_{M-t+1}, \ldots, C_M) \). Then, the backward sampling algorithm is based on the equation

\[
P(C_{M-t} = j | C_{[M-t+1:M]}, Y, \theta) \propto P(C_{M-t} = j, C_{[M-t+1:M]}, Y | \theta) \\
= P(C_{M-t} = j, Y_{[1:M-t]} | C_{M-t+1}, \theta) \\
\times P(C_{[M-t+2:M]}, Y_{[M-t+1:M]} | C_{M-t+1}, \theta) P(C_{M-t+1} | \theta) \\
\propto P(C_{M-t} = j, C_{M-t+1} | Y_{[1:M-t]}, \theta) \\
= P(C_{M-t} = j | Y_{[1:M-t]}, \theta) (C_{M-t+1} | C_{M-t} = j) \\
= \alpha_{M-t}(j) \lambda_{j,C_{M-t+1}}.
\]

Specifically, given the current drawn values of \( \theta \), denoted by \( \theta^{(t)} \), update \( C \) using the following backward sampling algorithm.

1. For \( j = 1, 2, 3 \), initialize the forward probabilities: \( \alpha_1(j) = \pi_j f_j(Y_1) \).

2. For \( j = 1, 2, 3 \), \( t = 2, \ldots, M \) and given \( \alpha_1(j) \), compute the remaining forward probabilities using the recursive equation \( \alpha_t(j) = f_j(Y_t) \sum_{i=1}^{3} \alpha_{t-1}(i) \lambda_{ij} \).

3. Sample \( C_M \) from Multinomial(\( B_M \)), where \( B_M = (\frac{\alpha_M(1)}{\sum_k \alpha_M(k)}, \frac{\alpha_M(2)}{\sum_k \alpha_M(k)}, \frac{\alpha_M(3)}{\sum_k \alpha_M(k)}) \).

4. Sample \( C_{t-1} \) recursively for \( t = M, M - 1, \ldots, 2 \) from

\[
C_{t-1} \sim \text{Multinomial}(B_{t-1}(C_t)),
\]

where \( B_{t-1}(C_t) = (\frac{\alpha_{t-1}(1)}{\sum_k \alpha_{t-1}(k)} \lambda_{1C_t}, \frac{\alpha_{t-1}(2)}{\sum_k \alpha_{t-1}(k)} \lambda_{2C_t}, \frac{\alpha_{t-1}(3)}{\sum_k \alpha_{t-1}(k)} \lambda_{3C_t}) \).

This algorithm is similar in structure to the Viterbi algorithm, except that now the backward pass of the algorithm is stochastic so that in repeated iteration it will explore more of the space of possible copy number sequences. This provides a basis for developing measures of uncertainty. Chapter 6 proposes a few such measures within this framework.
### 3.4.3 Sampling the model parameters

The conditional posterior distribution of a particular parameter given the most recently drawn copy number sequence, denoted $C^{(t)}$, the data $Y$ and other model parameters, is proportional to the product of its prior distribution and the likelihood. For the Bayesian HMM all parameters have conjugate prior distributions so we can utilize Gibbs sampling. Our algorithm iteratively samples new states for each quantity as follows:

1. For $c \in \{1, 2, 3\}$, given the data, other model parameters, $Out^{(t)}$ and $C^{(t)}$, sample $\mu_c$ from

$$
\mu_c \sim N\left(\frac{b_c m_c + \frac{n_c^{(t)}}{\sigma_c^2} \bar{y}_c^{(t)}}{b_c \sigma_c^2 + \frac{n_c^{(t)}}{\sigma_c^2}}, \left(\frac{b_c}{\sigma_c^2} + \frac{n_c^{(t)}}{\sigma_c^2}\right)^{-1}\right),
$$

where $n_c^{(t)} = \sum_{i=1}^{M} 1(C_i^{(t)} = c, Out_i^{(t)} = 0)$, the number of SNPs having $c$ copies of DNA for MCMC scan $t$; and $\bar{y}_c^{(t)} = \left(\sum_{i: C_i^{(t)} = c, Out_i^{(t)} = 0} y_i^{(t)}\right) / n_c^{(t)}$, the average of the raw CNs for SNPs with $c$ copies for MCMC scan $t$.

2. For $c \in \{1, 2, 3\}$, given the data, other model parameters, $Out^{(t)}$ and $C^{(t)}$, sample $\sigma_c^2$ from

$$
\sigma_c^2 \sim \left(d_{0,c} s_0^2 + s_c^2 + \frac{b_c n_c^{(t)}}{b_c + n_c^{(t)}} \left(\bar{y}_c^{(t)} - m_c\right)^2\right) / \chi^2_{\nu_c^{(t)} + d_{0,c}},
$$

where $s_c^2$ is $(n_c^{(t)} - 1)$ times the sample variance of state $c$.

3. For $i = 1, \ldots, M$, given the data, other model parameters and $C^{(t)}$, sample $Out_i$ from

$$
Out_i^{(t)} \sim \text{Bernoulli}\left(\frac{\pi_{out,c_i^{(t)}} / r}{\pi_{out,c_i^{(t)}} / r + (1 - \pi_{out,c_i^{(t)}}) N\left(y_i; \mu_{c_i^{(t)}}, \sigma_{c_i^{(t)}}\right)}\right),
$$

where $r = Y_{\text{max}} - Y_{\text{min}}$. 


4. Sample the transition probabilities from

\[
\lambda_1 \sim \text{Dir}(a_{11} + n_{11}^{(t)}, a_{12} + n_{12}^{(t)}, a_{13} + n_{13}^{(t)}),
\]

\[
\lambda_2 \sim \text{Dir}(a_{21} + n_{21}^{(t)}, a_{22} + n_{22}^{(t)}, a_{23} + n_{23}^{(t)}),
\]

\[
\lambda_3 \sim \text{Dir}(a_{31} + n_{31}^{(t)}, a_{32} + n_{32}^{(t)}, a_{33} + n_{33}^{(t)}),
\]

where \( n_{ij}^{(t)} = \sum_{m=2}^{M} 1(C_{m-1}^{(t)} = i, C_{m}^{(t)} = j) \) for \( i, j = 1, 2, 3 \), the number of transitions from state \( i \) to state \( j \) for MCMC scan \( t \).

5. Sample the outlier probabilities from

\[
\pi_{\text{out},1} \sim \text{Beta}(1 + n_{1,\text{out}}^{(t)}, 20 + n_{1,\text{nout}}^{(t)}),
\]

\[
\pi_{\text{out},2} \sim \text{Beta}(1 + n_{2,\text{out}}^{(t)}, 20 + n_{2,\text{nout}}^{(t)}),
\]

\[
\pi_{\text{out},3} \sim \text{Beta}(1 + n_{3,\text{out}}^{(t)}, 20 + n_{3,\text{nout}}^{(t)}),
\]

where \( n_{k,\text{out}}^{(t)} = \sum_{m=1}^{M} 1(C_{m}^{(t)} = k, \text{Out}_{m}^{(t)} = 1) \) and \( n_{k,\text{nout}}^{(t)} = \sum_{m=1}^{M} 1(C_{m}^{(t)} = k, \text{Out}_{m}^{(t)} = 0) \) for \( k = 1, 2, 3 \), the number of outlier and non-outlier SNPs, respectively, for copy state \( k \) for MCMC scan \( t \).

After running the MCMC algorithm, we have a set of sampled copy number sequences \( \{C^{(t)}\}_{1 \leq t \leq T} \), where \( C^{(t)} = (C_{1}^{(t)}, \ldots, C_{M}^{(t)}) \) and \( T \) is the number of MCMC scans, a set of sampled outlier states \( \{\text{Out}^{(t)}\}_{1 \leq t \leq T} \), where \( \text{Out}^{(t)} = (\text{Out}_{1}^{(t)}, \ldots, \text{Out}_{M}^{(t)}) \), as well as a set of sampled model parameters \( \{\theta^{(t)}\}_{1 \leq t \leq T} \), where

\[
\theta^{(t)} = (\mu_{1}^{(t)}, \mu_{2}^{(t)}, \mu_{3}^{(t)}, \sigma_{1}^{2(t)}, \sigma_{2}^{2(t)}, \sigma_{3}^{2(t)}, \Lambda^{(t)}, \pi_{\text{out},1}^{(t)}, \pi_{\text{out},2}^{(t)}, \pi_{\text{out},3}^{(t)}).
\]

The next section discusses approaches for summarizing \( \{C^{(t)}\}_{1 \leq t \leq T} \).
3.5 Bayesian copy number sequence estimation

The sampled copy number sequences \( \{C^{(t)}\}_{1 \leq t \leq T} \) typically will be stored in a large matrix of dimension \( T \times M \). Since each \( C_j \) is discrete, the distribution from which these sequences arise, \( P(C|Y) \), has support on a finite space of sequences. However, this space of all possible sequences is high-dimensional and very difficult to summarize. Based on \( \{C^{(t)}\}_{1 \leq t \leq T} \), the DNA copy number can be reconstructed by using either the marginal (SNP-specific) probabilities or the joint distribution. This section compares techniques for both approaches.

### 3.5.1 Marginal distribution of \( C_j \)

The most straightforward way of summarizing \( P(C|Y) \) is by reconstructing the sequence based on the marginal (SNP-specific) probabilities: \( P(C_j|Y) \). The marginal probabilities are computed from the MCMC scans by:

\[
\hat{P}(C_j = c|Y) = 1/T \sum_{t=1}^{T} 1(C_j^{(t)} = c)
\]

for \( j = 1, \ldots, M \).

An improved estimate can be obtained by summing the probabilities rather than the events that were generated by these probabilities. Namely,

\[
\hat{P}(C_j = c|Y) = 1/T \sum_{t=1}^{T} P(C_j^{(t)} = c|\theta^{(t)}, Y)
\]

for \( j = 1, \ldots, M \). The probabilities in the summation are computed using the forward-backward algorithm at each iteration, i.e.,

\[
P(C_j^{(t)} = c|\theta^{(t)}, Y) = \frac{\alpha_j(c; \theta^{(t)}) \beta_j(c; \theta^{(t)})}{P(Y)},
\]

where \( P(Y) = \sum_{k=1}^{3} \alpha_M(k; \theta^{(t)}) \) and note the dependence of the forward and backward variables on the model parameters. Computing the \( \alpha \) and \( \beta \) variables entails running a forward and backward pass at each MCMC iteration, which the algorithm is already
doing, thus, using these probabilities does not increase the computational demand. See Chapter 2 for more details about the forward and backward variables. This well known technique is called Rao-Blackwellization. It typically results in estimates with lower variability.

Due to the segmental structure of copy number data, however, we are more interested in the overall configuration of $C$ than the value of, for example, an individual SNP $j$. Thus, theoretically, it is preferable to focus on the joint distribution of the hidden states when reconstructing the DNA copy number.

### 3.5.2 Marginal distribution of $C$

Recall that the Viterbi algorithm maximizes the conditional posterior distribution $p(C|Y, \theta)$. As pointed out in Chapter 2, this doesn’t fully account for the uncertainty in the model parameters. A better approach involves examining the probability of the DNA copy number sequence conditional on the data but marginal over the values of the model parameters. Namely, the marginal posterior distribution of $C$,

$$p(C|Y) = \int p(C|Y, \theta) p(\theta) \, d\theta. \quad (3.10)$$

This averages over the model parameters, that is, it takes into account the uncertainty of the model parameters. As pointed out by Scott (2002), it is more difficult to maximize $p(C|Y)$ than $p(C|Y, \theta)$ because averaging over $\theta$ destroys the Markov structure of the model. Consequently, there is no Viterbi-style algorithm for maximizing the marginal distribution. However, Markov chain Monte Carlo can be used to examine $p(C|Y)$, for example, using Monte Carlo,

$$p(C|Y) = \int p(C|y, \theta) p(\theta) \, d\theta \approx \frac{1}{T} \sum_{t=1}^{T} p(C|Y, \theta^{(t)}), \quad (3.11)$$

where $\{\theta^{(t)}\}_{1 \leq t \leq T}$ is a sequence of sampled parameters from a Markov chain. Still though, the function in 3.11 is difficult to maximize.
The simplest method is to compute the frequencies of each of the sampled copy number sequences \( C(t) \), where \( C(t) = (C_1^{(t)}, \ldots, C_M^{(t)}) \) for \( 1 \leq t \leq T \), and then the sampled sequences can be ranked according to their frequencies. The problem with this approach is that for long sequences any single sampled sequence will not reappear multiple times. So, the set of distinct \( C(t) \) will be too large.

An alternative approach designed specifically for long sequences, which was suggested by Scott (2002) in the context of HMMs, is to examine the set \( \{ \hat{C}^{(1)}, \ldots, \hat{C}^{(T)} \} \), where \( \hat{C}^{(t)} = \hat{C}(\theta^{(t)}) = \arg \max C \ p(C|Y, \theta^{(t)}) \), instead of \( \{ C^{(1)}, \ldots, C^{(T)} \} \). (That is, separately for each \( t \), the Viterbi algorithm is used to maximize \( p(C|Y, \theta^{(t)}) \).) Meaning that \( \hat{C}^{(t)} \) is the most likely copy number sequence corresponding to \( \theta^{(t)} \). So the most frequent sequence among this set of sequences can be described as the 'most often most likely'.

The reason for doing this is that the set of distinct \( \hat{C}^{(t)} \) will be smaller than the set of distinct \( C^{(t)} \). This allows us to compute observed frequencies of most likely sequences \( \hat{C}^{(t)} \) with which we can approximately maximize \( p(C|Y) \). However, for SNP genotyping arrays the number of SNPs on a chromosome is often over 10,000 so this approach is computationally prohibitive and, based on our experience, not effective for such long sequences.

Since we cannot reliably estimate MCMC frequencies of sampled copy number sequences, it may be useful to compute joint probabilities based on the marginal posterior \( P(C|Y) \). The marginal posterior distribution of the Bayesian HMM is proportional to the likelihood, which is given in equation 3.4, times the prior. Therefore

\[
P(C|Y) \propto \prod_{k=1}^{3} \left( \frac{\sqrt{b_k}}{n_k + b_k} (d_{0,k} s_{0,k}^2)^{d_{0,k}/2} \frac{\Gamma(n_k + d_{0,k})}{\Gamma(d_{k}/2)} \frac{B(a_k + n_k)}{B(a_k)} \right) \\
\times \left[ d_{0,k} s_{0,k}^2 + s_k^2 + \frac{n_k b_k (\bar{y}_k - m_k)^2}{n_k + b_k} \right],
\]

where the factor \( \pi^{n_k} \) was dropped and \( B(a_j) \) is the multinomial Beta function, \( \frac{\Gamma(a_{11})\Gamma(a_{12})\Gamma(a_{13})}{\Gamma(a_{11}+a_{12}+a_{13})} \).

The relative probability of a copy number sequence can be computed by evaluating equation 3.12 for the sequence, then dividing by the total probability which is computed by
summing equation 3.12 over all of the sampled sequences. That is, the relative probability of sequence $C^{(t)}$ is

$$P(\hat{C}^{(t)}) = \frac{1}{T} \frac{P(C^{(t)}|Y)}{\sum_{t=1}^{T} P(C^{(t)}|Y)},$$

(3.13)

where $P(C^{(t)}|Y)$ is computed up to a constant using equation 3.12. These relative probability estimates can be used as a measure of the strength of evidence, or the confidence level, of each sampled sequence. Furthermore, the relative probabilities can be used to rank the sampled sequences, for example, we can save the top $K$ most probable sequences for further examination. We denote the top sequences by $\{\hat{C}^{(k)}\}_{1 \leq k \leq K}$ where the sequences are ranked such that $P(\hat{C}^{(1)}) > P(\hat{C}^{(2)}) > \ldots > P(\hat{C}^{(K)})$. Based on simulation studies, this approach typically results in very similar reconstructions to the one based on the marginal probabilities, so we prefer to restore the sequence by using the marginal probabilities due to its computational simplicity.

### 3.6 Choosing the number of states

For data analysis the first task is to pick the number of copy states in the model. For DNA from normal cell lines, the largest possible number of states considered in the model is six, corresponding to the 0-, 1-, 2-, 3-, 4- and 5-copy states. Copy numbers beyond 5 are not only extremely rare, but are also indistinguishable from the 5-copy state for SNP genotyping array data. The marginal distribution of the data, $P(Y|C)$, can be used to help discriminate between models with a different number of copy number states. Redefine $\theta = (\theta_1, \ldots, \theta_6)$ where $\theta_k$ is the vector of model parameters corresponding to the model with $k$ copy states. For example, $\theta_1$ corresponds to the model with only the 2-copy state, whereas $\theta_6$ corresponds to the model with all 6 copy states. The following procedure can be used as a crude method for picking the number of copy states in the model:

1. For $k \in (1, 2, 3, 4, 5, 6)$,
(a) Estimate $C$ by maximizing $P(C|Y, \hat{\theta}_k)$, where $\hat{\theta}_k$ is chosen using the EM-algorithm, giving $\hat{C}_k$.

(b) Compute $P(Y|\hat{C}_k)$ using equation 3.4.

2. Let $\hat{K} = \arg \max_k P(Y|\hat{C}_k)$.

Alternatively, the distribution of values of $P(C|Y)$ from the MCMC algorithm can be examined for models with a different number of states, and comparisons between these models can reveal information about the true number of hidden states.

### 3.7 Implementation of the algorithm

The Bayesian approach is computationally intensive due to having to store thousands of long sequences of copy number data. All the preprocessing of the raw data (i.e. probe-level data) was done in R using the package `aroma.affymetrix` (see Chapter 1 for more details). However, to improve computational performance the entire MCMC algorithm was implemented in the C programming language. Due to the large amount of looping required by the algorithm, C is much faster than R is, and since there are tens of thousands of SNPs on a given chromosome, the computations must be done efficiently for the algorithm to be computationally feasible. Furthermore, it is necessary to bound the memory usage, so during runtime MCMC results are written to file, and the storage problem is reduced by only storing the breakpoints of the sequence. Code for the algorithm will be made available online.

### 3.8 Analysis of HapMap data

This section illustrates the use of the Bayesian HMM by analyzing Affymetrix SNP array 6.0 data from the HapMap study. The analysis of a single sample on the second arm of chromosome 10 is shown here. Since the Bayesian HMM was developed for finding rare
CNVs, or low-frequency CNPs, and high-frequency CNPs can be easily detected using methods that utilize multiple samples, regions harboring high-frequency CNPs, defined as those having a frequency greater than 0.9, were removed from the data. We used the map of copy of number variation delineated in McCarroll et al. (2008) to identify regions harboring high-frequency CNPs. Also, in computing the SNP-specific reference signals, any samples that were known to harbor CNPs were removed from the computation of the reference single so that the estimation of the reference signal would not be skewed by non-diploid observations.

As a representative analysis that shows the inadequacy of the current model, the HapMap sample NA18550 was analyzed by the model. This consisted of data from 25,793 SNPs on the second arm of chromosome 10 (between 40 Mb and 80 Mb in Figure 1.2). Other samples had similar results for this model. As reported by McCarroll et al. (2008), this particular sample contains two amplifications (three copies of DNA), 180 and 17 SNPs, respectively; four short hemizygous deletions (one copy of DNA), 6, 4, 3 and 7 SNPs, respectively; and two homozygous deletions, 27 and 3 SNPs, respectively. The approach used by McCarroll et al. (2008) to find these CNPs is described in Section 3.2.3. The raw data and known CNVs are shown in Figure 3.3, where the amplified SNPs are in green and the deleted region of SNPs are in red. It is quite possible that this sample contains additional rare CNVs that were not reported in McCarroll et al. (2008).

For the emission model parameters, the hyperparameters used are \( m = (-1.887, -0.629, 0, 0.432) \), \( b = (10, 10, 10, 10) \), \( s_2^2 = (0.2, 0.1, 0.05, 0.05) \), and \( d_0 = (10, 10, 10, 10) \). These values were specified according to the analysis of the titration and X chromosome data in Section 3.2.3. For the transition probability matrix, hyperparameters used are \( a_0 = (18, 1, 1, 1) \), \( a_1 = (1, 18, 1, 1) \) \( a_2 = (1, 1, 18, 1) \) \( a_3 = (1, 1, 1, 18) \), which is a weakly informative prior, favouring transition matrices that have a higher probability of remaining in the current state, while still allowing for the possibility of transition matrices with lower probability of remaining in the current state, if the data supports it. The results are based on 5,000
MCMC scans after a burnin period of 1,000 scans. The DNA copy number sequence was reconstructed as follow. First, the marginal (SNP-specific) probabilities are estimated across MCMC iterations. Then, each SNP is assigned a copy number value according the rule: \( \max_k P(C_i = k | Y) \). Finally, a CNV was defined as a sequence of SNPs with non-two copy values within the reconstruction. Tables 3.2 and 3.3 show the posterior summary of the model parameters and Figure 3.4 shows the resulting copy number reconstruction for the first 2000 SNPs analyzed. This subplot of SNPs was chosen as a representative segment to demonstrate the primary problem with the model: An enormous number of called 1-copy deletions within a segment that does not appear to harbor
any CNVs. The posterior of the BDHMM supports frequent transitions between the 1- and 2-copy states. Biologically, for a normal DNA sample, there should only be a small to moderate number of CNVs per chromosome. Thus, this result is biologically implausible. The reason for this unrealistic copy number reconstruction is that there is not much separation between the posterior distributions of the 1- and 2-copy states, so that the 1-copy state merges with the 2-copy state. This is readily seen in Table 3.3 which shows that the posterior mean values of the class mean for the 1- and 2-copy states are 0.02 and -0.02, respectively. Based on the analysis of the X chromosome data in Section 3.2.4, we expect the class mean the 1-copy state to be around -0.63, not near 0. The posteriors of the class means for the 0- and 3-copy states are more in line with biological expectations, however, the class variance for the 0-copy state is extremely large, with a posterior mean of 0.90. Furthermore, the high posterior mean value of 0.10 for the outlier rate of the 0-copy state suggests that the model for this state is inadequate, that is, the 0-copy state emission distribution does not match well the actual data.

3.9 Discussion of Results

Although the parameter estimates from BDHMM are statistically valid, they are not biologically plausible. This indicates that there is a problem with the model and, in general, an inadequacy with the use of the standard HMM for analysing Affymetrix SNP array data. Examination of the posterior distributions showed that there was very little separation between the posteriors of the 1- and 2-copy states, that is, the posterior distribution of the 1-copy state collapsed to form another 2-copy state in the sense that the resulting emission distributions for these two states overlapped. This suggests that the emission distribution for the 2-copy state is inadequate for these data and a more complicated model using substates within the 2-copy state is likely to be more effective. Chapter 5 explores the use of such extensions as well as other model extensions.
Figure 3.4: Subplot of 2000 SNPs. The top display shows the observed raw CNs for the first 2000 SNPs on the chromosome and the bottom display shows the estimated copy number of these SNPs based on the BDHMM.

The Bayesian HMM presented in this chapter is based on a discrete-time Markov process and thus it does not take into account the distance between adjacent SNPs. Taking into account the physical distance between adjacent SNP markers is desirable because it is likely that SNPs closer to one another are more likely to share the same DNA copy number (i.e. lie in the same segment) than SNPs that are farther apart. A natural framework for accounting for this distance is using a continuous-time Markov process. Chapter 4 will discuss the implementation and use of a novel Bayesian continuous-time HMM.
Table 3.2: Posterior summary of the parameters of the transition probability matrix for BDHMM.
<table>
<thead>
<tr>
<th></th>
<th>2.5%</th>
<th>25%</th>
<th>50%</th>
<th>Mean</th>
<th>75%</th>
<th>97.5%</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>-2.23</td>
<td>-1.99</td>
<td>-1.86</td>
<td>-1.86</td>
<td>-1.73</td>
<td>-1.46</td>
<td>0.20</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.02</td>
<td>-0.02</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>0.28</td>
<td>0.38</td>
<td>0.43</td>
<td>0.43</td>
<td>0.48</td>
<td>0.59</td>
<td>0.08</td>
</tr>
<tr>
<td>$\sigma^2_0$</td>
<td>0.45</td>
<td>0.67</td>
<td>0.84</td>
<td>0.90</td>
<td>1.06</td>
<td>1.63</td>
<td>0.31</td>
</tr>
<tr>
<td>$\sigma^2_1$</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td>0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>$\sigma^2_2$</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>$\sigma^2_3$</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>$\pi_{out,0}$</td>
<td>0.01</td>
<td>0.06</td>
<td>0.09</td>
<td>0.10</td>
<td>0.14</td>
<td>0.24</td>
<td>0.06</td>
</tr>
<tr>
<td>$\pi_{out,1}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>$\pi_{out,2}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\pi_{out,3}$</td>
<td>0.00</td>
<td>0.01</td>
<td>0.03</td>
<td>0.05</td>
<td>0.07</td>
<td>0.17</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 3.3: Posterior summary of the parameters of the emission distribution for BDHMM.
Chapter 4

A Continuous-time HMM

4.1 Introduction

The Bayesian discrete-time HMM was covered in detail in Chapter 3. The ability of this model to infer DNA copy number changes was assessed using chromosome 10 HapMap data. This chapter discusses the continuous-time HMM (CHMM) which accounts for the distance between adjacent SNPs in a model-based framework. CHMMs are defined in terms of instantaneous transition intensities rather than probabilities. CHMMs have not previously been proposed for analyzing Affymetrix SNP arrays, however, they have been previously used by Wang et al. (2009) for analyzing data from Illumina SNP arrays. The Bayesian CHMM (BCHMM) proposed in this chapter differs from the model of Wang et al. (2009) in a few important ways: (i) it is a more general and arguably more realistic continuous-time model, containing more parameters in the transition intensity matrix; (ii) the distance-$d$ transition probability matrix is derived using a different method and, as shown in Section 4.2.3, this results in a different and novel implementation of the CHMM; and (iii) it uses a Bayesian approach which accounts for full-parameter uncertainty and allows the incorporation of prior information in the form of mean CNV lengths.

The chapter begins with the most general parameterization of the transition intensity
matrix and a solution for deriving the corresponding distance-\(d\) transition probability matrix for this model is presented. A parsimonious version of the model which leads to an interesting, novel specification of the transition probability matrix is also discussed. The remainder of the chapter details an MCMC algorithm for parameter estimation of the BCHMMs and the HapMap data are analysed using both models.

### 4.2 Continuous-time HMMs

The actual CNV breakpoints may, and most likely will, occur between the observed SNP markers. The discrete-time model assumes that breakpoints happen at the observed SNP markers, whereas the continuous-time model admits that breakpoints may occur anywhere along the genome. Of course, there are only a discrete number of base pairs along the genome, so, strictly speaking, the occurrence of breakpoints along the genome is not continuous in space. But, due to the extremely long length of the genome – approximately 3 billion bps – the approximation error in treating the genome as continuous in space is negligible.

This last point also relates to the specification of the transition probability matrix: SNPs are not uniformly spread over the genome. Therefore, we should specify the transition probabilities as a function of the distance between SNPs. The continuous-time model accomplishes this in a model-based framework. That is, for the transition intensity matrix \(Q\), which defines a continuous-time model, there exists a corresponding distance-\(d\) transition probability matrix \(T(d)\). In Section 4.2.2 we discuss the problem of deriving \(T(d)\) for our model.

This section outlines a continuous-time HMM for the copy number analysis of a single SNP array whose raw CN measurements are denoted by the vector \(Y = (Y_1, \ldots, Y_M)\) where \(Y_i\) is the observed raw CN at SNP \(i\) with corresponding genomic location \(d_i\) in base pairs (bps). Following a description of the model for the underlying DNA copy
number values, the problem of deriving the transition probabilities corresponding to the copy number model is discussed, and then a parsimonious submodel that is more practical for samples with a large number of copy states is discussed.

Figure 4.1: Graphical representation of the continuous-time HMM.

4.2.1 Copy number model

We model $C$ as a continuous-time Markov process. The continuous-time Markov process is parameterized in terms of a $3 \times 3$ transition intensity matrix of copy number changes $Q = \{q_{ij}\}_{i=1,2,3; j=1,2,3}$, where, for $i \neq j$, $q_{ij} > 0$ and $q_{kk} = -\sum_{l \neq k} q_{kl}$, so that $\sum_j q_{ij} = 0$ and $q_{kk} \leq 0$. Unlike a discrete-time Markov process whose state transitions are defined in terms of transition probabilities, the continuous-time Markov process is defined in terms of its instantaneous transition intensities $q_{ij}$. The intensity $q_{ij}$ represents the instantaneous risk of moving from state $i$ to state $j$:

$$q_{ij} = \lim_{h \to 0} P(C(t + h) = j|C(t) = i)/h. \quad (4.1)$$

The complete specification of the transition intensity matrix is given by

$$Q = \begin{pmatrix}
-q_{12} - q_{13} & q_{12} & q_{13} \\
q_{21} & -q_{21} - q_{23} & q_{23} \\
q_{31} & q_{32} & -(q_{31} + q_{32})
\end{pmatrix}.$$
This model assumes that, for example, the rate of deletion in a normal (2-copy) region may be different than the rate of deletion in an amplified (3-copy) region. More specifically, this model allows for the possibility that the rate of deletions from the normal state is larger than the rate of deletions from the amplified state, that is, \( q_{21} > q_{31} \). This is plausible since deletions may be significantly less likely to follow amplifications.

The parameters of the Q matrix govern the occurrence of copy number changes along the chromosome. Perhaps a better way of understanding the evolution of the Markov chain is through the time it spends in a state or, in the context of genomics, the number of bases pairs before a copy number change occurs. (Note that the number of bps before a copy number change occurs simply corresponds to the length of the region, for example, the length of a copy number deletion.) Assuming that the chain begins in the diploid (2-copy) state, the chain stays in the diploid state for a length of time (or distance) that is exponentially distributed with rate parameter \( \nu_2 = q_{21} + q_{23} \). Thus, under this model, the expected length of a diploid region is \( 1/\nu_2 \). Once the stay in the diploid state is complete, the chain then moves to either the deleted (1-copy) state with probability \( q_{21}/(q_{21} + q_{23}) \) or the amplified (3-copy) state with probability \( q_{23}/(q_{21} + q_{23}) \). This process repeats itself, over and over. We denote the transition probabilities of the chain by \( p_{ij} = q_{ij}/(\sum_{k \neq j} q_{ik}) \) and the transition rates by \( \nu_j = \sum_{k \neq j} q_{jk} \) for \( j = 1, 2, 3 \). Once the process leaves state \( i \) it must proceed to a new state, so that \( p_{ii} = 0 \) for \( i = 1, 2, 3 \). Also, \( p_{13} = 1 - p_{12}, p_{23} = 1 - p_{21} \) and \( p_{32} = 1 - p_{12} \). As we will see later in this chapter, it is convenient to reparameterize the model in terms of the \( p \)- and \( \nu \)-parameters. Under this parameterization for the CHMM, the collection of all parameters is \( \theta = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, p_{12}, p_{21}, p_{31}, \nu_1, \nu_2, \nu_3) \) and we sometimes also distinguish between parameters governing the emission distribution, \( \theta_E = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2), \) and the parameters governing the Q-matrix, \( \theta_Q = (p_{12}, p_{21}, p_{31}, \nu_1, \nu_2, \nu_3) \).

This is the most general model in the sense that it places no restrictions on the transition rates, however, we may also be interested in comparing more parsimonious
versions of this model. We discuss such models in detail in Section 4.2.3.

4.2.2 Computing the transition probabilities

We define $T(\Delta)$ to be the transition probability matrix of transition probabilities over a distance $\Delta$. Specifically, the distance $\Delta$ transition probabilities are $T_{ij}(\Delta) = P(C(t + \Delta) = j|C(t) = i)$; that is, the probability that a process presently in state $i$ will be in state $j$, $\Delta$ base pairs later. It is important to compute these probabilities because we only observe data at a discrete number of locations along the genome, specifically, at the genomic positions $d_1, \ldots, d_M$. Therefore, the likelihood contribution of the data depends on these probabilities. To compute the transition probability matrix $T(\Delta)$ we must first derive its form, which depends on the specification of the transition intensity matrix $Q$. This is done by using a set of differential equations for the transition probabilities, known as the Kolmogorov equations. The following treatment follows standard theory that can be found in Ross (1993). There are two sets of differential equations. First, the Kolmogorov backward equations are given by

$$T'_{ij}(\Delta) = \sum_{k \neq i} q_{ik} T_{kj}(\Delta) - \nu_i T_{ij}(\Delta). \quad (4.2)$$

Second, the Kolmogorov forward equations are given by

$$T'_{ij}(\Delta) = \sum_{k \neq j} q_{kj} T_{ik}(\Delta) - \nu_j T_{ij}(\Delta). \quad (4.3)$$

It is informative to express these differential equations in matrix form. We define the matrices $T(\Delta)$ and $T'(\Delta)$ to be composed of the elements $T_{ij}(\Delta)$ and $T'_{ij}(\Delta)$, respectively. Then the backward equations are

$$T'(\Delta) = QT(\Delta) \quad (4.4)$$

and the forward equations are

$$T'(\Delta) = T(\Delta)Q. \quad (4.5)$$
It can be shown, see for example Ross (1993), that the solution to these equations are

\[ T(\Delta) = T(0)e^{Q\Delta} \]  \hspace{1cm} (4.6)

where

\[ e^{Q\Delta} = \inf_{n=0} \sum Q^n \frac{\Delta^n}{n!} \]  \hspace{1cm} (4.7)

Of course, \( T(0) \) is the identity matrix, so

\[ T(\Delta) = e^{Q\Delta}. \]  \hspace{1cm} (4.8)

For the transition intensity matrix \( Q \) with 6 parameters defined in Section 4.2.1, simple analytic expressions for the transition probability matrix do not exist. Furthermore, computing equation (4.7) is computationally expensive. Instead, to derive a closed form for \( T(\Delta) \), we assume that there is at most one state transition between adjacent SNPs. This is a reasonable assumption as we do not expect more than one transition between adjacent SNPs due to the close proximity of SNPs on the arrays. With this assumption, if we let \( L_i \) be a random variable representing the length of the stay in state \( i \), then, assuming that we are in state \( i \) at SNP \( j - 1 \), the probability that we are still in state \( i \) at SNP \( j \) is

\[ T_{ii}(d_j - d_{j-1}) = P(C_j = i|C_{j-1} = i, d_j, d_{j-1}) \]
\[ = P(L_i > d_j - d_{j-1}) \]
\[ = e^{-\nu_i(d_j - d_{j-1})} \]
\[ = e^{-\nu_i\Delta_j} \]

where \( \Delta_j = d_j - d_{j-1} \), the distance between SNP \( j \) and \( j - 1 \). This holds because we know that \( L_i \) is exponentially distributed with parameter \( \nu_i \). Similarly, the probability
that we transition to state $k$ from state $i$ between SNPs $j$ and $j - 1$ is

$$T_{ik}(\Delta_j) = P(C_j = k | C_{j-1} = i)$$

$$= P(C_j = k | C_{j-1} = i) P(L_i \leq \Delta_j)$$

$$= (q_{ik} / \nu_i)(1 - e^{-\nu_i \Delta_j})$$

$$= p_{ik}(1 - e^{-\nu_i \Delta_j}),$$

where $p_{ik} = q_{ik} / \nu_i$, the probability of transitioning from state $i$ to state $k$. Therefore, the distance-$\Delta$ transition probability matrix is given by

$$T(\Delta) = \begin{pmatrix}
    e^{-\nu_1 \Delta} & p_{12}(1 - e^{-\nu_1 \Delta}) & (1 - p_{12})(1 - e^{-\nu_1 \Delta}) \\
    p_{21}(1 - e^{-\nu_2 \Delta}) & e^{-\nu_2 \Delta} & (1 - p_{21})(1 - e^{-\nu_2 \Delta}) \\
    p_{31}(1 - e^{-\nu_3 \Delta}) & (1 - p_{31})(1 - e^{-\nu_3 \Delta}) & e^{-\nu_3 \Delta}
\end{pmatrix}.$$  

### 4.2.3 Parsimonious submodel

Special cases of the model presented in Section 4.2.1 that correspond to more parsimonious parameterizations are sometimes of interest. Indeed, when there are many copy states for a given sample a more parsimonious model may be desired. For example, we may assume that the rate of transition into a specific state is the same regardless of the originating states. Under the general model, this corresponds to applying the constraints: $q_{21} = q_{31}$, $q_{12} = q_{32}$, and $q_{23} = q_{13}$. This results in a simplified Q-matrix with only three parameters instead of six:

$$Q = \begin{pmatrix}
    -(q_2 + q_3) & q_2 & q_3 \\
    q_1 & -(q_1 + q_3) & q_3 \\
    q_1 & q_2 & -(q_1 + q_2)
\end{pmatrix},$$

where $q_1$ denotes the rate of deletions, $q_3$ denotes the rate of amplifications, and $q_2$ denotes the rate of the diploid state. The difference in the number of parameters between the two models becomes more significant as the number of copy states in the model increases. This is due to the fact that the number of parameters in the general model grows quadratically.
with the number of states, whereas the number of parameters in this parsimonious model grows linearly with the number of states. For example, when there are four copy states, the difference is four parameters versus twelve parameters.

For this model the expected length of deletions is \(1/(q_2 + q_3) = 1/\nu_1\), the expected length of amplifications is \(1/(q_1 + q_2) = 1/\nu_3\), and the expected length of a diploid stretch is \(1/(q_1 + q_3) = 1/\nu_2\). Using simple algebra it can be shown that each of the transition intensities are functions of the \(\nu\)-parameters. That is,

\[
q_1 = (\nu_2 + \nu_3 - \nu_1)/2 \\
q_2 = (\nu_1 + \nu_3 - \nu_2)/2 \\
q_3 = (\nu_1 + \nu_2 - \nu_3)/2.
\]

Now we can derive the corresponding distance-\(d\) transition probability matrix using the same approach as in Section 4.2.2. The diagonal elements remain the same, while for the off-diagonal elements, the probability that we transition to copy state 2 from copy state 1 between SNPs \(j\) and \(j-1\) is

\[
T_{12}(\Delta_j) = P(C_j = 2|C_{j-1} = 1) \\
= P(C_j = 2|C_{j-1} = 1)P(L_1 \leq \Delta_j) \\
= (q_2/\nu_1)(1 - e^{-\nu_1 \Delta_j}) \\
= \frac{\nu_1 + \nu_3 - \nu_2}{2\nu_1}(1 - e^{-\nu_1 \Delta_j}).
\]

Using analogous arguments for transitions between the other states, the resulting matrix is

\[
T(d) = \begin{pmatrix}
e^{-\nu_1 d} & \frac{\nu_1 + \nu_3 - \nu_2}{2\nu_1}(1 - e^{-\nu_1 d}) & \frac{\nu_1 + \nu_2 - \nu_3}{2\nu_1}(1 - e^{-\nu_1 d}) \\
\frac{\nu_2 + \nu_3 - \nu_1}{2\nu_2}(1 - e^{-\nu_2 d}) & e^{-\nu_2 d} & \frac{\nu_1 + \nu_2 - \nu_3}{2\nu_2}(1 - e^{-\nu_2 d}) \\
\frac{\nu_2 + \nu_3 - \nu_1}{2\nu_3}(1 - e^{-\nu_3 d}) & \frac{\nu_1 + \nu_2 - \nu_3}{2\nu_3}(1 - e^{-\nu_3 d}) & e^{-\nu_3 d}
\end{pmatrix}.
\]

So, for this model, the transition probability matrix can be expressed as a function of parameters whose reciprocals are mean interval lengths, making it simpler to model. Note
that the transition rates are subject to the constraints $q_j > 0$ for $j = 1, 2, 3$. Therefore, we have the following constraint on the $\nu$-parameters

$$
\nu_1 < \nu_2 + \nu_3,
$$

$$
\nu_2 < \nu_1 + \nu_3,
$$

$$
\nu_3 < \nu_1 + \nu_2.
$$

We employ this constraint in our sampling algorithm which we describe in Section 4.3.2.

To our knowledge this is a novel specification of the distance-$d$ transition probability matrix. Such models were also considered for Illumina SNP genotyping arrays in Wang et al. (2009), however, our resulting transition probability matrix, $T(d)$, differs from theirs because we derive this matrix differently; see Section 4.2.2 for more details on our approach. In Wang et al. (2009), the matrix exponential of $Q$ was derived in closed-form, giving the following transition probability matrix,

$$
T(d) = \begin{pmatrix}
1 - (1 - \pi_1)(1 - e^{-\eta d}) & \pi_2(1 - e^{-\eta d}) & (1 - \pi_1 - \pi_2)(1 - e^{-\eta d}) \\
\pi_1(1 - e^{-\eta d}) & 1 - (1 - \pi_2)(1 - e^{-\eta d}) & (1 - \pi_1 - \pi_2)(1 - e^{-\eta d}) \\
\pi_1(1 - e^{-\eta d}) & \pi_2(1 - e^{-\eta d}) & 1 - (\pi_1 + \pi_2)(1 - e^{-\eta d})
\end{pmatrix},
$$

where $\pi_1 = \frac{q_1}{q_1 + q_2 + q_3}$, the frequency of deletions; $\pi_2 = \frac{q_2}{q_1 + q_2 + q_3}$, the frequency of the diploid state; $\pi_3 = \frac{q_3}{q_1 + q_2 + q_3}$, the frequency of amplifications; and $\eta = q_1 + q_2 + q_3$, a rate parameter which does not have a biological interpretation. Since there are no closed-form maximum likelihood solutions for the parameters of this matrix, Wang et al. (2009) proposed using a minorization of this matrix to develop an EM-algorithm, however, when analyzing real data they fix model parameters. We prefer using our version of $T(d)$ because, specifying sensible priors for $\nu_1, \nu_2$ and $\nu_3$ is, arguably, more straightforward than specifying priors for $\pi_1, \pi_2$ and $\eta$, since the $\nu$-parameters have direct biological interpretations as the reciprocals of the expected interval length. In addition, the structure of the diagonal elements of $T(d)$ derived by Wang et al. (2009) do not support conjugate
prior distributions and thus would require non-Gibbs samplers for all parameters of the transition probability matrix.

4.2.4 Heterogeneous model for CNPs

A location dependent process, i.e. a heterogeneous process, allows the transition intensities $q_{ij}$ to vary as a function of the location $t$, denoted by $q_{ij}(t)$. This is often referred to as a heterogeneous continuous-time Markov process. In reality, a homogeneous process is not appropriate for copy number data because of the existence of polymorphic regions along the genome. These regions contain copy number polymorphisms (CNPs) which are inherited CNVs and thus appear at higher-than-normal frequencies in a given population. Chapter 7 proposes a framework for modelling CNPs for studies with multiple samples. However, for the analysis of a single sample we do not have sufficient information to fit heterogeneous models. So it is necessary to assume homogeneity of the process.

4.3 Bayesian continuous-time model

The continuous-time HMM allows for model based inference on the mean length of CNVs, avoiding the need to compile this information post-analysis. This section presents a Bayesian framework for full-parameter estimation for the general model proposed in Section 4.2.1 as well as for the parsimonious submodel.

The continuous-time model uses the same specification for the emission distributions, including the outlier components, that was used for the BDHMM, which is given in Section 3.3.1.

4.3.1 Priors

For the general model with three DNA copy states the collection of all parameters is

$$\theta = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, p_{12}, p_{21}, p_{31}, \nu_1, \nu_2, \nu_3, \pi_{out,1}, \pi_{out,2}, \pi_{out,3}).$$
For the parameters of the distance-$d$ transition probability matrix, or equivalently the parameters of the $Q$-matrix, we use the priors:

\[
p_{12} \sim \text{Beta}(a_1, b_1),
\]
\[
p_{21} \sim \text{Beta}(a_2, b_2),
\]
\[
p_{31} \sim \text{Beta}(a_3, b_3),
\]
\[
\nu_1 \sim \text{Gamma}(2, 1/l_1),
\]
\[
\nu_2 \sim \text{Gamma}(2, 1/l_2),
\]
\[
\nu_3 \sim \text{Gamma}(2, 1/l_3),
\]

where, for the Gamma distributions, 2 is the value of the shape parameter and $1/l_j$ is the value of the scale parameter. We chose the Gamma($2, 1/l_j$) priors for the rate parameters because, as we explain below, they readily allow one to incorporate prior biological information into the model. Recall that the reciprocal of $\nu_1$ represents the expected length of a deletion and, under the above prior specification, it can be shown that $E(1/\nu_1) = l_1$. Therefore, $l_1$ is the prior expected length of a deletion. The other states have analogous interpretations. Furthermore, as will be shown in Section 4.3.2, these priors contribute information to the posterior distribution that is equivalent to one class-$j$ segment of length $l_j$. Although \textit{a priori} we usually never know the true expected length of a deletion in the population being studied, we do have a rough idea of the average length of CNVs from previous studies. For example, most studies report deletions and duplications whose lengths fall between 1 kb and 50 kb. Indeed, in a large study McCarroll et al. (2008) reported the mean length of detected CNPs was approximately 26 kb, however, as can be seen from Figure 3.1, a small proportion of CNVs exceed 100 kb in length. On the other hand, we usually observe long gaps between CNVs which corresponds to long stretches in the 2-copy state, so reasonable values of $l_2$, the prior expected length of the 2-copy state, could be between 2 Mb and 10 Mb.
The Gamma$(2,1/l_j)$ prior is noninformative enough to allow the data to adjust the prior means if the sample and/or chromosome being analyzed were to fall in the small proportion of samples harboring large CNVs (i.e. they are weakly informative). We will see an example of this in the data analysis of the HapMap sample which contains small deletions (all less than 20 kb) but large amplifications, including one greater than 90 kb.

Our Beta prior parameters $(a_1, a_2, a_3)$ and $(b_1, b_2, b_3)$ can be interpreted as pseudo-counts of the number of transitions between states. For example, $a_1$ is the prior number of observations for transitions between the 1-copy state and the 2-copy state, and $b_1$ is the prior number of observations for transitions between the 1-copy state and the 3-copy state. Since it is rare for a duplication to follow a deletion, we expect the probability $p_{12}$ to be close to 1. Thus, the specification of $a_1$ should be large relative to $b_1$.

For the outlier probabilities we use the priors:

\[
\pi_{\text{out},1} \sim \text{Beta}(1,20), \\
\pi_{\text{out},2} \sim \text{Beta}(1,20), \\
\pi_{\text{out},3} \sim \text{Beta}(1,20),
\]

which favours low outlier probabilities while still allowing the possibility of higher probabilities, if the data supports it.

For the parameters of the emission distribution we can use the same priors as for the discrete-time model which are given in Section 3.3.2. These priors are conjugate for all parameters except for the $\nu$-parameters for which no conjugate prior exists. This leaves us with a set of hyperparameters,

\[
\phi = (m_c, s_0^2, a_1, a_2, a_3, b_1, b_2, b_3, l_1, l_2, l_3),
\]

to specify.

For the parsimonious submodel, the collection of all parameters is

\[
\theta = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, \nu_1, \nu_2, \nu_3, \pi_{\text{out},1}, \pi_{\text{out},2}, \pi_{\text{out},3}).
\]
We can use the same priors for this model as above with the exception of the $p$-parameters which disappear for this model and subject to the constraint given in equation 4.9 for the $\nu$-parameters.

### 4.3.2 Sampling the model

The algorithm for sampling the copy number sequences remains the same as in Chapter 3 except that now $\lambda_{ij}$ is replaced by $T_{ij}(\Delta_t)$. Readers are referred to Section 3.4.2 for the details. The algorithm for sampling the parameters of $Q$ involves a Gibbs step for the transition probabilities and a Metropolis-Hastings step (Hastings 1970) for sampling the rate parameters. Namely,

- **Sample the transition probabilities from**

  $$
  p_{12} \sim \text{Beta}(a_1 + n_{12}^{(t)}, b_1 + n_{13}^{(t)}),
  $$

  $$
  p_{21} \sim \text{Beta}(a_2 + n_{21}^{(t)}, b_2 + n_{23}^{(t)}),
  $$

  $$
  p_{31} \sim \text{Beta}(a_3 + n_{31}^{(t)}, b_3 + n_{32}^{(t)}),
  $$

  where $n_{ij}^{(t)} = \sum_{m=2}^{M} 1(C_{m-1}^{(t)} = i, C_{m}^{(t)} = j)$, the number of transitions from state $i$ to state $j$ for MCMC scan $t$.

- For $c = \{1, 2, 3\}$, given the data, other model parameters and the most recently drawn value of $\nu_c$, denoted by $\nu_c^{(t-1)}$, sample $\nu_c$ using a Metropolis-Hastings algorithm:

  1. Sample a proposal $\nu^*_c$ from a $\text{Gamma}(1, \nu_c^{(t-1)})$ distribution. This proposal is centered around the most recently drawn value of $\nu_c$.

  2. Accept $\nu^*_c$ with probability

     $$
     \min \left\{ \frac{p(\nu^*_c \mid \sim) \Gamma(\nu_c^{(t-1)}, 1, \nu^*_c)}{p(\nu_c^{(t-1)} \mid \sim) \Gamma(\nu_c^{(t-1)}, 1, \nu_c^{(t-1)})}, 1 \right\},
     $$

     where

     $$
     p(\nu_c \mid \sim) \propto \nu_c^{-\nu_c} e^{-\nu_c \left( l_c + \sum_{j \in A_{ck}} \Delta_j \right)} \prod_{k=1, k \neq c, c \in A_{ck}} \prod_{j \in A_{ck}} (1 - e^{-\nu_c \Delta_j}),
     $$

     and $A_{ck}$ is the set of states not in state $c$. The sum $\sum_{m=2}^{M} 1(C_{m-1}^{(t)} = i, C_{m}^{(t)} = j)$ represents the number of transitions from state $i$ to state $j$ for MCMC scan $t$. This formulation allows for the continuous-time nature of the model, where the transitions are not constrained to integer times, and thus the rate parameters must be updated using a continuous distribution, such as a Gamma distribution, to account for the time between events.
where $\Gamma(\nu; a, b)$ is the density function for the Gamma$(a, b)$ distribution, $A_{cc} = \{ j : C_j = C_{j-1} = c \}$ for $c = 1, 2, 3$, and $A_{kl} = \{ j : C_j = l, C_{j-1} = k \}$ for $k, l = 1, 2, 3$.

We can see from the conditional posterior distribution of $\nu_j$ that the prior contributes information to the posterior distribution that is equivalent to one segment in class $j$ of length $l_j$, as well as a factor of $\nu_j$. To avoid computational underflows, which result due to the product of a large number of probabilities, we perform the computations on the log-scale.

After running the MCMC algorithm, we have a set of sampled copy number sequences \( \{C^{(t)}\}_{1 \leq t \leq T} \), where $C^{(t)} = (C_1^{(t)}, \ldots, C_M^{(t)})$ and $T$ is the number of MCMC scans, as well as a set of sampled model parameters \( \{\theta^{(t)}\}_{1 \leq t \leq T} \), where

\[
\theta^{(t)} = (\mu_1^{(t)}, \mu_2^{(t)}, \mu_3^{(t)}, \sigma_1^{2,(t)}, \sigma_2^{2,(t)}, \sigma_3^{2,(t)}, p_{12}^{(t)}, p_{21}^{(t)}, p_{31}^{(t)}, \nu_1^{(t)}, \nu_2^{(t)}, \nu_3^{(t)}, \pi_{\text{out,}1}^{(t)}, \pi_{\text{out,}2}^{(t)}, \pi_{\text{out,}3}^{(t)}).
\]

Note that based on these MCMC scans we can obtain a sample for the parameters of the Q-matrix using the transformation:

\[
q_{ij} = \nu_i p_{ij} \tag{4.11}
\]

for $i, j = 1, 2, 3$.

For the parsimonious submodel, introduced in Section 4.2.3, sampling can be done in the same fashion except we drop the steps for sampling the $p$-probabilities and the Metropolis-Hastings algorithm for the $\nu$-parameters needs to be modified since the conditional posterior, $p(\nu^{(t)}_c | \sim)$, changes for this model. The details are given in Appendix A. For this model the transition intensities can be recovered by

\[
q_1 = (\nu_2 + \nu_3 - \nu_1)/2
\]
\[
q_2 = (\nu_1 + \nu_3 - \nu_2)/2
\]
\[
q_3 = (\nu_1 + \nu_2 - \nu_3)/2.
\]
4.4 Data analysis

This section reanalyzes HapMap sample NA18550, which was first analyzed using the BDHMM in Section 3.8, with the Bayesian continuous-time HMMs. The more general model, which we refer to as BCHMM1, is compared to the parsimonious submodel, which we refer to as BCHMM2.

4.4.1 Model 1 (BCHMM1)

For the hyperparameters of the emission distribution parameters we used the same values as in the BDHMM. For the parameters of the transition intensity matrix we used $l_0 = 25000, l_1 = 35000, l_2 = 10^7, l_3 = 35000$, reflecting a prior mean length of hemizygous deletions and amplifications of 35 kb, a prior mean length of homozygous deletions of 25 kb, and a prior mean length of stays in the 2-copy state of 10 Mb. Results are based on 5,000 MCMC scans, after a burn-in period of 1,000 scans. CNVs were called by first using the estimated marginal probabilities to assign copy number values to each SNP, then a CNV was defined as a sequence of SNPs with non-two copy values.

The middle panel of Figure 4.2 shows the estimated copy number along the genome for BCHMM1. For comparison with BDHMM, Figure 4.3 includes the estimated copy number for both the BDHMM and BCHMM1 for the first 2000 SNPs analyzed. While BCHMM1 has a smaller number of apparent false positives compared to BDHMM, there are still too many short visits between the 1- and 2-copy states in a region that does not appear to harbor any CNVs. Table 4.1 shows a summary of the posterior distribution of the model parameters. The important thing to notice is that, similar to the results for BDHMM, there is very little separation between the 1- and 2-copy states which had posterior mean values for their class means of 0.03 and -0.01, respectively.
Figure 4.2: Whole chromosome display. The top display shows the observed raw CNs for sample NA18550 on the second arm of chromosome 10, the middle and bottom displays show the estimated copy number of these SNPs for BCHMM1 and BCHMM2, respectively.

4.4.2 Model 2 (BCHMM2)

We also analyzed the data using the submodel described in Section 4.2.3. For this model, the mixing of the Markov chain for the \( \nu \)-parameters was very slow and showed extremely high autocorrelation. This can be attributed to the more constrained nature of the transition intensity matrix for this model. We found that sampling the transition intensities, i.e. the \( q \)-parameters, directly, worked much better. Note that the transition intensities are only subject to the constraint that \( q_j > 0 \) for \( j = 0, 1, 2, 3 \), which is not as restrictive as the constraint required for the \( \nu \)-parameters. So we present results based
Figure 4.3: Subregion of 2000 SNPs. The top display shows the observed raw CNs for the first 2000 SNPs on the chromosome, the middle display shows the estimated copy number of these SNPs for BDHMM, and the bottom display shows the estimated copy number of these SNPs for BCHMM1.

on this approach which used the following priors on the transition intensity parameters:

\[
q_0 \sim \text{Gamma}(1, 10^{-7}),
\]
\[
q_1 \sim \text{Gamma}(1, 10^{-7}),
\]
\[
q_2 \sim \text{Gamma}(1, 10^{-4}),
\]
\[
q_3 \sim \text{Gamma}(1, 10^{-7}),
\]

and we also use a similar Metropolis-Hastings algorithm to sample from the posterior.
Figure 4.4: Trace plots of the $\nu$-parameters for BCHMM2.

We took 30,000 MCMC scans, discarded the first 5,000, then thinned the chain by 5, leaving 5,000 scans for the analysis. Table 4.1 shows a summary of the posterior results for the model parameters of BCHMM2 and Figure 4.4 shows the trace plots for the $\nu$-parameters, which now show good mixing of the Markov chain. The estimated copy number sequence for this model is given in the lower panel of Figure 4.2. Although there are fewer transitions between the 1- and 2-copy states for BCHMM2 compared to BCHMM1, the copy number reconstruciton is biologically implausible: there are still way too many short visits to the 1-copy state than would be realistically possible and, as shown in Table 4.1, the 1-copy state again merges with the 2-copy state.
Chapter 4. A Continuous-time HMM

4.5 Chapter Summary and Discussion of Results

This chapter introduced a novel Bayesian continuous-time HMM for inferring DNA copy number changes from SNP arrays. A Bayesian framework was proposed as a method by which to estimate model parameters and incorporate prior information into the model. The remainder of the chapter discussed a parsimonious submodel, the implementation of the MCMC algorithm for the Bayesian model, and data analysis of the HapMap data.

Although an improvement over the discrete-time HMM, the Bayesian continuous-time HMM has not fixed the problem of the merging of the 1-copy state with the 2-copy state, and ultimately the model is still predicting more CNVs than we would expect according to biological knowledge. Chapter 5 will introduce a modified continuous-time HMM which leads to biologically plausible results. Thus, a more detailed comparison between the general model and the submodel is postponed until the next chapter.
Table 4.1: Posterior summary of the parameters of BCHMM1, first two columns, and BCHMM2, last two columns.

<table>
<thead>
<tr>
<th></th>
<th>BCHMM1</th>
<th></th>
<th>BCHMM2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>$\mu_0$</td>
<td>-1.894</td>
<td>0.156</td>
<td>-1.917</td>
<td>0.162</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>0.032</td>
<td>0.024</td>
<td>-0.003</td>
<td>0.033</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>-0.014</td>
<td>0.001</td>
<td>-0.015</td>
<td>0.001</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>0.304</td>
<td>0.019</td>
<td>0.305</td>
<td>0.016</td>
</tr>
<tr>
<td>$\sigma^2_0$</td>
<td>0.791</td>
<td>0.282</td>
<td>0.821</td>
<td>0.288</td>
</tr>
<tr>
<td>$\sigma^2_1$</td>
<td>0.234</td>
<td>0.026</td>
<td>0.155</td>
<td>0.021</td>
</tr>
<tr>
<td>$\sigma^2_2$</td>
<td>0.038</td>
<td>0.000</td>
<td>0.038</td>
<td>0.001</td>
</tr>
<tr>
<td>$\sigma^2_3$</td>
<td>0.050</td>
<td>0.005</td>
<td>0.050</td>
<td>0.006</td>
</tr>
<tr>
<td>$\pi_{out,0}$</td>
<td>0.133</td>
<td>0.065</td>
<td>0.113</td>
<td>0.065</td>
</tr>
<tr>
<td>$\pi_{out,1}$</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>$\pi_{out,2}$</td>
<td>0.002</td>
<td>0.001</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td>$\pi_{out,3}$</td>
<td>0.014</td>
<td>0.009</td>
<td>0.011</td>
<td>0.008</td>
</tr>
<tr>
<td>$1/\nu_0$</td>
<td>11201.717</td>
<td>4918.330</td>
<td>10124.221</td>
<td>3884.416</td>
</tr>
<tr>
<td>$1/\nu_1$</td>
<td>2114.045</td>
<td>399.749</td>
<td>10449.013</td>
<td>3965.128</td>
</tr>
<tr>
<td>$1/\nu_2$</td>
<td>155880.801</td>
<td>19791.970</td>
<td>265130.692</td>
<td>70839.676</td>
</tr>
<tr>
<td>$1/\nu_3$</td>
<td>89502.842</td>
<td>46863.016</td>
<td>10171.444</td>
<td>4044.153</td>
</tr>
<tr>
<td>$p_{01}$</td>
<td>0.198</td>
<td>0.164</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{02}$</td>
<td>0.360</td>
<td>0.190</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{10}$</td>
<td>0.004</td>
<td>0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{12}$</td>
<td>0.984</td>
<td>0.011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{20}$</td>
<td>0.013</td>
<td>0.007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{21}$</td>
<td>0.971</td>
<td>0.013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{30}$</td>
<td>0.360</td>
<td>0.161</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{31}$</td>
<td>0.244</td>
<td>0.186</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 5

A Continuous-Time HMM with Copy-Number-Specific Emissions

5.1 Introduction

While the Bayesian continuous-time HMM presented in Chapter 4 had some desirable properties, including accounting for the physical distance between SNPs in a model-based framework and allowing for inference on the expected length of CNVs, the results produced by the model when analyzing real Affymetrix SNP array 6.0 data were still biologically implausible. This chapter addresses two issues regarding the modelling of the 0- and 2-copy states. The improvement gained is assessed by analysing the HapMap data with these modifications.

5.2 Novel HMM

As outlined in Chapter 2, previous applications of HMMs to SNP array data for copy number inference made numerous assumptions and simplifications, including (i) that the emission probabilities of the model, which specify the probability distribution of the data given the unobserved DNA copy number, are symmetric about their mean, (ii) that the
measurement variances are constant across the copy number classes, (iii) fixing model parameters to arbitrary values, and (iv) using the Viterbi algorithm based on fixed model parameters to reconstruct the DNA copy number sequence. Chapters 3 and 4 developed models that dealt with (ii) to (iv). It is also clear that assumption (i) is not always tenable, particularly for the 0-copy state which often exhibits skewness. This section proposes modifications to the emission probabilities of the 0- and 2-copy states that are more robust to the noisy genomic data.

Figure 5.1: Estimated density using kernel density estimation for data in the 1- and 2-copy states grouped according to copy number values for a randomly selected MCMC iteration for BCHMM1.
5.2.1 Emission probabilities, 2-copy state

In the previous analysis we noticed that there was very little separation between the posteriors of the 1- and 2-copy states, that is, the posterior distribution of the 1-copy state collapsed to form another 2-copy state in the sense that the resulting emission distributions for these two states overlapped. This is readily seen in Figure 5.1 which shows the estimated density of the raw CNs in the 1- and 2-copy states grouped according to copy number values for a randomly selected MCMC iteration based on BCHMM1 (see Chapter 4). The distributions of the clustered data for other MCMC samples showed very similar features. This appears to resemble a mixture of two distributions both centered near zero. The first is a high-density, low-variance distribution, the second is a small-density, high-variance distribution. Note also that the slight bump in the density of the 1-copy state observations between -0.5 and -1.0 probably corresponds to the real 1-copy state observations. Thus, the posterior distribution indicates that a single distribution is not sufficient for modelling the 2-copy state. To address this, we use a mixture of two Normal distributions and a Uniform distribution,

\[ Y_i | C_i = 2 \sim \pi_{out,2} / (Y_{max} - Y_{min}) + (1 - \pi_{out,2})(\pi_{2A} N(Y_i; \mu_{2A}, \sigma_{2A}^2) + (1-\pi_{2A})N(Y_i; \mu_{2B}, \sigma_{2B}^2)) \]

where there are two substates with different means and variances as well as a frequency mixture weight for the normal substates, \( \pi_{2A} \), within the non-outlier component.

5.2.2 Emission probabilities, 0-copy state

A common assumption is that the emission distributions for all states are symmetric about their mean. While this may be approximately correct for some copy states, it certainly is not true for the 0-copy state. The reason for this is that observations in the 0-copy state are derived from taking the logarithm of near-zero values which has the effect of substantially blowing up the noise. This can be seen in Figure 5.2 which shows
a histogram of raw CNs from 29 HapMap samples that share a homozygous deletion (0-copy state) containing 27 SNPs, which were identified as 0-copy state observations in a previous study of CNPs by McCarroll et al. (2008). (This is an example of an inherited CNV that is prevalent in the HapMap population at a frequency of 20 out of 270 samples.) The distribution is very much asymmetric with a heavy tail favouring very
small values, that is, the distribution is skewed left. We propose using an extreme value
type-distribution for this state. Specifically, we use the emission distribution

\[ Y_i | C_i = 0 \sim \pi_{out,0}/(Y_{max} - Y_{min}) + (1 - \pi_{out,0})f_0(Y_i; \beta), \]  

(5.2)

where

\[ f_0(y; \beta) = \beta 2^y \ln(2) \exp(-\beta 2^y), \]

which is the density corresponding to a random variable, \( Y_i \), that is defined as \( Y_i = \log_2(X_i) \) where \( X_i \sim \text{Exp}(\beta) \), \( \text{Exp}(\beta) \) denoting the Exponential distribution with parameter \( \beta \). Thus, the reciprocal of \( \beta \) can be seen to be the mean on the non-logged scale.
Specifically, recall that the raw CN is computed as \( Y_i = \log_2(\psi_i/\psi_{R_i}) \), the logarithm of the ratio of the sample signal to the SNP-specific reference signal. Refer to Chapter 1 for more details. Then our emission distribution for the 0-copy state is equivalent to the assumption that

\[ [\psi_i/\psi_{R_i} | C_i = 0] \sim \text{Exp}(\beta). \]  

(5.3)

Taking base-2 logarithms results in data having distribution of the form given by \( f_0(y; \beta) \).
Figure 5.2 compares this distribution, shown in red, to a Normal distribution, shown in blue. While not a perfect fit for the data, it fits better than a Normal and, as we will see in the data analysis, this is critical in avoiding a abnormally high outlier rate which can lead to incorrect state predictions.

5.2.3 Priors

For the emission distributions parameters, with the exception of the 0- and 2-copy states, we continue to use the same Normal-inverse-Chi-squared conjugate priors which are given
in Section 3.3.2. For the emission distribution parameters of the 2-copy state we use

\[ \mu_{2A} \sim N(0, \sigma_{2A}^2/b), \]

\[ \mu_{2B} \sim N(0, \sigma_{2B}^2/b), \]

\[ \frac{1}{\sigma_{2A}^2} \sim \frac{1}{d_{0.2}0.05 \chi_{d_{0.2}}^2}, \]

\[ \frac{1}{\sigma_{2B}^2} \sim \frac{1}{d_{0.2}0.05 \chi_{d_{0.2}}^2}, \]

and a flat Beta prior for the 2-copy substate frequency mixture weight,

\[ \pi_{2A} \sim \text{Beta}(1, 1). \]

The prior means of \( \mu_{2A} \) and \( \mu_{2B} \) are both set to zero to reflect the fact they are both modelling the 2-copy state.

For \( \beta \), the parameter of the non-outlier component for the 0-copy state, we found that a Normal prior on the logarithm of \( \beta \) works quite well in practice,

\[ \log(\beta) \sim N(\log(5), 1), \tag{5.4} \]

where \( \log(5) \) was chosen to reflect a copy number ratio somewhere between 0/2 (zero copy number) and 1/2 (one copy number), but closer to 0/2. So, we chose 1/5. Recall that the reciprocal of \( \beta \) corresponds to the mean of the 0-copy state observations on the non-logged scale, so 1/2 is the theoretical ratio of 1-copy state observations and 0/2 is the theoretical ratio of 0-copy state observations. Alternatively, a conjugate Gamma prior distribution could also be used.
5.2.4 Sampling

The sampling algorithm is the same as given in Section 4.3.2 with the following exceptions. To sample $\beta$, we first derive the conditional posterior distribution of $\beta$:

$$p(\beta | \sim) \propto p(y | C, \theta) p(\beta) \propto \beta^{n_{0,nout} - 1} \exp \left( -\beta \sum_{i:C_i=0,Out_i=0} 2^{y_i} \right) \exp \left( -\frac{(\log(\beta) - \log(5))^2}{2} \right),$$

where $n_{0,nout}$ is the number of observations in the 0-copy state that are not outliers. Given the data, other model parameters and the most recently drawn value of $\beta$, denoted by $\beta^{(t-1)}$, sample $\beta$ using a Metropolis-Hastings algorithm:

1. Sample a proposal $\beta^*$ from a $N(\beta^{(t-1)}, 0.5)$ distribution. This proposal is centered around the most recently drawn value of $\beta$.

2. Accept $\beta^*$ with probability $\max \left\{ \frac{p(\beta^* | \sim)}{p(\beta^{(t-1)} | \sim)}, 1 \right\}$.

5.3 Data Analysis

This section reanalyzes the HapMap data, which we originally analyzed in Section 3.8, using the modifications proposed in this Chapter. First, it is shown that the BCHMM with a mixture of two Normals and a Uniform distribution for the 2-copy state, which we refer to as the Bayesian continuous-time hidden Markov model with diploid mixture (BCHMMDmix), leads to improved results that are biologically plausible. Then, it is shown that employing the extreme value type-distribution for the 0-copy state further improves the results. We refer to this latter model as BCHMMDmix2. Finally, these results are compared to a parsimonious version of BCHMMDmix2 based on the submodel for the $Q$-matrix introduced in Section 4.2.3. The submodel assumes four transition intensity parameters ($q_1, q_2, q_3, q_4$) instead of twelve. Specifically, the submodel makes the following constraints on the transition intensity parameters: $q_{10} = q_{20} = q_{30} = q_0$. 
\[ q_{01} = q_{21} = q_{31} = q_1, \quad q_{02} = q_{12} = q_{32} = q_2, \quad \text{and} \quad q_{03} = q_{13} = q_{23} = q_3. \]

We refer to this submodel as BCHMMDmix3.

Figure 5.3: Whole chromosome display. The top display shows the observed raw CNs for sample NA18550 on the second arm of chromosome 10 and the bottom display shows the estimated copy number of these SNPs for BCHMMDmix.

5.3.1 Model 1 (BCHMMDmix)

The data are reanalyzed using the BCHMM with a mixture of two Normals and a Uniform distribution for the 2-copy state. We used the same priors as in the previous analysis except for the emission model parameters of the 2-copy state for which we use the priors given in Section 5.2.3.

As can be seen from Figures 5.3 and 5.4, the estimated copy number sequence of
Figure 5.4: Subregion of 2000 SNPs. The top display shows the observed raw CNs for the first 2000 SNPs on the chromosome, the middle display shows the estimated copy number of these SNPs for BCHMM, and the bottom display shows the estimated copy number of these SNPs for BCHMMDmix.

the current model does not transition uncontrollably between the aberrant and normal states like the reconstructed sequences of the previous versions of the BCHMM. This reconstructed copy number sequence is more biologically plausible, and the class means of the states, given in Table 5.2 as posterior means, now show good separation between the 1- and 2-copy states with values of -0.53 and -0.026 (substate 2A), respectively. Thus, modelling by a mixture of two Normals and a Uniform was necessary for the 2-copy state. Within the 2-copy state, the posterior supports a model whose component 2A corresponds to a larger, low variance subclass (posterior mean of 0.025 for $\sigma^2_{2A}$); whereas, component
2B corresponds to a smaller, high variance subclass (posterior mean of 0.094 for $\sigma_{2B}^2$).

![BCHMMDmix](image1.png)

![BCHMMDmix2](image2.png)

Figure 5.5: Subregion of 200 SNPs containing two SNPs that appear to have been misclassified by BCHMMDmix. Green points correspond to 3-copy state predictions, whereas red points correspond to 0-copy state predictions.

Before interpreting the rest of the results, we address another issue with the model. Although this version of the model results in a substantial improvement in the results, there still exists another modelling issue: two adjacent SNPs having large positive values, corresponding to the observed values of 2.42 and 1.18, were classified to the 0-copy state. The region encompassing 200 SNPs surrounding the two SNPs in question is shown in Figure 5.5. Clearly, these two SNPs have been wrongly classified into the 0-copy state. The cause of this is the inadequacy of the Normal emission distribution for the 0-copy state. The model attributes a high variance for the 0-copy state (posterior mean
of 0.81 for $\sigma_0^2$) and, as a result, this state is picking up SNPs that have large, outlying positive values. Also, the high outlier rate for the 0-copy state (posterior mean of 0.12 for $\pi_{\text{out},0}$) attributed to the model suggests that the model is still inadequate for the data. For example, using the posterior mean values for illustration, the estimated emission probability of this observation for the 0-copy state is

$$0.12/(Y_{\text{max}} - Y_{\text{min}}) + (1 - 0.12) \times N(y = 1.18, \mu_0 = -1.92, \sigma_0 = \sqrt{0.81}) = 0.015,$$

whereas the estimated emission probability of this observation for the 3-copy state is

$$0.02/(Y_{\text{max}} - Y_{\text{min}}) + (1 - 0.02) \times N(y = 1.18, \mu_3 = 0.27, \sigma_3 = \sqrt{0.05}) = 0.003.$$

So, under the posterior, it is more likely these large positive values are 0-copy state outliers rather than 3-copy state outliers. Thus, the 0-copy state is serving as an outlier state that is centered around 0-copy state observations for all other states. This is inappropriate and in the next section we show that using the extreme value type-distribution for the 0-copy state addresses this problem.

### 5.3.2 Model 2 (BCHMMDmix2)

The bottom panel of Figure 5.5 shows the subregion containing the two SNPs in question for BCHMMDmix2. For this model the two questionable SNPs are now classified as being in the 3-copy state rather than the 0-copy state. For BCHMMDmix2, the estimated marginal probabilities of being in the 3-copy state at the questionable SNPs are 0.71 and 0.74, whereas for BCHMMDmix the probabilities are 0.39 and 0.42. Also, its 0-copy state outlier rate is more reasonable with a posterior mean of 0.04 versus 0.12 for BCHMMDmix, see Table 5.2 for posterior details. Therefore, modelling the 0-copy state emission distribution with an extreme value type-distribution succeeded in addressing this problem and improving the results. Note that aside from the estimated copy number at these two SNPs, there was only one other SNP at which the two models disagreed on.
Figure 5.6: Posterior distributions for the expected interval length in base pairs. Prior densities are superimposed in blue. The prior density for the 3-copy state (lower right panel) was rescaled to fit inside the plotting region.

This SNP is located on the boundary of the largest detected amplification seen in Figure 5.3 between the loci at 56 and 60 Mb. We discuss this region in more detail in Chapter 6.

With the continuous-time model we can also infer the expected lengths of the copy states for this sample by examining the posterior distribution of the $\nu$-parameters. For example, Figure 5.6 shows the posterior distributions of the mean length for each region along with their superimposed prior densities. We see good separation between the priors and the posterior, indicating that we are not just sampling from the prior distributions.
The posterior mean of the expected length of homozygous deletions (copy number 0), hemizygous deletions (copy number 1), diploid regions (copy number 2) and amplifications (copy number 3) is 10962 bps, 7926 bps, 2336463 bps, and 131511 bps, respectively. In sum, this tells us that this sample, on average, has much longer amplifications than deletions. This is consistent with what we know about this sample from previous studies: it contains one short and one moderate homozygous deletion (average length of 9000 bps); four short hemizygous deletions (average length of 3774 bps); and one moderate and one long amplification (average length of 94530 bps). Thus, the model does a decent job of recovering known information on CNV lengths. In addition, our model-based estimates of mean length may differ slightly from those reported in McCarroll et al. (2008) since we found additional CNVs that were not reported in their study. In total, BCHMMDmix2 found 16 CNVs, eight more than were reported in McCarroll et al. (2008). Seven out of the sixteen CNVs we found overlapped with CNVs reported in McCarroll et al. (2008). The additional CNVs that we found were probably rare CNVs or low-frequency CNPs since McCarroll et al. (2008) focused their method on finding CNPs (i.e. inherited CNVs). We will discuss these CNVs and their significance in more detail in Chapter 6 where we delve into the topic of statistical inference for subsequences, or detected CNVs.

Figure 5.7 shows the posterior and prior density of $\beta$. This also shows good separation between the prior and posterior distributions. From Table 5.2 we see that the posterior mean of $\beta$ is 4.84, corresponding to a non-logged copy number ratio of 0.21 for the 0-copy state. Other class means had posterior means of -0.53, -0.026, 0.026, and 0.266 for the 1-copy state, 2-copy-A substate, 2-copy-B substate, and 3-copy state, respectively.

### 5.3.3 Model 3 (BCHMMDmix3)

BCHMMDmix3, the parsimonious submodel, produced similar results to BCHMMDmix2, however, there were some key differences. Table 5.1 shows the number of SNPs with differing estimated copy numbers for the two models. In total, 14 SNPs (out of
25,793) had different copy number, including a segment consisting of 8 SNPs which was identified as a hemizygous deletion by BCHMMDmix2 but was not identified as a deletion by BCHMMDmix3. In Chapter 6, we will see that, in fact, this segment is likely to be a deletion. Furthermore, as we can see in Table 5.2, the posterior means for the model parameters are very similar for the two models except for the mean interval lengths. The posterior means of the mean interval lengths for BCHMMDmix3 were 64739 bps, 64948 bps, 2130764 bps, and 65448 bps for homozygous deletions, hemizygous deletions, diploid regions, and amplifications, respectively. This is noticeably less consistent with the results of previous studies in that this sample is known to have longer amplifications than deletions. We suspect the reason for this is that the model for the transition intensity
Table 5.1: Estimated copy number, per SNP, for BCHMMDmix2 (row) and BCHMMDmix3 (column).

<table>
<thead>
<tr>
<th></th>
<th>BCHMMDmix2</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>26</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>25348</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>366</td>
<td></td>
</tr>
</tbody>
</table>

matrix is overly simplified. This can be seen by realizing that the difference between any two \( \nu \)-parameters depends on only two parameters, whereas for the more general model it depends on six parameters. For example, for the submodel

\[
\nu_1 - \nu_3 = q_3 - q_1,
\]

whereas for the more general model

\[
\nu_1 - \nu_3 = q_{10} + q_{12} + q_{13} - q_{30} - q_{31} - q_{32}.
\]

So model-based mean interval lengths are likely to be more similar for the submodel than the more general model due to the more restrictive nature of the submodel. Interestingly, although the inferred mean interval lengths are very close for BCHMMDmix3, the posterior probability that amplifications are longer than homozygous deletions, \( P(1/\nu_3 > 1/\nu_0 | Y) \), computed as a frequency across the MCMC scans, is 0.95, still quite large. Therefore, we believe the constraints on the submodel preclude accurate inferences on the mean interval lengths, and thus, we prefer using the more general model.
5.4 Chapter Summary and Discussion of Results

This chapter proposed modifications to the Bayesian continuous-time HMM, including the modelling of the 0- and 2-copy states. In the analysis of the HapMap data, it was shown that these modifications lead to biologically plausible results. Specifically, it was shown that for the emission probability of the 2-copy state, a mixture of two Normals and a Uniform distribution was necessary in preventing the merging of the posterior of the 1-copy state with that of the 2-copy state. The two Normal distributions within the non-outlier component of the 2-copy state were seen to serve complementary roles: the first, as a high-density, low variance subclass; the second, as a low-density, high variance subclass. Also, using an extreme value type distribution for the 0-copy state resulted in a reduced outlier rate suggesting a better model fit, as well as more sensible classification of a couple SNPs that had been previously misclassified. The framework laid out in this chapter should provide researchers with a HMM that can be used to estimate the DNA copy number and the model parameters from the data, thus, avoiding the need to fix model parameters to arbitrary values.

Up to this point, the thesis has focused on a HMM for effectively reconstructing the DNA copy number sequence of a whole chromosome. The next chapter discusses Bayesian inference for subchromosomal regions.
Table 5.2: Posterior summary of the parameters of BCHMMDmix2, first two columns, and BCHMMDmix3, last two columns.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>BCHMMDmix2</th>
<th></th>
<th>BCHMMDmix3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>-0.530</td>
<td>0.055</td>
<td>-0.523</td>
<td>0.072</td>
</tr>
<tr>
<td>$\mu_2$A</td>
<td>-0.026</td>
<td>0.002</td>
<td>-0.026</td>
<td>0.002</td>
</tr>
<tr>
<td>$\mu_2$B</td>
<td>0.026</td>
<td>0.006</td>
<td>0.024</td>
<td>0.006</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>0.266</td>
<td>0.014</td>
<td>0.269</td>
<td>0.016</td>
</tr>
<tr>
<td>$\sigma^2_1$</td>
<td>0.089</td>
<td>0.023</td>
<td>0.106</td>
<td>0.030</td>
</tr>
<tr>
<td>$\sigma^2_2$A</td>
<td>0.025</td>
<td>0.001</td>
<td>0.025</td>
<td>0.001</td>
</tr>
<tr>
<td>$\sigma^2_2$B</td>
<td>0.005</td>
<td>0.005</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td>$\sigma^2_3$</td>
<td>0.054</td>
<td>0.005</td>
<td>0.053</td>
<td>0.005</td>
</tr>
<tr>
<td>$\beta$</td>
<td>4.842</td>
<td>0.805</td>
<td>4.937</td>
<td>0.839</td>
</tr>
<tr>
<td>$\pi_{out,0}$</td>
<td>0.043</td>
<td>0.041</td>
<td>0.037</td>
<td>0.037</td>
</tr>
<tr>
<td>$\pi_{out,1}$</td>
<td>0.023</td>
<td>0.023</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>$\pi_{out,3}$</td>
<td>0.020</td>
<td>0.010</td>
<td>0.022</td>
<td>0.010</td>
</tr>
<tr>
<td>$\pi_{dip,A}$</td>
<td>0.744</td>
<td>0.021</td>
<td>0.734</td>
<td>0.025</td>
</tr>
<tr>
<td>$\pi_{dip,B}$</td>
<td>0.255</td>
<td>0.021</td>
<td>0.264</td>
<td>0.025</td>
</tr>
<tr>
<td>$1/\nu_0$</td>
<td>10961.660</td>
<td>5120.590</td>
<td>64739.319</td>
<td>25913.025</td>
</tr>
<tr>
<td>$1/\nu_1$</td>
<td>7025.961</td>
<td>3676.309</td>
<td>64948.231</td>
<td>26085.569</td>
</tr>
<tr>
<td>$1/\nu_2$</td>
<td>2336463.409</td>
<td>664698.100</td>
<td>2130764.176</td>
<td>541640.278</td>
</tr>
<tr>
<td>$1/\nu_3$</td>
<td>131510.774</td>
<td>59707.262</td>
<td>65447.997</td>
<td>26408.833</td>
</tr>
<tr>
<td>$p_{01}$</td>
<td>0.043</td>
<td>0.040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{02}$</td>
<td>0.860</td>
<td>0.072</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{10}$</td>
<td>0.038</td>
<td>0.036</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{12}$</td>
<td>0.917</td>
<td>0.055</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{20}$</td>
<td>0.162</td>
<td>0.087</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{21}$</td>
<td>0.444</td>
<td>0.133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{30}$</td>
<td>0.107</td>
<td>0.061</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_{31}$</td>
<td>0.078</td>
<td>0.052</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 6

Bayesian Inference for Subsequences

6.1 Introduction

After running the MCMC algorithm, we have a set of sampled copy number sequences 
\( \{ C^{(t)} \}_{1 \leq t \leq T} \), where \( C^{(t)} = (C^{(t)}_1, \ldots, C^{(t)}_M) \) and \( T \) is the number of MCMC scans. Chapter 3 introduced Bayesian techniques for estimating the DNA copy number sequence based on these results. However, we may also be interested in answering questions about subchromosomal regions or subsequences. For example, we may want to quantify the evidence that an individual segment within the estimated sequence is in fact a CNV. Such questions are often of interest to practitioners who require the ability to rank detected CNVs from SNP arrays so that they can focus further experimental efforts on the most promising regions. This chapter proposes measures of uncertainty for detected regions from the HMM.

6.2 Inference on subsequences

This section discusses the use of posterior probabilities as measures of uncertainty as well as effective plotting techniques for summarizing interesting subsequences. In this treatment, particular attention is paid to the issue of effectively accounting for the uncertainty
around the boundaries of the CNVs, which typically can be quite high.

### 6.2.1 Estimating posterior probabilities of CNVs

Define a segment to be a sequence of SNPs containing aberrant values \((\neq 2)\) of the same state; for example, if \(C_{i-1}^{(t)} = 2, C_i^{(t)} = \cdots = C_j^{(t)} \neq 2, C_{j+1}^{(t)} = 2\), then there is a sampled segment from SNP \(i\) to SNP \(j\) for sequence \(t\). We denote such a segment as \(C_{ij}\). For a given candidate segment, we can quantify the evidence that it is a CNV by computing its posterior probability, \(P(C_{ij}|Y)\), based on the MCMC scans. For example, given a candidate segment \(C_{ij}\) that is suspected of being a deletion from position \(d_i\) to \(d_j\), we can use the sample average based on the MCMC scans to estimate the probability that it is a deletion,

\[
P(C_{ij}|Y) \approx \frac{1}{T} \sum_{t=1}^{T} 1(C_{i-1}^{(t)} = 2, C_i^{(t)} = \cdots = C_j^{(t)} = 1, C_{j+1}^{(t)} = 2). \tag{6.1}
\]

We refer to these estimates as frequency-based estimates and denote them by \(\hat{P}(C_{ij}|Y)\).

Alternatively, we can also use Rao-Blackwellized estimates by noticing that

\[
P(C_{ij}|Y, \theta) = \frac{\alpha_{i-1}(2; \theta)(\prod_{l=i}^{j} f_1(y_l) T_{c_{i-1},c_l(\Delta_l)}) f_2(y_{j+1}) \beta_{j+1}(2; \theta)}{P(Y)}. \tag{6.2}
\]

The computation of \(P(C_{ij}|Y)\) using this approach entails running a forward and backward pass at each MCMC iteration, which we are already doing as part of the sampling algorithm (described in detail in Section 3.4.2), thus, this does not increase the computational demand. So, we can estimate \(P(C_{ij}|Y)\) by summing probabilities rather than the events that were generated by these probabilities. Namely,

\[
P(C_{ij}|Y) \approx \frac{1}{T} \sum_{t=1}^{T} P(C_{ij}|Y, \theta^{(t)}), \tag{6.3}
\]

where \(P(C_{ij}|Y, \theta^{(t)})\) is computed using equation 6.2. See Appendix B for a proof of equation 6.2.

In practice, as pointed out by Lai et al. (2008), there is usually some uncertainty about the endpoints \(i, j\) of the segment. Therefore, they suggested that in assessing
the evidence for \( C_{ij} \) it is more appropriate to consider a set of segments with variation
around the endpoints \( i \) and \( j \):

\[
A_{ij}(k) = \{(i^*, j^*) : |i - i^*| \leq k, |j - j^*| \leq k\}, \tag{6.4}
\]

where \( k \) is the number of locations by which the endpoints of the segment can differ from
the endpoints of the segment of interest. Since the segments belonging to the set \( A \) are
disjoint (in the probability sense, not the physical sense),

\[
P(\bigcup_{l,m \in A_{ij}} C_{lm} | Y) = \sum_{l,m \in A_{ij}} P(C_{lm} | Y).
\]

Therefore, we can estimate this probability from the MCMC iterations as \( \hat{P}(C_{ij} | Y) \),
where

\[
\hat{P}(C_{ij} | Y) = \sum_{l,m \in A_{ij}} P(C_{lm} | Y)
= \sum_{l,m \in A_{ij}} \left( \frac{1}{T} \sum_{t=1}^{T} 1(C^{(t)}_{l-1} = 2, C^{(t)}_{l} = \ldots C^{(t)}_{m} = 1, C^{(t)}_{m+1} = 2) \right)
\]

where each \( P(C_{lm} | Y) \) is computed using equation 6.1. So these refined estimates, denoted
by \( \hat{P}(C_{ij} | Y) \), are based on a frequency approach with adjustments to the endpoints by
admitting some uncertainty around the endpoints. In practice, \( k \) can be chosen so that
the set \( A \) includes all segments that overlap with the segment of interest, \( C_{ij} \). Note also
that by construction \( \hat{P}(C_{ij} | Y) > P(C_{ij} | Y) \), and significant discrepancy between the two
estimates would indicate uncertainty in the endpoints of the segment.

Alternatively, we could use equation 6.2 to get Rao-Blackwellized estimates. These
are given by

\[
P(C_{ij} | Y) \approx \sum_{l,m \in A_{ij}} P(C_{lm} | Y)
= \sum_{l,m \in A_{ij}} \left( \frac{1}{T} \sum_{t=1}^{T} P(C_{lm} | Y, \theta^{(t)}) \right),
\]
where $P(C_{lm}|Y,\theta^{(t)})$ is computed using equation 6.3. In the next section, we propose an approach for visualizing the segmental structure of the MCMC results which is useful for gaining more insight into the probable structure of a subsequence.

### 6.2.2 Approach to visualizing the MCMC results

While computing and comparing probabilities for each segment of interest is useful, ultimately, more information can be conveyed by plotting the results. Due to the long length of the genome and the high-dimensionality of the space of possible sequences on which the posterior distribution $P(C|Y)$ is defined, it is not feasible to visualize $P(C|Y)$. Instead, we find it more informative to focus on posterior results within subsequences of interest. For example, the simplest approach to plotting results for a subsequence of interest is to plot the marginal probabilities, $P(C_{j}|Y)$, versus the genomic location within the subsequence. If we are interested in quantifying the evidence for $C_{ij}$ we could plot the marginal probabilities in a subsequence containing $C_{ij}$. The primary drawback of this approach is that information about the segmental structure, in the form of the sampled segments which reflect the dependence in copy number values between the start and end point of the segment, is lost since we are plotting SNP-specific probabilities. Thus, we should also examine the joint probabilities within the subsequence of interest.

We propose an analogous approach that preserves the segmental structure of the data by plotting the copy number of all sampled segments within the subsequence times their MCMC frequency, $P(C_{ij}|Y)$. That is, we can consider a subregion containing $C_{ij}$ between SNP $i - k$ and $j + k$, and plot the MCMC frequencies for all segments belonging to the set \( \{C_{lm} : l \geq i - k, m \leq j + k; \exists t : C_{l-1}^{(t)} = 2, C_{l}^{(t)} = \ldots = C_{m}^{(t)} \neq 2, C_{m+1}^{(t)} = 2 \} \).

This provides more insight into the plausible structure of the CNV region and reflects the information used to compute $P(C_{ij}|Y)$, the uncertainty measure of segment $C_{ij}$. An example of such a plot is given in Figure 6.1, which we explain in more detail in the next section.
6.2.3 Data analysis of a subsequence

Figure 6.1: Subregion of 16 SNPs containing CNV 5. Left plot, the sampled segments times their MCMC frequency. Right plot, marginal probabilities at each SNP in the subregion.

Table 6.1 gives a summary of the sixteen CNVs that were detected in the analysis of HapMap sample NA18550 in Chapter 5 for BCHMMDmix2. Recall that the DNA copy number was reconstructed by first using the estimated marginal probabilities to assign copy number values to each SNP, then a CNV was defined as a sequence of SNPs with non-two copy values. Two singleton CNVs were ignored. Table 6.1 includes both posterior probabilities that account for uncertainty in the endpoints, \( \hat{P}(C_{ij} | Y) \); and posterior probabilities that do not account for uncertainty in the endpoints, \( \tilde{P}(C_{ij} | Y) \). Next, we do
a further analysis of one of these regions which demonstrates that, alone, the frequency-based probabilities are inadequate summaries.

We examine CNV 5 which did not overlap with any known CNVs and contained four SNPs, spanning 68370 bps. For this CNV, \( P(C_{ij}|Y) = 0.16 \), whereas \( \tilde{P}(C_{ij}|Y) = 0.83 \) with \( k = 1 \). The reason for the discrepancy is readily seen in the left plot of Figure 6.1 which plots each of the sampled segments (across the 5000 MCMC scans) in a region surrounding the original CNV times their MCMC frequency. In this plot, CNV 5, the putative CNV of interest, is highlighted in blue and includes SNP 7 to SNP 10. In addition to this segment, we can see that many of the segments that differ from this one by one or two SNPs also have sizable posterior probabilities, with one segment between SNP 7 and SNP 11 actually having a higher frequency than CNV 5. So, although the frequency of CNV 5 is quite small, the overall evidence that there is a CNV in this region is actually much greater. This is reflected by the large value of \( \tilde{P}(C_{ij}|Y) \). The small probability estimate based on the frequency approach to estimating probability reflects more the uncertainty in the endpoints of CNV 5 rather than the uncertainty that there is a CNV here. Furthermore, as mentioned, one of the overlapping, sampled segments, comprising the five SNPs between SNP 7 and 11, actually had a higher frequency (0.21) than CNV 5 (0.16). Thus, a reconstruction based on joint probabilities yields a slightly more probable segment than the one based on marginal probabilities, which are given in the right panel of Figure 6.1. The right panel shows that in the original DNA copy number reconstruction SNP 11 was not included in the segment since its SNP-specific posterior probability was slightly less than 0.5, despite the fact that the longer segment including SNP 11 is actually more probable. Additionally, the posterior suggests that the CNV residing in this subsequence is likely no larger than 6 SNPs, corresponding to SNP 6 to SNP 11 in the plot, as the probability drops off to zero beyond this segment. Thus, if this region was selected for further validation in the lab, or was to be used as a susceptibility loci in a follow-up study, our recommendation would be to focus on the
genomic information between SNPs 5 and 12.

This underscores the importance of examining such plots as they contain useful information on the segmental structure of the subsequence. In practice, we find that first reconstructing the copy number sequence according to the marginal, SNP-specific, posterior probabilities, then examining plots like Figure 6.1 for each region of interest, is an effective approach for summarizing the MCMC results.

### 6.2.4 Further data analysis of identified CNVs

In this section we compare the CNVs found by BCHMMDmix2 in the analysis of HapMap sample NA18550 in Chapter 5 to those reported in McCarroll et al. (2008). Here, reference to posterior probabilities refers to the posterior probabilities that adjust for uncertainty in the endpoints.

Among the 16 CNVs in Table 6.1, 7 overlapped with CNVs that were found in McCarroll et al. (2008), whose study identified 8 CNVs in total. Among the 7 overlapping CNVs, all had posterior probabilities greater than 0.90 with the exception of one short hemizygous deletion (CNV 13) which was associated with a posterior probability of 0.59. The one CNV that our method did not find which was found in their study was a hemizygous deletion consisting of 6 SNPs. It had an associated posterior probability of $8 \times 10^{-4}$ (with $k = 2$). The mean of the raw CNs for these six SNPs was -0.21, which is larger than the typical mean of 1-copy state observations and actually closer to the mean of the 2-copy state. Given the weak signal, our HMM could not identify this region.

Our method also found 9 CNVs that were not reported in McCarroll et al. (2008), including one long amplification (CNV 1). This amplification consisted of 88 SNPs, spanning 138396 bps, and it was located approximately 0.5 Mb away from the large amplification between 47 and 47.2 Mb that was reported in McCarroll et al. (2008). The additional regions that we found may be rare CNVs or low-frequency CNPs since McCarroll et al. (2008) focused their method on finding CNPs (i.e. inherited CNVs).
6.3 Bayes factor for assessing copy number segments

This section discusses the use of the Bayes factor as a measure of uncertainty for evaluating subsequences. The use of the Bayes factor for assessing CNVs from HMMs was first suggested in Colella et al. (2007). However, they did not provide a closed-form solution, which we now derive for our Bayesian HMM. We note that closed-form solutions are not available for the Bayesian continuous-time HMM due to the non-conjugacy of the parameters of the transition probability matrix. So here we derive Bayes factors for segments based on the conjugate BDHMM, which we described in detail in Chapter 3.

The Bayes factor, defined as the ratio of the probability of the data under the null hypothesis to the probability of the data under the alternative hypothesis, can be used to assess the confidence, or degree of evidence, of a segment of interest. For a given segment $C_{ij}$ the null hypothesis is

$$H_0 : C_i = C_{i+1} = \ldots = C_j = 2$$

versus the alternative hypothesis

$$H_1 : C_i = C_{i+1} = \ldots = C_j = 3.$$  

Therefore, the Bayes factor for candidate segment $C_{ij}$ is defined as

$$BF_{ij} = \frac{P(Y_{ij}|H_1)}{P(Y_{ij}|H_0)},$$

where $Y_{ij} = (Y_i, Y_{i+1}, \ldots, Y_j)$, the observed data in the segment of interest. So a Bayes factor of 1 indicates that the segment is as likely to be a normal segment as an amplification, while larger values of $BF_{ij}$ indicate that the segment is more likely to be an
amplification. For the BDHMM without the outlier component,

$$P(Y_{ij}|H_0) = \int \prod_{t=i}^{j} f(y_t; c_t = 2)p(\mu)p(\sigma^2)p(\Lambda) \, d\mu d\sigma^2 d\Lambda$$

$$= \pi^{-n_{ij}/2} \frac{\sqrt{b_2}}{n_{ij} + b_2} (d_{0,2} s_{0,2}^2)^{d_{0,2}/2} \frac{\Gamma\left(\frac{n_{ij}+d_{0,2}}{2}\right)}{\Gamma(d_{0,2}/2)}$$

$$\times \left[ d_{0,2} s_{0,2}^2 + s_{ij}^2 + \frac{n_{ij} b_2 (\bar{y}_{ij} - m_2)^2}{n_{ij} + b_2} \right]^{-\frac{(n_{ij}+d_{0,2})}{2}},$$

and similarly,

$$P(Y_{ij}|H_1) = \int \prod_{t=i}^{j} f(y_t; c_t = 3)p(\mu)p(\sigma^2)p(\Lambda) \, d\mu d\sigma^2 d\Lambda$$

$$= \pi^{-n_{ij}/2} \frac{\sqrt{b_3}}{n_{ij} + b_3} (d_{0,3} s_{0,3}^2)^{d_{0,3}/2} \frac{\Gamma\left(\frac{n_{ij}+d_{0,3}}{2}\right)}{\Gamma(d_{0,3}/2)}$$

$$\times \left[ d_{0,3} s_{0,3}^2 + s_{ij}^2 + \frac{n_{ij} b_3 (\bar{y}_{ij} - m_3)^2}{n_{ij} + b_3} \right]^{-\frac{(n_{ij}+d_{0,3})}{2}},$$

where $\bar{y}_{ij}$ is the mean of the data in the segment from SNP $i$ to SNP $j$; $n_{ij} = j - i + 1$, the number of SNPs in the segment; and $s_{ij}^2$ is $(n_{ij} - 1)$ times the sample variance of the data in the segment. Solutions for the BDHMM with the outlier component are similar and can be found in the Appendix.

If we assume equivalence between classes of the variance hyperparameters, that is, $d_{0,1} = d_{0,2} = d_{0,3}$ and $s_{0,1}^2 = s_{0,2}^2 = s_{0,3}^2$, then the Bayes factor evaluates to

$$BF_{ij} = \left[ \frac{b_{m_{ij}} (\bar{y}_{ij} - m_2)^2}{n_{ij} + b_2} + s_{ij}^2 + d_{0,2} s_{0,2}^2 \right]^{\frac{n_{ij}+d_{0,1}}{2}}.$$

The greater the value of $B_{ij}$, the more confidence we have that the segment $C_{ij}$ is in fact a CNV. The Bayes factor for segments is largest for segments with values of $\bar{y}_{ij}$ close to $m_3$, and small values of the within-segment sample variance, $s_{ij}^2$. The Bayes factor also favours longer segments, that is, larger values of $n_{ij}$. According to Jeffreys (1961), the strength of evidence is substantial if $BF_{ij} > 3$, strong if $BF_{ij} > 10$, and very strong if $BF_{ij} > 30$. 
The Bayes factor for segments does not depend on the model parameters $\theta$, as these are integrated out, so it is best suited for cases when we do not intend to run the entire MCMC algorithm. For example, when we wish to assess \textit{a priori} regions of interest without having to analyze the whole chromosome. Alternatively, we could use the Bayes factor for segments to specify initial values for $C$ by computing $BF_{i-k-1,i}$ (all segments of length $k$, e.g. $k = 4$) for $i = 4, \ldots, M$, then classifying SNPs as non-two-copy state SNPs if $B_{ij} > 3$. 
## Table 6.1: Summary of the 16 detected CNVs for the analysis of sample NA18550 given in Chapter 5. Length is the length of the CNV in bps, defined as difference between the end position and start position. Percent overlap refers to the number of SNPs in the CNV that overlap with those reported in McCarroll et al. (2008).

| CNV ID | Copy number | Length (bps) | % Overlap | $P(C_{lm}|Y)$ | $P(\hat{C}_{lm}|Y)$ |
|--------|-------------|--------------|-----------|---------------|-------------------|
| 1      | 3.00        | 138396       | 0.00      | 0.79          | 0.98              |
| 2      | 1.00        | 196          | 0.00      | 0.76          | 0.86              |
| 3      | 3.00        | 428698       | 0.87      | 0.04          | 0.99              |
| 4      | 3.00        | 208085       | 0.00      | 0.26          | 0.91              |
| 5      | 3.00        | 68370        | 0.00      | 0.16          | 0.83              |
| 6      | 1.00        | 396          | 0.00      | 0.81          | 0.98              |
| 7      | 1.00        | 364          | 0.00      | 0.65          | 0.97              |
| 8      | 1.00        | 3574         | 0.88      | 0.50          | 1.00              |
| 9      | 1.00        | 486          | 0.80      | 0.74          | 0.90              |
| 10     | 3.00        | 15488        | 0.00      | 0.16          | 0.76              |
| 11     | 0.00        | 6523         | 0.96      | 1.00          | 1.00              |
| 12     | 0.00        | 370          | 0.00      | 0.99          | 1.00              |
| 13     | 1.00        | 11           | 1.00      | 0.40          | 0.59              |
| 14     | 3.00        | 44960        | 0.00      | 0.54          | 0.97              |
| 15     | 0.00        | 12325        | 0.60      | 0.97          | 0.99              |
| 16     | 3.00        | 90828        | 0.59      | 0.32          | 0.98              |
Chapter 7

Discussion and Future Research

This chapter provides a summary and discussion of the research presented in this thesis, as well as ideas and suggestions for future areas of research.

7.1 Summary

This thesis introduced and explored the use of Bayesian HMMs for inferring DNA copy number changes from SNP arrays, and observed important limitations of current implementations of HMMs. In particular, this thesis showed that when accounting for full parameter uncertainty standard implementations of HMMs do not work well for the copy number analysis of real Affymetrix SNP array 6.0 data. Indeed, it is necessary that researchers fix some key model parameters in order to obtain solutions that are biologically plausible. For example, applications of HMMs for copy number data typically fix parameters of the transition probability matrix to control the resolution of the model, often treating these parameters as tuning parameters that are specified by controlling the false discovery rate (e.g., see Colella et al. (2007)). For the continuous-time HMM (CHMM) this would correspond to fixing the $\nu$-parameters, thereby specifying the mean length of CNVs, or the average time that the HMM spends in each hidden state, with large values of $\nu_1$ and $\nu_3$ corresponding to better capability of detecting short CNVs but
at the expense of reporting more false positives, whereas small values of $\nu_1$ and $\nu_3$ would result in a lower number of false positives but at the expense of missing short CNVs. From a statistical perspective, fixing HMM model parameters is not adequate because (i) model parameters may differ between samples and/or chromosomes, and (ii) fixing model parameters to arbitrary values does not properly account for model parameter uncertainty and may result in skewed inferences. In addition, from a biological perspective, the current information on the distribution of lengths of CNVs is incomplete in the sense that only a small number of putative CNVs have been validated, and mean lengths are likely to differ between different arrays and/or studied populations. This thesis proposed a novel Bayesian continuous-time HMM (BCHMM) that provides DNA copy number reconstructions that fully account for model parameter uncertainty, so that one does not need to fix these parameters to arbitrary values.

Although prior information is useful, it is more important to use appropriate models. Previous implementations of HMMs for the analysis of SNP arrays use Normal emission distributions for the non-outlier component of all copy states, which is a convenient choice, and often assume a constant variance across states. However, as was shown in Chapter 3 and 5, for real data the distribution of raw CNs differs among the different copy states. Indeed, the analysis of samples with known DNA copy number of the X chromosome in Chapter 3 showed larger variance for deletions than amplifications. Furthermore, in the analysis of HapMap sample NA18550 in Chapters 3 and 4, we observed several signs that indicated that the conventional models were inadequate. The first sign was the chaotic meandering between the 1- and 2-copy states of the reconstructed copy number sequences. This was due to lack of separation in the posterior distributions between the 1- and 2-copy states, whereby the 1-copy state was acting as a smaller, high-variance diploid state, and the 2-copy state was acting as a larger, low-variance diploid state. Therefore, the posterior distribution indicated that a single Normal emission distribution for the non-outlier component of the diploid state was insufficient. To address this we
proposed a model with a mixture of three distributions for the diploid state: a mixture of two Normal distributions, representing the non-outlier component; and a Uniform distribution, representing the outlier component. The two Normal distributions within the non-outlier component can be considered to be substates within the 2-copy state with different means and variances. This allows the model to account for the heterogeneous group of apparent diploid SNPs that we observed in the data analysis. As was shown in Chapter 5, this modification resulted in substantially improved results: the posterior now distinguished between the 1- and 2-copy states, and led to reconstructed DNA copy number values that did not meander chaotically between the two states.

Still though, a few minor problems persisted: The posterior attributed a very high outlier rate to the 0-copy state and a few outlying, positive valued SNPs had been apparently misclassified to the 0-copy state. Given the high outlier rate and high variance, the 0-copy state was acting as an outlier class for other copy states in addition to true 0-copy state observations, which is inappropriate. Thus, we sought a more outlier-resistant distribution for the non-outlier component of the 0-copy state. Raw CNs in the 0-copy state can be seen to arise as the logarithm of near-zero values, resulting in observed data that is skewed and has an extremely high variance relative to the other states. We proposed using a mixture of an extreme-value type distribution (representing the non-outlier component) and a Uniform distribution (representing the outlier component) for the 0-copy state. As shown in Chapter 5, the extreme value distribution has a higher facility for dealing with skewed data and a wide range of extreme values than the Normal distribution does. This more complicated model lead to a reduced outlier rate, which was much more reasonable than the HMM using Normal emissions, and apparently correct classification of outlying SNPs which had previously been misclassified to the 0-copy state. We note, however, that since the 0-copy state corresponds to the minimum number of copies of DNA, estimated 0-copy state distributions may look more skewed than they actually are due to the bias incurred from the misclassification of some of the higher
valued 0-copy state observations into the 1-copy state. Further work involves comparing this model’s performance to other competitive models besides HMMs.

BCHMM also models the DNA copy number along the genome as a continuous-time Markov process, rather than using a discrete-time process. The primary reason for using the continuous-time model is that it takes into account the distance between SNPs in a model-based framework, making it a more realistic model for copy number data. To summarize the lengths of detected CNVs, previous approaches report the size of each putative CNV as the difference between the start and end position, then compute the mean or median CNV size. In general, this may tend to underestimate the size of CNVs since the true breakpoints will be somewhere between the SNP markers. The BCHMM improves on this framework by directly modelling the sample-level, mean size of each DNA copy number state, allowing for model-based inferences on mean sizes of CNVs.

In general, such model-based, or unified, approaches are preferable since they tend to be more accurate by placing more relative weight on regions that are more likely to be variants. The analysis of HapMap sample NA18550 in Chapter 5 showed fairly good agreement between the inferred mean sizes of CNVs and the reported lengths of known CNVs for the sample of interest. Although our case study only consisted of analyzing data on a single chromosome, one could extend the use of the model to the whole genome so that model parameters, including parameters representing the mean size of CNVs, would then reflect genome-wide values.

Our copy number model is more general than previous approaches that applied continuous-time processes to copy number data since it does not assume that the rate of a certain type of aberration is the same in all states. For example, our model allows for the possibility that the rate of deletions from the diploid state to be larger than the rate of deletions from the amplified state, that is, \( q_{21} > q_{31} \). Biologically, this is plausible as certain regions of the genome are known to be hotspots for structural rearrangements, encompassing higher frequencies of variants. Previous models, such as Wang et al. (2009),
assumed that $q_{21} = q_{31}$ as well as analogous constraints for transitions between the other states. In the data analysis, we compared the BCHMM to a Bayesian version of the more commonly used parsimonious model, which is a submodel of BCHMM with the above constraints on the parameters of the transition intensity matrix. Although the two models produced similar results, the more general model had superior performance when it came to inferring the mean size of CNVs. Also, the submodel missed one CNV that was detected by the more general model. Therefore, our analysis suggested that the assumptions made in the submodel are too restrictive in practice and may not be warranted for real data.

Previous applications of HMMs to copy number data use the Viterbi algorithm based on maximum likelihood estimates of the model parameters to reconstruct the DNA copy number. The primary drawback of this technique is that it does not properly account for model parameter uncertainty since the Viterbi reconstruction is conditional on a single point estimate of the model parameters, $\theta$. Moreover, it does not produce uncertainty measures that can be used to summarize the uncertainty of a putative CNV. Researchers are often interested in such measures so that they can rank detected CNVs, and ultimately, focus further experimental efforts on the most promising regions. This motivated our choice to develop a Bayesian framework which, by using a stochastic version of the forward-backward algorithm, provides a basis on which to develop effective uncertainty measures. Chapter 6 introduced a measure of uncertainty derived from the MCMC algorithm that was designed to account for the uncertainty in the endpoints of the subchromosomal region of interest. Since a high level of uncertainty was observed in the endpoints of these regions, it was important to account for this uncertainty as not doing so would lead to uncertainty estimates that were overly pessimistic.

The Bayesian framework has the further advantage of allowing researchers to incorporate biological information into the model via prior distributions. For example, for well-studied populations we may have good information on mean sizes of deletions and
duplications. Perhaps, it is known that the mean size of deletions is 10 kb, whereas the mean size of amplifications is 40 kb. This information can be readily incorporated into the model through the specification of the prior expected length of each DNA copy state.

7.2 Future Work

7.2.1 Multiple sample analysis

The BCHMM could also be extended to the analysis of multiple samples where we are interested in finding copy number polymorphisms (CNPs) in addition to rare CNVs. Specifically, CNPs are inherited CNVs that occur at a frequency larger than 1% in a given population. Therefore, regions containing CNPs tend to be more variable in copy number than the rest of the genome. We commonly refer to the regions harboring CNPs as polymorphic regions and the surrounding ones as non-polymorphic regions. Researchers are interested in mapping these regions and, ultimately, assessing them for their role in disease susceptibility. Because CNPs are inherited, CNPs in the same polymorphic region will share common start- and end-points. Furthermore, CNP frequencies vary, ranging from 0.01, low-frequency CNPs, to above 0.90, high-frequency CNPs. To find CNPs, typically, researchers use methods for detecting individual-level CNVs, then they look for evidence of significant overlap using (non-statistical) crude approaches. While this may be sufficient for finding common CNPs, a method utilizing information across samples may increase our power of finding low frequency CNPs. We are aware of only one statistical method for finding CNPs. Wang et al. (2009) use a SNP-specific model that fits multiple samples using the EM algorithm. In what follows we provide some ideas on how to extend our HMM to this setting.

Using matrix notation, we denote the observed data as $Y = \{Y_{ij}\}_{1 \leq i \leq N, 1 \leq j \leq M}$ where $i$ denotes the individual and $j$ denotes the SNP being measured. Associated with this is an unobserved matrix of copy number values $C = \{C_{ij}\}_{1 \leq i \leq N, 1 \leq j \leq M}$. We
introduce the latent random variable $Z(t)$ that takes the value 1 if location $t$ is in a region of copy number polymorphism, and 0 otherwise. Since data is sampled at discrete points in space, we define a set of latent variables given by $Z = (Z_1, \ldots, Z_M)$, where

$$Z_i = \begin{cases} 
1, & \text{if SNP } i \text{ is in a CNP}, \\
0, & \text{if SNP } i \text{ is not in a CNP}. 
\end{cases}$$

We can assume that the values of $Z$ are generated according to a continuous-time Markov process with its own transition intensity parameter, $\alpha$. Then the complete data likelihood of the raw CNs for multiple samples, given by $L(\theta) = P(Y, C, Z|\theta)$, is

$$L(\theta) = P(Y|C, \theta_E)P(C|Z, \theta_Q)P(Z|\alpha)$$

$$= P(Z_1) \prod_{i=1}^{N} \{f(y_{i1}|c_{i1})P(C_{i1}|Z_1)\} \prod_{i=1}^{N} \prod_{t=2}^{M} \{f(y_{it}|c_{it})P(C_{i,t}|C_{i,t-1}Z_t, Z_{t-1})\}$$

$$\times \prod_{t=2}^{M} T^{Z}_{z_{t-1},z_{t}}(\Delta t),$$

where $T^{Z}_{i,j}(d)$ is the $ij$th element of the transition probability matrix at a distance $d$ for the latent variable $Z$. The model assumes the conditional independence, given the copy number, of raw CN values between adjacent SNPs and across samples. For each sample, the likelihood contribution of the first SNP is the product of the emission density of the raw CN, the initial probability of the copy number and the initial probability of the latent variable $Z_1$. The likelihood contribution of the subsequent SNPs is the product of the emission density of the raw CN, the transition probability of the copy number between adjacent SNPs and the transition probability of the latent variable $Z$ between adjacent SNPs. The goal here is to infer both the individual-level copy number $C$ and the population-level CNP status $Z$. Notice that the transition probabilities of the copy number states now depend on not only the copy number at the previous SNP, but also the value of $Z$ at the current and previous SNPs. For example, when $Z_{j-1} = 0$ and $Z_j = 1$, there is a breakpoint between SNP $j - 1$ and SNP $j$, marking the beginning of a polymorphic region. Thus, we can treat the copy number state at SNP $j$ as being
independent of the copy number at SNP $j - 1$. Namely,

$$P(C_{i,j} = k | C_{i,j-1} = l, Z_j = 1, Z_{j-1} = 0) = P(C_{i,j} = k | Z_j = 1, Z_{j-1} = 0).$$

Then $P(C_{i,j} = k | Z_j = 1, Z_{j-1} = 0) = \lambda_{1k}$ is the frequency of copy number state $k$ in a CNP. We can also allow for CNP-specific frequencies.

Similarly, when $Z_{j-1} = 1$ and $Z_j = 0$ there is a breakpoint between SNP $j - 1$ and SNP $j$, marking the beginning of a non-polymorphic region. Thus, we can treat the copy number state at SNP $j$ as being independent of the copy number at SNP $j - 1$. Namely,

$$P(C_{i,j} = k | C_{i,j-1} = l, Z_j = 0, Z_{j-1} = 1) = P(C_{i,j} = k | Z_j = 0, Z_{j-1} = 1).$$

Then $P(C_{i,j} = k | Z_j = 0, Z_{j-1} = 1) = \lambda_{0k}$ is the frequency of copy number state $k$ in a non-polymorphic region. We refer to these frequencies as the background frequencies. Since CNPs are characterized by higher frequencies of deletions and duplications than the rest of the genome, $\lambda_{11}$, the frequency of deletions in CNPs, and $\lambda_{13}$, the frequency of amplifications in CNPs, will tend to be larger than the corresponding background frequencies, $\lambda_{01}$ and $\lambda_{03}$. To simplify the problem, we could fix the parameters of $\Lambda_0$ (e.g. $\lambda_{02} = 0.995$) since the vast majority of the genome is background noise and de novo CNVs are very rare.

When both SNPs are not in a region of CNP, that is $Z_t = 0$ and $Z_{t-1} = 0$, we can determine the transition probabilities by assuming that the copy number follows a different continuous-time Markov process whose parameters represent the de novo rate of CNVs, so that the model has separate rates for de novo and inherited CNVs.

Finally, when both SNPs are in a region of CNP, that is $Z_t = 1$ and $Z_{t-1} = 1$, the transition probabilities are given by

$$P(C_{i,t} = j | C_{i,t-1} = k, Z_t = 1, Z_{t-1} = 1) = \begin{cases} 1, & \text{if } k = j \\ 0, & \text{otherwise.} \end{cases}$$

This enforces shared boundaries of CNPs in the same polymorphic region.
Then, the detailed complete data likelihood is

\[
L(\theta) = P(Z_1) \prod_{i=1}^{N} \{ f(y_{i1}|c_{i1})P(C_{i1}|Z_1) \} \prod_{i=1}^{N} \prod_{t=2}^{M} \{ f(y_{it}|c_{it}) \} \prod_{i=1}^{N} \prod_{t=2}^{M} \{ f(y_{it}|c_{it}) \} = \prod_{i=1}^{N} \prod_{t=2}^{M} \{ f(y_{it}|c_{it}) \}
\]

(7.1)

The Bayesian approach, although computationally intensive, is more straightforward to implement. Starting from an initial value \( \theta(0) \) and \( C(0) \), for \( t = 0, 1, \ldots \), iteratively sample from the conditional posterior distributions:

1. Sample \( Z(t) \sim P(Z|Y, C(t), \theta(t)) \) (stochastic backward sampling),

2. Sample \( C(t+1) \sim P(C|Y, Z(t), \theta(t)) \) (stochastic backward sampling),

3. Sample \( \theta(t+1) \sim P(\theta|Y, Z(t), C(t+1)) \).

### 7.2.2 Family trio data

This type of model can help us gain insight into which CNVs are more likely to be inherited (i.e., those in polymorphic regions) and which CNVs are more likely to be de novo, or non-inherited. Of course, family trio data, where we have copy number data from both parents and their offspring, are better suited to distinguish between these types of CNVs. The multiple sample BCHMM could be modified to account for the correlation structure inherent in family trio data by allowing the transition probabilities of the offspring to depend on the copy number status of both its parents. For example, say we have data along the genome for a single parent-offspring trio, which we denote by \( Y_j = (Y_{j,o}, Y_{j,m}, Y_{j,f}) \), where \( Y_{j,o} \) is the raw CN of the offspring at SNP \( j \), \( Y_{j,m} \) is the raw CN of the mother at SNP \( j \), and \( Y_{j,f} \) is the raw CN of the father at SNP \( j \). Similarly, there exist unobserved copy number states for each member of the trio, which we denote \( C_j = (C_{j,o}, C_{j,m}, C_{j,f}) \) for SNP \( j \). We can utilize the biological relationships between
family members to infer whether or not the CNVs of the offspring were inherited from their parents. Inherited CNVs are much more common than de novo CNVs. In order to distinguish between inherited CNVs and de novo CNVs of the offspring, we redefine the variable $Z_j$ to take the value 1 for de novo CNVs and 0 for inherited CNVs. We will refer to this variable as the de novo status of the offspring.

Then, we can reexpress the transition probabilities as

$$P(C_j|C_{j-1}, Z_j, Z_{j-1}) = P(C_{j,o}|C_{j,m}, C_{j,f}, C_{j-1,o}, C_{j-1,m}, C_{j-1,f}, Z_j, Z_{j-1})$$

$$\times P(C_{j,m}, C_{j,f}|C_{j-1,o}, C_{j-1,m}, C_{j-1,f})$$

$$= P(C_{j,o}|C_{j,m}, C_{j,f}, C_{j-1,o}, Z_j, Z_{j-1})P(C_{j,m}|C_{j-1,m})P(C_{j,f}|C_{j-1,f}),$$

where we have made use of the fact that the copy number of the mother is independent of the copy number of the father, and conditional on the copy number of the parents and the copy number of the offspring at the previous SNP, the copy number of the offspring is independent of the copy number of the parents at the previous SNP. In the following sections, we describe the computation of these probabilities.

**Transition probabilities of the copy number states of the offspring**

The transition probabilities of the offspring depend on both the copy number of the parents and the de novo status of the offspring at the current and previous SNPs, as well as the copy number of the offspring at the previous SNP. Thus, we examine the transition probabilities for each combination of $Z_j$ and $Z_{j-1}$. When both SNPs are de novo, that is $Z_j = 1$ and $Z_{j-1} = 1$, then

$$P(C_{j,o}|C_{j,m}, C_{j,f}, C_{j-1,o}, Z_j = 1, Z_{j-1} = 1) = P(C_{j,o}|C_{j-1,o}, Z_j = 1, Z_{j-1} = 1),$$

so that the copy number of the offspring does not depend on the copy number of its parents. We can determine these probabilities by assuming a different continuous-time Markov process on the copy number of the offspring whose parameters, for the case when $Z_j = 1$ and $Z_{j-1} = 1$, would represent de novo rates of aberration.
If both SNPs are inherited, that is \( Z_j = 0 \) and \( Z_{j-1} = 0 \), then the copy number of the offspring is directly determined from the copy number of its parents according to the laws of Mendelian inheritance. We omit the mathematical details here.

If SNP \( j-1 \) is inherited and SNP \( j \) is de novo, that is \( Z_j = 0 \) and \( Z_{j-1} = 1 \), there is a breakpoint between the SNPs and we can assume that the copy number at SNP \( j \) is independent of the copy number at SNP \( j-1 \). Furthermore, since SNP \( j \) is not inherited, the copy number does not depend on the parents. Thus,

\[
P(C_{j,o}|C_{j,m}, C_{j,f}, Z_{j-1} = 0) = P(C_{j,o}|Z_j = 1, Z_{j-1} = 0).
\]

We can denote the unknown probability of the de novo event by \( \lambda_{1,k} = P(C_{j,o} = k|Z_j = 1, Z_{j-1} = 0) \) for \( k = 1, 2, 3 \).

Finally, if SNP \( j-1 \) is de novo and SNP \( j \) is inherited, again, there is a breakpoint between the SNPs and we can assume that the copy number at SNP \( j \) is independent of the copy number at SNP \( j-1 \). Thus,

\[
P(C_{j,o}|C_{j,m}, C_{j,f}, Z_{j-1} = 1) = P(C_{j,o}|C_{j,m}, C_{j,f}, Z_j = 0, Z_{j-1} = 1),
\]

where \( P(C_{j,o}|C_{j,m}, C_{j,f}, Z_j = 0, Z_{j-1} = 1) \) is evaluated based on the model for Mendelian inheritance.

Without going into the details, this model could be implemented with an MCMC algorithm. Letting \( C_o = (C_{1,o}, \ldots, C_{M,o}) \), \( C_f = (C_{1,f}, \ldots, C_{M,f}) \) and \( C_m = (C_{1,m}, \ldots, C_{M,m}) \), and starting from an initial value \( \theta^{(0)} \) and \( C_o^{(0)} \), for \( t = 0, 1, \ldots \), iteratively sample from the conditional posterior distributions:

1. Sample

\[
C_m^{(t)} \sim P(C_m|Y_m, \theta^{(t)}) \quad \text{(stochastic backward sampling)},
\]

\[
C_f^{(t)} \sim P(C_f|Y_f, \theta^{(t)}),
\]

2. Sample \( Z^{(t)} \sim P(Z|Y, C_o^{(t)}, C_m^{(t)}, C_f^{(t)}, \theta^{(t)}) \) (stochastic backward sampling),
3. Sample $C_o(t^{(t+1)}) \sim P(C_o|Y, Z(t), C_m(t), C_f(t), \theta(t))$ (stochastic backward sampling),

4. Sample $\theta(t^{(t+1)}) \sim P(\theta|Y, Z(t), C_o^{(t^{(t+1)})}, C_m(t), C_f(t))$.

Future work involves implementing this algorithm, analyzing real data in order to assess the improvement gained by using such a model over the standard approach, which is to apply single-sample methods to each member of the family trio separately, then compare the reconstructed DNA copy numbers between trio members, looking for evidence of overlap of CNVs. We suspect our model-based approach will improve our power of not only finding short CNVs, but also distinguishing between whether they are inherited or de novo CNVs. Additionally, such a model-based approach would reduce the need of the researcher to make subjective decisions about what constitutes significant overlap.
Appendix A

Sampling $\nu$ for the submodel

In Chapter 4 we introduced a Bayesian continuous-time HMM, referred to as BCHMM, as well as a submodel of the $Q$-matrix for this model. Here we provide the details of the Metropolis-Hastings algorithm for sampling the $\nu$-parameters of the transition probability matrix for the submodel.

For the submodel, sampling the $\nu$-parameters is more difficult because of the constraints imposed on these parameters and the fact that each parameter appears in more elements of the transition probability matrix. The constraints imposed on the parameters of the submodel are:

\begin{align*}
\nu_1 &< \nu_2 + \nu_3, \\
\nu_2 &< \nu_1 + \nu_3, \\
\nu_3 &< \nu_1 + \nu_2.
\end{align*} \quad (A.1)

\begin{align*}
\nu_1 &< \nu_2 + \nu_3, \\
\nu_2 &< \nu_1 + \nu_3, \\
\nu_3 &< \nu_1 + \nu_2. \\
\end{align*} \quad (A.2)

- For $c = \{1, 2, 3\}$, given the data, other model parameters and the most recently drawn value of $\nu_c$, denoted by $\nu_c^{(t-1)}$, sample $\nu_c$ using a Metropolis-Hastings algorithm:

  1. Sample a proposal $\nu_c^{(t)}$ from a $\text{Gamma}(1, \nu_c^{(t-1)})$ distribution. This proposal is centered around the most recently drawn value of $\nu_c$. 

123
2. If \( \nu^{(t)}_c \) satisfies the constraints given in A.1, accept \( \nu^{(t)}_c \) with probability

\[
\min \left\{ \frac{p(\nu^{(t)}_c | \sim) \Gamma(\nu^{(t-1)}_c; 1, \nu^{(t)}_c)}{p(\nu^{(t-1)}_c | \sim) \Gamma(\nu^{(t)}_c; 1, \nu^{(t-1)}_c)}, 1 \right\},
\]

where

\[
p(\nu_c | \sim) \propto \nu_c e^{-\nu_c (l_c + \sum_{j \in Acc} \Delta_j)} \prod_{i=2}^M (T_{C_{i-1}, C_i}(\Delta_i))^{(1-I(C_{i-1} = C_i))},
\]

where \( T(\Delta_i) \) is specified in Section 4.2.3, \( \Gamma(\nu_c; a, b) \) is the density function for the Gamma\((a, b)\) distribution, and \( Acc = \{ j : C_j = C_{j-1} = c \} \) for \( c = 1, 2, 3 \). Otherwise, reject \( \nu^{(t)}_c \).

To avoid computational underflows, which result due to the product of a large number of probabilities, we actually perform the computations on the log-scale.
Appendix B

In this section we prove equation 6.2 in Chapter 6, which expresses the posterior probability of a subsequence, \( P(C_{ij}|Y, \theta) \), in terms of the forward and backward variables, allowing for straightforward computations. In the following, for convenience we drop \( \theta \) from the notation.

\[
P(C_{ij}|Y) = P(Y|C_{ij}) P(C_{ij}) / P(Y)
\]

\[
= P(Y_{1:i-1}|C_{i-1} = 2) P(Y_{i:j+1}|C_{i:j+1}) P(Y_{j+2:M}|C_{j+1} = 2) P(C_{ij}) / P(Y)
\]

\[
= \frac{\alpha_{i-1}(2)}{P(C_{i-1} = 2)} \left( \prod_{k=1}^{j+1} f(Y_k; C_k) \right) \beta_{j+1}(2) P(C_{ij}) / P(Y)
\]

\[
= \alpha_{i-1}(2) \left( \prod_{k=1}^{j+1} f(Y_k; C_k) T_{C_{k-1},C_k}(\Delta_k) \right) \beta_{j+1}(2) / P(Y)
\]

where the first step follows from Bayes theorem, the second step follows from the properties of the Markov chain (Figure 4.2 may help), and the third step follows from conditional independence and the definitions of the forward and backward variables. Finally, \( P(Y) = \sum_{i=1}^{K} \alpha_M(i) \).
Appendix C

Approximate E-step using variational inference

Variational methods (Jordan et al. 1999) are a relatively new technique for inference in probabilistic models. In the context of HMMs, they can be used as an approximating framework whereby an approximation for \( P(C|Y) \) is provided by a simpler distribution \( Q(C) \). This is useful in more complex HMMs where \( P(C|Y) \) has an intractable form and direct application of the standard algorithms for parameter estimation are infeasible. More specifically, we often need to marginalize over the hidden variables \( C \) to compute expectations. For simpler models the FB algorithm, which takes advantage of the simple structure of the model, can be used to compute these expectations. However, for more complicated models summation over an exponentially large number of terms is unavoidable. Variational inference provides an approximation by making use of the following inequality on the log likelihood

\[
\log L(\theta) = \log \sum_C P(C, Y) = \log \sum_C Q(C) \left( \frac{P(C, Y)}{Q(C)} \right) \geq \sum_C Q(C) \log \left( \frac{P(C, Y)}{Q(C)} \right),
\]

where the last step follows from Jensen’s inequality and \( Q \) is any distribution. Taking the difference between the left- and right-hand side of this inequality yields the Kullback-
Leibler divergence between $Q$ and $P$:

$$KL(Q||P) = \sum_{C} Q(C) \log \frac{Q(C)}{P(C|Y)}.$$  \hspace{1cm} (C.2)

The distribution $Q$ is called the variational approximation and its parameters are called the variational parameters. Again, the idea is to pick a distribution $Q$ that has a tractable structure while at the same time providing a tight bound so as to minimize the KL divergence. (We discuss various choices for $Q$ in a subsequent chapter.)

In sum, the approach to approximating the E-step consists of minimizing equation C.2 with respect to the variational parameters. The needed expectations in the E-step, which were previously intractable, can now be computed as a simple function of the estimated variational parameters.

We now apply this technique to the hierarchical HMM for finding CNPs which was proposed in Chapter 7. In this case, we are required to maximize the bound on the likelihood given in equation C.1, which for the hierarchical HMM is

$$\log(P(Y|\theta)) \geq E_{Q}(\log P(Y|C)) + E_{Q}(\log P(C|Z)) + E_{Q}(\log P(Z))$$  \hspace{1cm} (C.3)

$$- E_{Q}(\log Q(C, Z)),$$  \hspace{1cm} (C.4)

where $Q$ is the approximating distribution (i.e. variational approximation) and the expectations are taken with respect to $Q$. $P(Y|C)$, $P(C|Z)$ and $P(Z)$ are all given in equation 7.1. In the next two sections we consider approximating distributions that factor the two levels of missing copy number data so that $Q(C, Z) = Q_{1}(C)Q_{2}(Z)$. 

Completely factorized variational approximation

The completely factorized distribution assumes conditional independence of the latent variables given the data $Y$, namely

$$
Q_1(C|h^C) = \prod_{i=1}^{N} \prod_{t=1}^{M} Q_1(C_{it}|h^C_{it}), \quad (C.5)
$$

$$
Q_2(Z|h^Z) = \prod_{t=1}^{M} Q_2(Z_{it}|h^Z_{it}), \quad (C.6)
$$

where the individual distributions are Bernoulli distributions: $Q_1(C_{it}|h^C_{it}) = \prod_{k=1}^{3} \phi(k h^C_{itk})$ and $Q_2(Z_{it}|h^Z_{it}) = (h^Z_{it})^{Z_{it}} (1 - h^Z_{it})^{1 - Z_{it}}$, where $C_{itk} \in \{0, 1\}$ and $\sum_{k=1}^{3} C_{itk} = 1$. The use of this approximation was motivated by Ghahramani & Jordan (1997) who used this approximation for fitting factorial hidden Markov models. The variational parameters $\{h^C_{it}\}$ and $\{h^Z_{it}\}$ are the means of the DNA copy number and copy number polymorphism status values, respectively. Thus, since all the hidden variables are discrete each parameter represents the probability of the variable taking on each of its possible values. So, for example, $h^C_{it1}$ is the probability that sample $i$ has a copy number deletion at SNP $t$, and $h^Z_{it}$ is the probability that SNP $t$ is in a polymorphic region.

We obtain a set of fixed point solutions for these parameters by taking the derivatives of $C.3$ with respect to the variational parameters $h^C_{itk}$ and setting these equal to zero (see the Appendix for complete derivations). This gives

$$
\begin{align*}
h^C_{it}^{\text{new}} &= \Phi(Y_{it} \mu / \sigma^2 - 0.5 \mu^2 / \sigma^2 + (1 - h^Z_{it})(1 - h^Z_{i,t-1})(\log T_C(\Delta_i)) h^C_{i,t-1} \\
&\quad + (1 - h^Z_{i,t+1})(1 - h^Z_{it})(\log T_C(\Delta_{t+1}))' h^C_{i,t+1} \\
&\quad + h^Z_{it} h^Z_{i,t-1}(\log(I_3)) h^C_{i,t-1} + h^Z_{i,t+1} h^Z_{it}(\log(I_3))' h^C_{i,t+1} \\
&\quad + h^Z_{it} (1 - h^Z_{i,t-1}) \log(\lambda_1) + (1 - h^Z_{it}) h^Z_{i,t-1} \log(\lambda_0),
\end{align*}
$$

where $\Phi$ is the softmax operator which ensures that the elements of the vector sum to one. Similarly, taking the derivatives of $C.3$ with respect to the variational parameters
and setting these equal to zero, we obtain the following fixed point solution:

\[
\text{logit}(h_t^{Z_{\text{new}}}) = (1 - h_{t-1}^Z) \left[ \sum_{i=1}^{N} (h_{i,t}^{C'} \log \lambda_1 - \text{tr}\{h_{i,t-1}^{C} h_{i,t}^{C'} \log T_C(\Delta_t)\}) \right] \\
+ \log T_{01}^{Z}(\Delta_t) - \log T_{00}^{Z}(\Delta_t)
\]

\[
(1 - h_{t+1}^Z) \left[ \sum_{i=1}^{N} (h_{i,t+1}^{C'} \log \lambda_0 - \text{tr}\{h_{i,t}^{C} h_{i,t+1}^{C'} \log T_C(\Delta_{t+1})\}) \right] \\
+ \log T_{10}^{Z}(\Delta_{t+1}) - \log T_{00}^{Z}(\Delta_{t+1})
\]

\[
h_{t-1}^{Z} \left[ \sum_{i=1}^{N} (\text{tr}\{h_{i,t-1}^{C} h_{i,t}^{C'} \log I_3\} - h_{i,t}^{C'} \log \lambda_0) + \log T_{11}^{Z}(\Delta_t) - \log T_{10}^{Z}(\Delta_t) \right]
\]

\[
h_{t+1}^{Z} \left[ \sum_{i=1}^{N} (\text{tr}\{h_{i,t}^{C} h_{i,t+1}^{C'} \log I_3\} - h_{i,t+1}^{C'} \log \lambda_1) + \log T_{11}^{Z}(\Delta_{t+1}) - \log T_{01}^{Z}(\Delta_{t+1}) \right] .
\]

Despite the independence structure of the approximation distribution, the terms comprising \(h_{it}^{C_{\text{new}}}\) still preserves the Markov structure of the HMM.


Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter, N. M., Scherer,

Markov models approach to the analysis of array CGH data’, *Journal of Multivariate 

second edn, Chapman and Hall.


Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, 
R. J., Freedman, B. I., Quinones, M. P., Bamshad, M. J., Murthy, K. K., Rovin, B. H., 
Bradley, W., Clark, R. A., Anderson, S. A., O’Connell, R. J., Agan, B. K., Ahuja, 
CCL3L1 gene-containing segmental duplicaitons on HIV-1/AIDS susceptibility’, *Sci- 

Griffiths, A. J., Wessler, S. R., Lewontin, R. C., Gelbart, W. M., Suzuki, D. T. & Miller, 

Gu, W., Zhang, F. & Lupski, J. R. (2008), ‘Mechanisms for human genomic rearrange-
ments’, *Pathogenetics* **1**(4).


and population-genetic analysis of SNPs and copy number variation', *Nature Genetics* **40**(10), 1166–1174.


resolution copy number variation detection in whole-genome SNP genotyping data’, *Genome Research* 17, 1665–1674.
