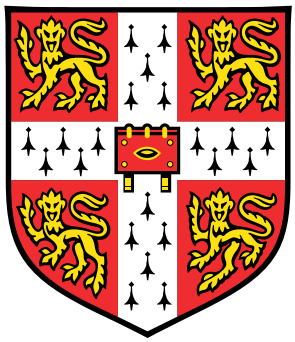
Variational Mixture Models for non-Gaussian observations: Applications to molecular data



Stavroula D. Gerontogianni

Department of Medical Genetics

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Homerton College September 2021

"Nothing takes place in the world whose meaning is not that of some maximum or minimum."

Leonhard Euler (1707-1783).

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Stavroula D. Gerontogianni

September 2021

Abstract

Epigenetics is the field of biology that studies the changes in organisms due to alteration of gene expression rather than modification of the DNA sequence itself. DNA methylation is a well-studied type of epigenetic change, which results in gene silencing and can be dangerous when occurs at tumor suppressor gene loci. Many techniques have been developed to map the methylation pattern of individuals at several genetic loci, such as the HumanMethylation450 BeadChip, the EPIC BeadChip and the whole-genome bisulfite sequencing. Each of these DNA profiling platforms quantifies methylation occurrence in different ways, either continuously (rates of methylation intensity) or discretely (counts of methylated reads). Identifying subgroups of individuals with similar methylation patterns, as well as those genetic loci that discriminate the subgroups, is a crucial procedure that helps linking diseases to specific methylation patterns. Clustering analysis and posterior feature selection of the most important genetic loci that discriminate each subgroup of individuals are the two tools we suggest for achieving this venture. Clustering DNA methylation data though is not a trivial procedure since they are platform-specific and not normally distributed.

In this thesis, we propose clustering DNA methylation data based on the data type (continuous or discrete) by fast model-based clustering methods, while we select the most important/discriminatory genetic loci by an *a posteriori* feature selection measure. Specifically, we apply variational non-Gaussian Dirichlet Process mixture models because they have infinite number of components that allow model-determination and are flexible to model any discrete or continuous data type. We also employ Variational Inference with the "annealing" extension that accounts for poor initialization of the algorithm, due to its high speed in estimating the model parameters and its scalability to high-dimensional data. Our real applications on neonatal DNA methylation data measured in three different ways show that the discrete data types - number of aberrantly methylated genetic loci (counts) and whether a genetic locus is abnormally methylated or not (binary) - can be more informative than its continuous version (intensity of methylation *per* genetic locus) for revealing the association of artificial conception with the predisposition of developmental disorders.

Acknowledgements

First of all, I would like to thank my supervisor Leonardo Bottolo who gave me this amazing opportunity to be part of the incredible academic environment of the University of Cambridge. His persistence, perfectionism and constant attention contributed radically in the completion of this doctoral research, thus I am grateful to him. A great thank you goes as well to my examiners Prof. Marc Chadeau-Hyam and Dr. Paul Lyons for their invaluable contribution in meticulously reading and correcting my thesis.

Subsequently, I need to express my gratitude to Petros Dellaportas and Nikos Demiris. Petros played an important role during my postgraduate studies by offering me his unconditional support and guidance. Nikos was the first academic person who believed in me by inspiring me to pursue this postgraduate route. Both of them have been the initial motive to kick off my PhD journey.

Next, I would like to thank the Alan Turing Institute and its fantastic faculty who never stop prioritising students' needs. I feel privileged to have done my PhD at this amazing place. In this institute, I also had the chance to meet and collaborate with wonderful people, such as Deniz and Andreas. Deniz was like an unofficial advisor to me by trying to help me in difficult scientific moments offering his vast source of knowledge. Andreas, as the coolest person he is, managed to turn work into a fun game and I cannot wait to carry on with our research collaboration.

From the university, I would like to thank Eguz Ochoa for offering me her real datasets for my analysis. I am grateful to her for working really hard to explain from a biological perspective a lot of the results in this thesis. Without her a significant segment of the real analysis interpretation would have been a long shot. Furthermore, I could not forget Amanda Goldsmith who played her best role as PhD administrator by always trying to solve all of my bureaucratic questions.

Regarding my precious family, there are no words to express my deepest gratitude. My father Dionisis, my mother Roula and my brother Dimitris have been my rock all my life and nothing would have been accomplished without their limitless love and support.

I cannot thank them enough for being life mentors and for offering me the sentimental and physical tools to achieve my goals.

As for my dearest friends, I owe a big thank you to Aglina for always being there for me and supporting me at any moment of my life for almost a decade now. Moreover, I would like to thank Albesa for being a beloved friend and the best PhD buddy who would always stick up for me. Additionally, I could not leave behind Dimitra for being a sweet friend and also the inspiration for entering the field of Statistics.

Last but not least, I am deeply thankful to Ares. Without him this PhD journey would have been a rutty path to take. His intelligence, optimism, genuine love and support made these years feel like a pleasant adventure.

Contents

List of Figures

List of Tables

List of Algorithms

Nomenclature

Acronyms / Abbreviations

CA: Coordinate Ascent

DAG: Directed Acyclic Graph

DP: Dirichlet Process

ELBO: Evidence Lower BOund

EM: Expectation Maximization

MCMC: Markov Chain Monte Carlo

PCA: Principal Component Analysis

PG: Pólya Gamma

RJMCMC: Reversible Jump Markov Chain Monte Carlo

VB-DPBerM: Variational Bayes - Dirichlet Process Bernoulli Mixtures

VB-DPBM: Variational Bayes - Dirichlet Process Beta Mixtures

VB-DPGM: Variational Bayes - Dirichlet Process Gaussian Mixtures

VB-DPPM: Variational Bayes - Dirichlet Process Poisson Mixtures

VB: Variational Bayes

VI: Variational Inference

Introduction

There are several conundrums around the biology of life that seem inconceivable but in reality they do have a reasonable explanation. For example, imagine of having two identical twins with exactly the same genetic code and one of them develops a disease that has a genetic component, while the other one is healthy. How is this possible since both have the same genetic code?

This question can be answered based on a similar example studied in Waterland and Jirtle [145]. In this example, we consider two twin mice with the same genetic material, however one is of average size and brown in colour, while the other one is yellow and obese. How is this feasible? The answer is that a gene called the *agouti* gene, which is a colour- and obesity-associated gene in mice, as well as probably related to specific diseases, is actively expressed in the yellow mouse but silenced in its twin brown mouse because it is methylated. DNA methylation is a heritable biological process that changes the expression of genes rather than modifying the genetic code itself. Epigenetic alterations (modification on top or around the genetic code) can happen during the differentiation of somatic cells. These alterations are then passed on to the descendants resulting in a phenomenon called epigenetic inheritance (Lind and Spagopoulou [77]). For example, Cytosine methylation is a process where methyl groups bind onto specific DNA segments changing the way the gene is read. Environmental influences and lifestyle factors such as smoking, diet or physical activity could be responsible for triggering DNA methylation (Lim and Song [74]).

In the mice example, methylation on the *agouti* gene results to a normal mouse, while demethylation to an abnormal mouse. On the other hand, when DNA methylation occurs at promoter regions of tumor suppressor genes the result can be human cancer (Esteller [40]). Therefore, there is great need in studying and analyzing methylation patterns in genes that are associated to deleterious diseases such as atherosclerosis (Dong et al. [35]) or rare developmental disorders like the Beckwith-Wiedemann Syndrome (Weksberg et al. [147]). For instance, atherosclerosis is an important inflammatory process that impairs the quality of life in the aging population. Hence, it is crucial to possess the knowledge around its formation mechanism for better prevention strategies

and treatment productions. Additionally, there is high demand in identifying the most important of the disease-associated genes that when abnormally methylated a disease predisposition could occur.

Analyzing DNA methylation data with the aim of discovering methylation patterns, for instance differences in a case/control study or clustering of healthy controls or group of patients along with the most salient genetic regions where these differences occur, is a rather challenging procedure. The reason is two-fold: a) the data produced by the DNA profiling platforms are non-normally distributed and b) each platform measures in a different way the methylation level. Thus, DNA methylation data are non-Gaussian and platform-specific. For example, Illumina Infinium HumanMethylation450 BeadChip platform (450K) and EPIC BeabChip derive methylation rates (beta-intensities *per* genetic locus), while whole-genome bisulfite sequencing (WGBS) methylation counts (methylated reads *per* genetic locus). For the non-Gaussian part although, there are mathematical functions like *log* or *logit* that usually conform the data to normality (*i.e.,* M-values in Illumina Infinium arrays). Nonetheless, data transformation may not always be the solution to ensure normality and thus the need to retain the original data space could be in demand.

In this thesis, we overcome the challenges of subgroups identification analysis by providing novel and fast model-based clustering tools for methylation data of bounded continuous and discrete type. In particular, we identify subgroups of individuals with similar methylation patterns through non-Gaussian Dirichlet Process mixture models estimated by variational algorithms. For instance, we propose modelling data from 450K or EPIC by the variational Dirichlet Process Beta mixture and data from WGBS by the variational Dirichlet Process Binomial mixture. Moreover, we propose doing posterior selection of the most important genetic loci *per* subgroup of individuals. Specifically, we exploit an *a posteriori* measure that indicates the genetic loci that are the most important for segregating individuals by methylation patterns. This key information is useful for discriminating genetic loci in three categories: 1) important for all the subgroups of methylation patterns, 2) not important for any or 3) exclusively important for a specific subgroup.

Overall, the reason we work on Dirichlet Process mixtures is due to their infinite number of components that allows model-determination and to their flexibility in choosing the appropriate discrete or continuous distribution according to the platform-specific methylation measure type. In addition, the motive for employing variational algorithms for inference is because the commonly used Markov chain Monte Carlo (MCMC) sampling algorithms are non-scalable in large datasets. For example, datasets produced by the EPIC platform have more than 800K genetic loci reported for each individual. MCMC could not scale on this example due to the repeated evaluations

of the likelihood function at each iteration. In cases of considerably lower dimensions, *i.e.*, 50 or 100 genetic loci, MCMC would also take several days to converge. On the other hand, Variational Inference is a family of optimization algorithms that scale well to high-dimensional data and provide fast parameter estimates, considering that optimization of a posterior distribution is a faster procedure than sampling from it.

This thesis is structured in six chapters. In Chapter 1, we briefly review the basic concepts of the molecular biology. The intention is to build a concise knowledge around the mechanism action of DNA methylation and the tools to quantify it, facilitating the implementation of the statistical analysis.

In Chapter 2, we present the theory of Variational Inference, as well as we derive popular variational regression models to assist the understanding around the derivation of more complex models such as those of our interest: the variational Dirichlet Process mixture models. We also explain the advantage of Dirichlet Process mixtures (infinite number of components) over the Finite mixtures (fixed and specified number of components) as model-determination tools. Moreover, we introduce a way to deal with poor initialization of the variational algorithm, called "annealing". Finally, we describe the *a posteriori* feature selection measure to detect discriminatory subgroups features, *i.e.*, CpG sites, Differentially Methylated Regions (DMRs) etc.

With regards to Chapter 3, we provide the full mathematical derivation of a substantial variety of variational Dirichlet Process mixture models as well as some variational Finite mixture models for reader's reference. We also introduce appropriate models when confounding parameters such as age, sex, ethnicity etc. could contaminate the clustering results. In a nutshell, all the models of this chapter can be used for clustering DNA methylation data of discrete or continuous type given the DNA profiling platform and the existence or non existence of confounding parameters. The provision of specialized models for each type of methylation measurement and discrete models that can take into account covariates/confounders is a novel contribution in the area of DNA methylation clustering analysis.

Regarding Chapter 4, we create synthetic scenarios and assess the clustering performance of three variational Dirichlet Process mixtures that are utilized for the real applications in Chapter 5. These are the variational Dirichlet Process Beta, Poisson and Bernoulli mixtures. We also compare their performance to the most commonly used non-probabilistic clustering algorithms (K-means, Hierarchical clustering and DBSCAN).

In Chapter 5, we perform clustering analysis on real data. More precisely, we analyze DNA methylation measured in three different ways in a dataset of artificially and naturally conceived neonates with and without a rare developmental disorder. The aim

is to find which of the three measures of methylation is more informative in revealing the association of artificial conception with the predisposition of developmental disorders, based on the recorded DNA methylation on genetic loci that are parental-specific (imprinted DMRs). The dimensions of this dataset of neonates is relatively low (228 neonates $\times$ 33 imprinted DMRs), however we prefer using Variational Inference over MCMC since it completes the inference in seconds, whereas MCMC requires days to converge.

Finally, in Chapter 6, we summarise the conclusions of the simulation tests and the real data analysis, discuss the contribution of the proposed models and finish with directions for future research.

Chapter 1

Overview of DNA Methylation

1.1 Molecular Biology in a Nutshell

Life is by rights the most complex and fascinating mechanism, from the moment a living organism is created until the time it ceases to exist. Life can be threatened, enhanced or generally amended in many different ways within the branch of genetics. Genetics is the biological field that studies genes, genetic variations and heredity in living organisms (Mather et al. [88]).

In the current chapter, we are interested in modifications on the genetic material and especially DNA methylation, which is a process that changes the activity of DNA regions without altering its sequence. In order to comprehend this modification, we first need to unravel the pieces of a living organism and then define the methylation's action. We begin with the definition of cells and their connection to DNA. We briefly explain the transcription process responsible for the protein production (essential elements to form life), aiming at highlighting the impact of DNA amendments on the resulted living creature. Subsequently, we discuss about Cytosine methylation and the contemporary method called Bisulfite Conversion liable for quantifying the level of methylation. We then report some of the most popular DNA profiling techniques accompanied by their main advantages and disadvantages. This molecular biology synopsis is important for Chapter 5, where real blood samples from neonates with differentially methylated gene regions are analyzed. Furthermore, due to additional real applications (data from BLUEPRINT epigenome project, Stunnenberg et al. [127]) on two methylated cell types of the immune system: a) neutrophils and b) monocytes, an extra short section is presented at the end regarding their function as white blood cells.

1.2 Cells and Genetic Material

Cells are the smallest living units of an organism and are categorized into eukaryotic and procaryotic. Both categories bear three common features: a) a cell membrane that separates the content of a cell from its environment, b) the cytoplasm, which is a jelly-like fluid under the membrane and c) the genetic material known as DNA (DeoxyriboNucleic Acid) where all the information for the cell functions is stored. Eykaryotic cells, found in plants and animals, possess a core called nucleus as well as membrane-enclosed organelles, in contrary to prokaryotic which include none of the above (Vellai and Vida [140]).

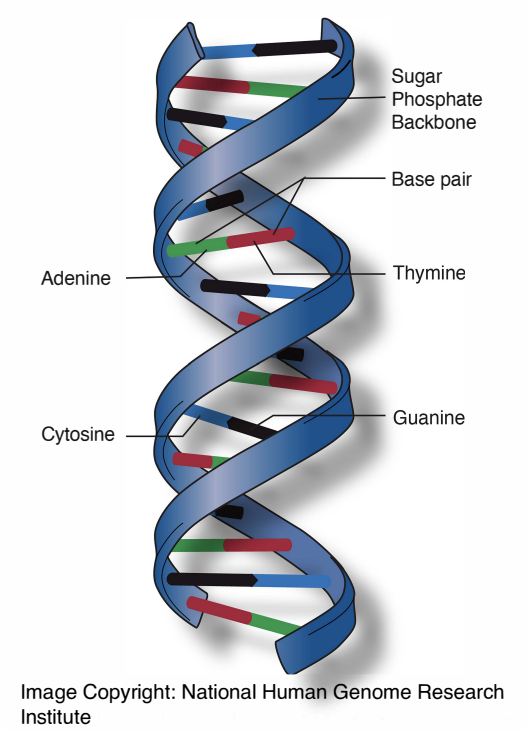


Figure 1.1 DNA double helix. The steps are the base pairs while the nucleobases on each chain are bonded together *via* sugar-phoshates.

DNA is made up of atoms that are combined together to form a long spiraling molecule with two strands, the double helix (Watson and Crick [146]). This helix reminds a ladder with steps the pairs of four different chemicals (bases or nucleobases), the Adenine (A) - Thymine (T) couple and the Cytosine (C) - Guanine (G) couple (Chargaff [24]) (Figure 1.1). A single DNA chain may have length up to 2 meters in human cells (Annunziato [3]) rendering difficult to package inside the small-scale nucleus. To achieve fitting, DNA is wrapped around proteins, the histone octamers (Peterson and Laniel [109]), forming compact packages known as the nucleosomes, with the whole derived fiber being the chromatin and the overall result formulating the chromosome (Hammond et al. [53]) (Figure 1.3). In total, humans have 46 chromosomes inside each one of their cells (Tjio and Levan [139]). During cell reproduction, chromosomes are paired into 23 chromosome couples (for humans) with one chromosome inherited from

the mother and the other from the father. The genetic information is written across these 23 pairs, with the two individuals containing principally the exact same genes at the same locations. Nonetheless, slight variations that carry brand new information could exist due to unique mutations on the genetic code.

1.3 Transcription and Proteins

Each cell has a specific activity based on the genes that are expressed. For example, the liver cells, despite of having the same DNA as muscle cells, read the liver genes and silence the muscle genes, whereas muscle cells perform the opposite. When a gene is switched on, an enzyme called RNA polymerase binds to the start of the gene called promoter region and crosses along the DNA by creating the single chained mRNA (messenger RNA) out of free bases in the nucleus. This process is known as transcription (Clancy [26]). mRNA is then set free to the cytoplasm from the tiny pores of the nucleus and enters the ribosome particle. Ribosome is the protein building machine that reads by a three-base step the RNA code with the help of tRNAs (transfer RNA). Each tRNA transfers the appropriate amino acid (20 different variants) according to the base triplet termed codon, until the whole chain has been read. This is the translation procedure (Clancy and Brown [27]). Eventually, the ribosome releases a sequence of amino acids, the peptide chain (Figure 1.2), which gets packed in a compact form to compose a very specific protein. Proteins are the essential ingredients to form the living cells, where living cells make up the tissues, tissues the organs and organs - when combined and set in function - create the living creatures (plants, animals etc.).

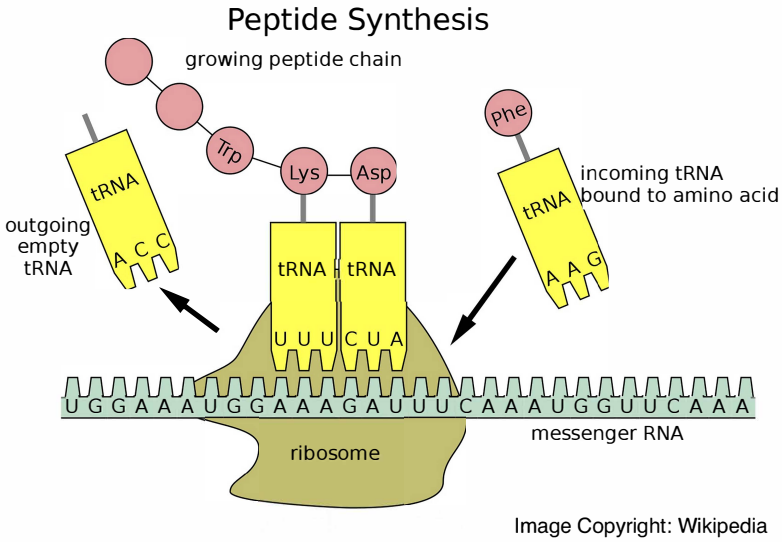


Figure 1.2 The peptide synthesis. The ribosome binds on the mRNA and starts reading the codons with the help of tRNAs. The incoming tRNA transfers the suitable amino acid which finally fastens to the growing peptide chain. tRNA is then released empty (outgoing empty tRNA).

To summarize, proteins are the source of life. They have to be produced at the right time, on the right shape and the right quantity, otherwise mutations may occur. DNA, as we already mentioned, is responsible for determining when and how these proteins are produced. Hence, any modifications on or in the DNA influence radically the proteins and thereby the resulted form of life.

1.4 Epigenetics and Cytosine Methylation

Epigenetics is the study of heritable alterations upon the genes and not within (no changes on the DNA sequence). This justifies the title "Epi-genetics", where "epi" is the greek prefix for "above" meaning modifications above the genetic material. Specifically, Epigenetics involves non-coding RNA (Morris [100]), histone modification (Strahl and Allis [126]) and DNA methylation (Moore et al. [99]). In this thesis, we conduct statistical analysis on the latter with focus on the Cytosine methylation. In Figure 1.3 we observe how histone modification and DNA methylation apply to affect DNA expression. Special chemical tags (yellow labels) attach to the histone (on the tails) and onto the double helix (DNA) concluding in alteration in gene expression. Histone modification defines how tight the double helix folds around the protein resulting in either expression or silencing of the genes in the wrapping area.

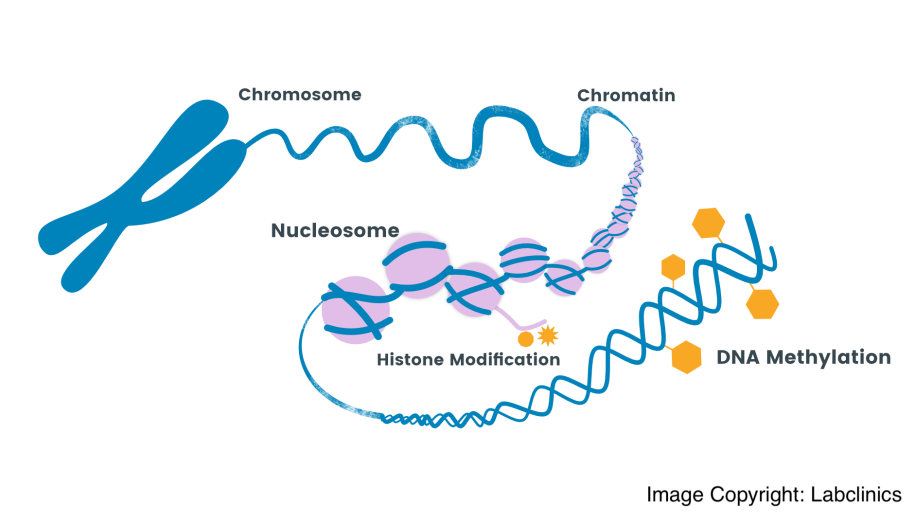


Figure 1.3 The structure of a chromosome pair modified by epigenetic marks (yellow tags). The chromosome is condensed to chromatin, which chromatin consists of nucleosomes. Nucleosome is a complex of DNA and histones (proteins). The histones (purple spheres) can be modified through chemical tags that bind on their tails, leading to histone modification and therefore alteration in gene expression in the folding DNA area. DNA methylation is the result of modifications onto the double helix with the attachment of chemical tags that do not alter the DNA sequence, although they affect the gene expression.

In Cytosine methylation, methyl groups are affixed directly to a Cytosine residue that exists in a CpG site by particular enzymes known as DNA methyltransferases. CpG stands for Cytosine-phospate-Guanine with phospate (in particular sugar-phospate) binding the CG pair in the single chain. A CpG site is a region of the DNA where a Cytosine is followed by a Guanine in the 5' to 3' direction. Segments with high frequency of CpG sites (> 50%) and sequence length greater than 200BP (base pairs) shape a CpG island (Illingworth and Bird [62]), where the addition of the methyl group CH_3 happens. Adjacent CpG sites are likely to share the same methylation status and therefore be correlated, whereas as the distance between CpG sites grows the co-methylation tends to decline (Affinito et al. [1]). Cytosine methylation is normally encountered in mammals, with approximately 70%-80% of their CpG Cytosines being methylated (Jabbari and Bernardi [63]) reshaping into 5-methylcytosines (Figure 1.4). This modification usually happens in the promoter region of a gene, blocking the RNA polymerase from binding and starting the transcription process. Therefore, the gene is deactivated (Bird [10]) and the corresponding protein is not produced.

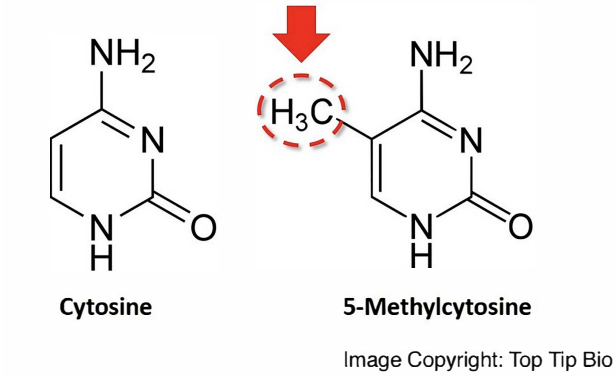


Figure 1.4 Cytosine and 5-methylcytosine after the addition of the methyl group CH_3. On the left the chemical structural formula of Cytosine's is presented, while on the right, the original structure is altered by the addition of the methyl-group CH_3 (dashed red circle) at the carbon 5 position, resulting in a methylated Cytosine (5-methylcytosine).

1.4.1 DNA Methylation Levels

DNA methylation on the right levels plays a crucial role in balancing the overall function of a living creature, for the reason that it silences tissue specific genes from being expressed in the wrong tissue. In cases the body detects large amounts of unmethylated DNA, it activates the immune system assuming there is a bacterial infection (bacterial DNA is mostly unmethylated). Hence, methylation has to exist for reasons of normal functioning. On the other hand, it can be severely dangerous when methyl groups are attached to promoter regions of tumor suppressor genes or within the gene, provoking cancer or likely other negative conditions. Consequently, there is urge to discover the type of methylation (what genes are prevented from expressing) and the level

of methylation: hypo- or hypermethylation. Hypomethylation has been accepted as a cause of oncogenesis (Das and Singal [30]), whilst a global hypomethylation with gene-specifc areas of hypermethylation can work as an early biomarker of atherosclerosis (Dong et al. [35]). In general, alteration in the methylation levels is an important indicator for cancer development and it may also be connected to autoimmune diseases such as lupus and multiple sclerosis (Wilson et al. [150]). Therefore, aberrant DNA modification can be responsible for majorly negative consequences on the quality and life duration of a being.

1.5 Bisulfite Conversion and DNA Profiling

Sodium Bisulfite conversion is the gold standard method for detecting DNA methylation (Frommer et al. [48]). During this process the unmodified Cytosines are deaminated (hydrolysis reaction) to Uracil (Figure 1.5), while the 5-methylcytosines remain unaffected.

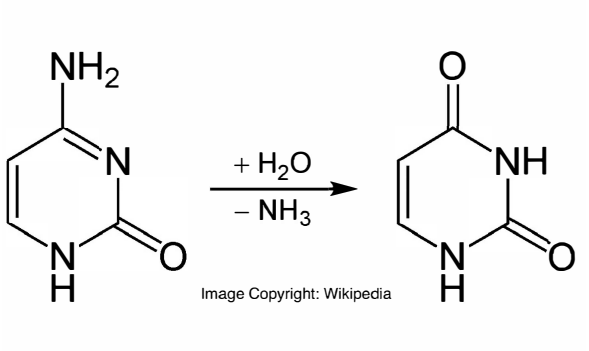


Figure 1.5 Deamination of Cytosine to Uracil. On the right, the structural formula of Cytosine is presented. A molecule of water H_2O breaks Cytosine's amino group NH_2 (hydrolysis reaction) resulting in the structural formula of Uracil on the right. During the process ammonia, NH_3, is released.

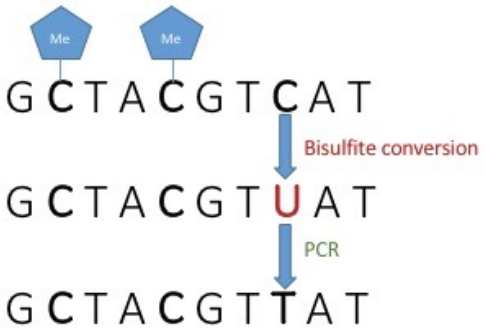


Figure 1.6 Bisulfite sequencing protocol that discovers Cytosine methylation. Blue tags above Cytosines indicate the effect of methylation. Through Bisulfite conversion, the unmethylated Cytosines transform to Uracils (deamination), while the methylated remain intact. In the next step, during PCR amplification Uracils covert to Thymines. Thus, the finally modified chain contains only the methylated Cytosines which are now detectable.

PCR amplification (Polymerase Chain Reaction) can follow after Bisulfite conversion where Uracil is converted to Thymine (Horváth and Vértessy [61]). The Cytosines that survive the process are the methylated ones (Figure 1.6).

With a focus on finding the DNA methylation pattern in a sample, scientists have to profile first the CpG islands of the individual across her genome and then implement the Bisulfite conversion and PCR amplification. Many techniques have been developed for DNA profiling purposes, which can be roughly categorized into two groups: a) bisulfite sequencing-based and b) array-based methods (Yang et al. [152]).

1.5.1 Bisulfite Sequencing-based Methods

In this category belong the whole-genome bisulfite sequencing (WGBS) and the reduced representation bisulfite sequencing (RRBS). WGBS offers coverage of $> 90\%$ of the CpG islands in the whole genome and it has been characterized as the benchmark method for profiling (Plongthongkum et al. [112], Farlik et al. [42]). RRBS interrogates only the CpG-rich regions, corresponding to $10 - 20\%$ of the CpGs in the human genome (Meissner et al. [93], Meissner et al. [94]). DNA methylation is measured in counts for these techniques and more precisely, in number of methylated reads per CpG, along with the total number of reads for this region (read depth).

An advantage of WGBS and RRBS is that they are able to detect methylation designs at single-base resolution (Cokus et al. [28]). RRBS can although show its drawback when an investigation in CpG-deficient regions is taking place (Yang et al. [152]), with WGBS excelling due to the whole genome examination. However, WGBS as well a RRBS are highly expensive *per* sample and therefore, they remain confined to small studies. This practically leads to having access in only few number of samples, incapacitating the credibility of standard statistical tools.

1.5.2 Array-based Methods

The first developed DNA profiling methods were the array-based. In this category belongs the most widely used profiling platform known as Illumina Infinium HumanMethylation450 BeadChip (450K), which covers over 480,000 CpGs (Morris and Beck [101]). A notable improvement to 450K is EPIC BeadChip (EPIC), a complementary platform to 450K that provides coverage of $> 850,000$ CpGs, involving more than 90% of the CpGs in 450K and a further 413,743 CpGs (Pidsley et al. [110]).

To measure the methylation level, both Illumina Infinium array-based methods use a pair of methylated/unmethylated probes and calculate their intensity individually. The

DNA methylation intensity can be quantified in two ways: a) the beta-value or beta-intensity and b) the M-value (Du et al. [36]). Beta-intensity corresponds to the level of methylation in an interrogated CpG site measured in the $[0, 1]$ interval (percentage of methylation), defined as: $\text{Meth}/(\text{Meth} + \text{Unmeth} + 100)$, where Meth signifies the intensity of the methylated probe and Unmeth the intensity of the unmethylated probe (Yang et al. [152]). The M-value corresponds to the $log2$ ratio of the intensities of the methylated probe versus the unmethylated probe (Du et al. [36]). In Du et al. [36], the relationship of the beta-intensity and M-value is easily proven to be a logistic function (base 2 logarithm), providing a convenient transformation of unrestricted support range.

In regard to the advantages and disadvantages of the Illumina platforms, both determine the methylation pattern at single-base resolution using probes on a micro-array, are cost-effective and simple to analyze (Yang et al. [152]). However, the limited coverage of genome, especially in 450K, is considered one of the main weaknesses of this field of profiling techniques. Nonetheless, array-based methods supply scientists with considerably larger amount of samples compared to Bisulfite Sequencing-based Methods, thanks to their lower price, increasing the credibility of the results derived by standard statistical methods.

Technical biases and corrections

Despite of the evident advantages, the analysis of DNA methylation data produced by array-based technologies presents challenges due to the existence of technical biases. Wang et al. [142] and Wilhelm-Benartzi et al. [149] introduce explicitly those biases, along with an analysis framework for corrections. For example, 450K BeadChip has two different probe types known as Infinium I and Infinium II that lead to a type design bias (Bibikova et al. [8]). In particular, based on the Dedeurwaerder et al. [32] study, the beta-intensities generated by Infinium II probes had a smaller range and were less sensitive to detect extreme methylation values than those obtained from Infinium I, triggering this type of bias and calling for normalisation/scaling actions.

Wang et al. [142] stress out the importance of pre-processing and normalising the BeadChip array data for performing a successful analysis. Important steps are the within-array normalisation and the consideration of batch effects. The within-array normalisation removes the background noise and corrects for technical dye-based intensity (red/green) and probe type differences (I/II). There are various techniques to perform the probe type correction, such as those described in Dedeurwaerder et al. [32] (peak-based correction) and Teschendorff et al. [134] (BIMQ - Beta-MIxture Quantile normalisation method). The BIMQ method in Teschendorff et al. [134] is a model-based strategy that adjusts for differences due to Infinium II and Infinium I beta-intensities.

BIMQ was reported as the best algorithm for tackling probe design bias, according to Marabita et al. [86].

The step regarding batch effects accounts for variation that is not caused by biological difference but by technical variation, *i.e.,* samples collected on different days or by different facilities, masking this way the true biological signal. Several methods are proposed to correct for those effects, with ComBat (Johnson et al. [66]) being one of the most frequently used since it adjusts for multiple confounders if needed and is not sensitive to outliers in small-sized samples (Sun et al. [129]). Specifically, ComBat uses empirical Bayes frameworks to adjust for batch effects. In our study of real datasets in Chapter 5, data normalisation was performed by the BMIQ method, while the batch correction was performed by ComBat.

1.5.3 Statistical Applications on Methylation Data

With respect to statistical tools for differential DNA methylation analysis, Robinson et al. [119] report a variety of techniques to find differential sites or regions. For example, there are advanced statistical methods for differential analysis in bisulfite sequencing data that exploit the Beta-Binomial model. MOABS (Sun et al. [128]), DSS (Feng et al. [43]) and methylSig (Park et al. [108]) are some of the packages that are based on the Beta-Binomial assumptions (given the methylation proportion at a CpG site, the observations follow a Binomial distribution, while the methylation proportion varies across the samples, *i.e.*, patients). In array-based data, such as Illumina 450K and EPIC array, the data for downstream analyses can be either beta-intensities or M-values as described earlier, with a preference on M-values (Du et al. [36]) because various statistical tools can be easily applied on them (*i.e.,* limma by Smyth [125]). For differential methylation tests, Wang et al. [141] suggest non-parametric tests on beta-intensities (like Wilcoxon test). ANOVA and t-tests are offered on the other hand in the COHCAP environment (Warden et al. [144]) - a package that applies on both beta-intensities of array-based data and methylation proportions of bisulfite sequencing data.

Nonetheless, in this thesis, we are not interested in determining the differentially methylated regions (DMRs) or sites. In particular, we are focused on clustering individuals according to their methylation profile in predefined differentially methylated regions (to be discussed in Chapter 5), seeking out any hidden heterogeneity between the subgroups. In practice, we do not have any label that permits the stratification of individuals into groups, besides demographic characteristics that are treated in our framework as confounding effects. Consequently, we propose an elegant and feasible way to model this scenario through hierarchical Bayesian mixtures.

To be specific, we cluster methylation beta-intensities drawn from the 450K and EPIC platform via mixtures of Beta densities, owing to their bounded $[0, 1]$ support range. In case of M-values, we suggest the mixture of Gaussian densities[1]. With regards to WGBS and RRBS, the derived type of data is counts (methylated reads and read depth). For these DNA related counts the advisable model is Beta-Binomial (Sun et al. [128]), which we level-up to mixtures when clustering is in demand inducing the hierarchical Binomial mixture model. This model takes in two parameters: the number of methylated reads and the read depth, while accounts for overdispersion in binomially distributed data (Kim and Lee [67]).

1.6 Discussion

In this chapter, we conducted a brief report on the very basic concepts of molecular biology with the intention to build an elementary knowledge around the action of Cytosine methylation. This thesis aims at revealing sub-populations established by specific Cytosine methylation patterns (real application in Chapter 5), however its ultimate purpose is to achieve providing all the necessary background work for constructing model-based clustering methods; methods that exploit the flexibility of the Bayesian mixture models and the scalability of Machine Learning inferential algorithms in order to efficiently cluster discrete data, such as counts or binary, and data with bounded support range (*i.e.*, $[a, b]$) in cases logit or log-transformations are not normally distributed (Changyong et al. [23]). An additional aim is to discover model-based clustering tools that can simultaneously control for the effect of confounding parameters (*i.e.*, batch effect, sex etc.). The final objective is to be capable of providing information regarding the most discriminative features that are responsible for leading the segregation into the estimated sub-populations.

[1]Model not described in the thesis, but code provided in the Appendix. More details in Chapter 3.

Chapter 2

Variational Bayes, Mixture Models and Feature Selection

2.1 Modern Computational Tools for High-structured Datasets

The process of estimating the parameters of a probabilistic model, known as statistical inference, constitutes a top discussion topic in Statistics and Machine Learning. Several algorithms have been proposed capable of making inference in intractable scenarios[1] either stochastically such as Markov chain Monte Carlo (Carlin and Chib [19]) or deterministically like the Expectation Maximization algorithm (McLachlan and Krishnan [91]) and Variational Inference (Blei et al. [14]).

MCMC bears a widespread reputation as a family of techniques that generates realizations from the invariant true posterior distribution. It builds a Markov chain that eventually settles on its equilibrium distribution. This final state distribution is the posterior from which the algorithm draws samples, proving that MCMC is an efficient and accurate mechanism to learn the model parameters. The most renowned of the MCMC algorithms are Metropolis-Hastings (Chib and Greenberg [25]) and Gibbs sampler (Casella and George [21]) while a wealth of variations exist. Some of them are Metropolis-adjusted Langevin, Hamiltonian Monte Carlo and non-reversible Zig-Zag. Metropolis-adjusted Langevin algorithms use Langevin dynamics to propose new states and Metropolis Hastings to accept or reject the proposals (Roberts and Tweedie [118]). Hamiltonian Monte Carlo utilizes the Hamiltonian dynamics and the Metropolis Hastings acceptance step to draw samples from the targeted distribution (MacKay

[1]The intractable scenario concerns the computation of a complex integral for the derivation of the marginal likelihood, the most important ingredient to perform models comparison and selection.

[81], Betancourt [7]) and non-reversible ZigZag omits the rejection step introducing a considerably faster rejection-free MCMC algorithm (Bierkens and Roberts [9]). Along with the great advantage of this class of exact sampling methods comes a noticeable drawback. This is the slow execution and consequently the slow convergence to the posterior due to the computationally demanding evaluation of the likelihood function at each one of the possibly high-dimensional observations, rendering this technique utterly time consuming, especially in cases fast results are in demand (*i.e.*, worldwide high-structured data analyzed for the evolution of a new pandemic).

To tackle this problematic situation, optimization algorithms widely used in Machine Learning problems, that scale well to large datasets are usually employed. In this thesis, we choose to enroll Variational Inference for inferring complex mixture of likelihood components, due to its significantly rapid convergence to the final estimates compared to MCMC. The variational algorithm shows a close connection to EM (Expectation-Maximization) because both follow the same E and M steps, with the difference lying on the output (Bishop [11]). EM results in point estimates, while Variational Inference in parameter distributions. The fact the variational method builds a complete information package is considered an extra advantage over EM which provides only the expected value skipping important information like variance and shape.

The goal of this chapter is two-fold: a) to survey variational approximations focusing on the density transform approach (Ormerod and Wand [107]) - probably the most common version - and b) to set out the theory around the complex family of mixture models where Variational Inference successfully applies. Both sections compose the foundation of Chapter 3, where their joint-presentation takes place. In addition, we present a simple extension to circumvent convergence issues of the optimization algorithm for non-convex instances, known as “annealing”, while we yield variational pseudocodes for the topmost regression models (linear regression, linear mixed model, probit regression and probit mixed model). Finally, we categorize the mixture models in finite and infinite, where in the former the number of components is fixed, whereas in the latter the components are infinite by construction allowing model-determination. The last section is devoted to selection of those features that discriminate one component from another based on a discriminative accuracy measure proposed by Lin et al. [76].

2.2 Kullback - Leibler Divergence

According to Variational Inference, the posterior probability of the unknown parameters is approximated by a distribution for which the normalizing constant is a more tractable probability function. The classical variational theory depends on the minimization of the Kullback - Leibler divergence (Kullback and Leibler [70]) while one of the most

common variational algorithms is Mean Field (alternatively, the density transform approach introduced by Ormerod and Wand [107]). Mean Field is responsible for the alternative name Variational Inference takes, admitted as Variational Bayes (Ormerod and Wand [107]). In this thesis, when we refer to Variational Inference we signify the Mean Field approximation, or alternatively Variational Bayes (VB).

Now, let us consider of having a Bayesian model with $\boldsymbol{\theta}$ the model parameters and $\boldsymbol{y}$ the observed data. The model might also introduce latent variables $\boldsymbol{z}$ which can be treated as parameters and be absorbed into $\boldsymbol{\theta}$. Hence, it is assumed that $\boldsymbol{\theta}$ are all the unknown parameters of the model. As specified by the Bayes' theorem, the posterior distribution is equal to

$$P(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{P(\boldsymbol{y}, \boldsymbol{\theta})}{P(\boldsymbol{y})}, \tag{2.1}$$

where $P(\boldsymbol{\theta} \mid \boldsymbol{y})$ denotes the posterior, $P(\boldsymbol{y}, \boldsymbol{\theta})$ the joint distribution and $P(\boldsymbol{y})$ the marginal likelihood. In the case where the $\boldsymbol{y}$ vector is continuous, the marginal likelihood on the denominator may be an integral of an intractable form. To deal with this, an equivalent decomposition is derived for the log-marginal likelihood after the introduction of an arbitrary distribution function $q(\cdot)$ over the parameter space of $\boldsymbol{\theta}$, which $q(\boldsymbol{\theta})$ will eventually work as a tractable approximation of the true posterior. The same situation might occur in discrete frameworks with the existence of complex summations. However, the continuous case is presented solely throughout this chapter, for the reason that the only difference with the discrete is the replacement of the integrals by summations.

$$\begin{aligned} \log P(\boldsymbol{y}) &= \log P(\boldsymbol{y}) \int q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \log \left(\frac{P(\boldsymbol{y}, \boldsymbol{\theta}) / q(\boldsymbol{\theta})}{P(\boldsymbol{\theta} \mid \boldsymbol{y}) / q(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \log \left(\frac{P(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \left(\frac{q(\boldsymbol{\theta})}{P(\boldsymbol{\theta} \mid \boldsymbol{y})} \right) d\boldsymbol{\theta}. \end{aligned} \tag{2.2}$$

The second term on the right hand side of equation (2.2) is the Kullback-Leibler divergence of distribution $q(\cdot)$ from the posterior distribution. This dissimilarity function belongs to a wider family of divergences called α-Divergence, also know as Rényi divergence (Li and Turner [73]), in which the parameter α can obtain values in $\mathbb{R}$. For $\alpha \to 0$ the $\mathrm{KL}(q \| P)$ is retrieved, whilst for $\alpha \to 1$ we obtain $\mathrm{KL}(P \| q)$ (when this divergence is minimized the algorithm is called Expected Propagation, Minka [97]). KL is not a symmetric function ($\mathrm{KL}(q \| P) \neq \mathrm{KL}(P \| q)$). For instance, from Minka [96], if the true posterior is a mixture of two univariate Gaussians, $q(\cdot)$ tends to capture one of the posterior's modes as we reduce the value of α. Conversely, higher α values force the approximated $q(\cdot)$ to cover the entire true distribution without defining the two modes.

For the Bayesian variational framework, KL with $a \to 0$ is utilized because it appears as a term in equation (2.2). According to Gibb's Inequality (MacKay [81]),

$$\mathrm{KL}\left(q \| P\right) = \int q\left(\boldsymbol{\theta}\right) \log\left(\frac{q\left(\boldsymbol{\theta}\right)}{P\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)}\right) d\boldsymbol{\theta} \geq 0. \tag{2.3}$$

Substituting KL into the log-marginal likelihood in (2.2), a new inequality arises. This result introduces a lower bound of the log-marginal likelihood known as Evidence Lower Bound (ELBO), which is denoted as $\mathcal{L}\left(\boldsymbol{y}; q\right)$

$$\log P\left(\boldsymbol{y}\right) \geq \int q\left(\boldsymbol{\theta}\right) \log\left(\frac{P\left(\boldsymbol{y}, \boldsymbol{\theta}\right)}{q\left(\boldsymbol{\theta}\right)}\right) d\boldsymbol{\theta} := \mathcal{L}\left(\boldsymbol{y}; q\right). \tag{2.4}$$

The previous derivation procedure resembles the EM algorithm (Dempster et al. [33]) in the part of exploiting the objective function (ELBO) to make inference. Despite that, their difference is found in the way $\boldsymbol{\theta}$ vector is treated. In Variational Inference, $\boldsymbol{\theta}$ is a random variable vector and therefore prior distributions are imposed upon its elements, while in the frequentist EM $\boldsymbol{\theta}$ is considered fixed with the data being the only random variables.

Regarding the ELBO, it obtains a more tractable form than the log-marginal likelihood if the $q(\cdot)$ distribution is restricted to a manageable family of distributions. At the same time, this family should be rich and flexible, aiming at approximating well the true posterior. The appropriate distribution of the selected family is defined by estimating its parameters through the minimization of the Kullback-Leibler divergence, which is equivalent to the maximization of the ELBO. The optimal solution is attained when $q(\boldsymbol{\theta}) = P(\boldsymbol{\theta} \mid \boldsymbol{y})$ (Ormerod and Wand [107]).

The most common restrictions for the selection of $q(\cdot)$ are the use of a parametric distribution or the factorization of $q(\cdot)$ into independent variational distributions of disjoint parameter groups. In the first case, the ELBO becomes a function of the assumed distribution's parameters and optimization techniques are applied to derive the optimal values. The latter case forms the Mean Field approach, presented in the next section.

2.3 Mean Field Approximation

The Mean Field approximation is considered to be a nonparametric restriction of the variational distribution by assuming independence between the parameter groups. These groups are created after the partition of the parameter vector $\boldsymbol{\theta}$ into disjoint groups $\boldsymbol{\theta}_i$ such that $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i\}_{i=1}^{I}$. Then, the $q(\cdot)$ distribution is factorized as follows

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{I} q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i). \tag{2.5}$$

For notation simplicity, we omit the distribution subscripts and let the parameter input define them, *i.e.,* $q_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) = q(\boldsymbol{\theta}_i)$. The independence restriction has advantages and disadvantages regarding the accuracy of the approximation. In cases where the posterior dependence between $\boldsymbol{\theta}_i$ is strong, the Mean Field approach leads to inaccurate approximation of the true posterior. On the other hand, in weak dependence situations Variational Bayes is a strong candidate algorithm for deriving considerably accurate results (Titterington et al. [137]). In general, it is unfortunately not known *a priori* what would be the dependence of the parameters in the posterior space, thus strong assumptions have to unavoidably be made without knowing their effects on the quality of Variational Bayes. Recent solutions are proposed in Tan and Nott [131] and Smith et al. [124].

In relation to the form of the variational distributions $q(\boldsymbol{\theta}_i)$, this is defined in two steps. The first step is the substitution of the product distribution (2.5) into the lower bound $\mathcal{L}(\boldsymbol{y}; q)$ in equation (2.4), and the second is the optimization of it. The general form of the ELBO in (2.4) is written equivalently as a sum of expected values in equation (2.6) to facilitate the derivation of Mean Field. The expected values are with respect to $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}; q) &= \int q(\boldsymbol{\theta}) \log\left(\frac{P(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right) d\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}}\left[\log\left(\frac{P(\boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}}\left[\log P(\boldsymbol{y}, \boldsymbol{\theta})\right] - \mathbb{E}_{\boldsymbol{\theta}}\left[\log q(\boldsymbol{\theta})\right].\end{aligned} \tag{2.6}$$

Two equations hold given the product transform in (2.5) (Blei [12]). The first one is true due to the probability chain rule and the second one because of the distribution factorization. After plugging equations (2.7) and (2.8) into (2.6), the new lower bound is obtained below and indicated as $\mathcal{L}_{\text{MF}}(\boldsymbol{y}; q)$, where index MF stands for Mean Field,

$$P(\boldsymbol{y}, \boldsymbol{\theta}) = P(\boldsymbol{y}) \prod_{i=1}^{I} P(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{i-1}, \boldsymbol{y}), \tag{2.7}$$

$$\mathbb{E}_{\boldsymbol{\theta}}[\log q(\boldsymbol{\theta})] = \sum_{i=1}^{I} \mathbb{E}_{\boldsymbol{\theta}_i}[\log q(\boldsymbol{\theta}_i)], \tag{2.8}$$

$$\mathcal{L}_{\text{MF}}(\boldsymbol{y}; q) = \log P(\boldsymbol{y}) + \sum_{i=1}^{I} \mathbb{E}_{\boldsymbol{\theta}}\left[\log P(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{i-1}, \boldsymbol{y})\right] - \sum_{i=1}^{I} \mathbb{E}_{\boldsymbol{\theta}_i}[\log q(\boldsymbol{\theta}_i)]. \tag{2.9}$$

Having formed the mathematically friendly objective function $\mathcal{L}_{\text{MF}}(\boldsymbol{y}; q)$, we proceed with the optimization *via* the Coordinate Ascent (Luo and Tseng [79]). This will be the algorithm for determining the optimal parameters throughout the thesis, since we work with models for which closed form variational updates can be obtained, and the number of observations is considerably less than millions (see DNA methylation applications in Chapter 5). Generally, Coordinate Ascent is not quite efficient in massive datasets such as in analysis of millions of articles in Wikipedia or DNA sequences of millions of people, considering that demands a pass through the full dataset at each iteration

(Hoffman et al. [59]). Hoffman et al. [59] introduce a scalable variational method by using stochastic optimization instead (Robbins and Monro [117]). In this method inference is based on subsamples of the data (minibatches) and not on the full dataset, achieving scalability and high speed. Nonetheless, Coordinate Ascent is the commonly used optimization algorithm in relatively smaller dataset analyses.

In particular, in Coordinate Ascent each parameter is optimized iteratively by holding the other parameters fixed. If each conditional distribution of the parameters is in the exponential family and the corresponding variational distribution belongs to the same exponential family, it is guaranteed that each coordinate can be optimized in closed form (Blei et al. [14]). The optimization starts with retaining all the parameters fixed except the j^{th} one, where $i = 1, ..., j, ..., I$. The Mean Field ELBO in equation (2.9) is treated as a function of $q(\boldsymbol{\theta}_j)$ (any other term rather than $q(\boldsymbol{\theta}_j)$ is absorbed into the constant term) and the chain rule is employed, with $\boldsymbol{\theta}_j$ being the last variable in the list. Note that $\boldsymbol{\theta}_{/j}$ denotes all the elements of $\boldsymbol{\theta}$ parameter vector excluding $\boldsymbol{\theta}_j$. The Mean Field ELBO can now be seen as in equation (2.10)

$$\begin{aligned}\mathcal{L}_{\mathrm{MF}}(\boldsymbol{y}; q) &= \mathbb{E}_{\boldsymbol{\theta}}\left[\log P(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{/j}, \boldsymbol{y})\right] - \mathbb{E}_{\boldsymbol{\theta}_j}[\log q(\boldsymbol{\theta}_j)] + \text{constant} \\ &= \int q(\boldsymbol{\theta}_j)\mathbb{E}_{\boldsymbol{\theta}_{/j}}\left[\log P(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{/j}, \boldsymbol{y})\right] d\boldsymbol{\theta}_j - \int q(\boldsymbol{\theta}_j)\log q(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j + \text{constant}.\end{aligned} \tag{2.10}$$

The Lagrange multipliers λ_i (Rockafellar [120]) are therefore applied to the ELBO in equation (2.10) resulting to (2.11), where $\mathcal{LE}$ stands for Lagrangian-equation. The derivative of equation (2.11) with respect to $q(\boldsymbol{\theta}_j)$ is then set equal to zero as shown below in equation (2.12)

$$\mathcal{LE}(\boldsymbol{y}; q) = \mathcal{L}_{\mathrm{MF}}(\boldsymbol{y}; q) - \sum_{i=1}^{I} \lambda_i \int q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i, \tag{2.11}$$

$$\frac{d\mathcal{LE}(\boldsymbol{y}; q)}{dq(\boldsymbol{\theta}_j)} = 0 \Rightarrow \mathbb{E}_{\boldsymbol{\theta}_{/j}}\left[\log P(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{/j}, \boldsymbol{y})\right] - \log q(\boldsymbol{\theta}_j) - 1 - \lambda_j = 0. \tag{2.12}$$

Thus, we solve equation (2.12) with respect to $q(\boldsymbol{\theta}_j)$ and obtain its optimal form in equation (2.13) defined as $q^{\star}(\boldsymbol{\theta}_j)$. This form resembles the behavior of Gibbs sampler in the sense that the latter collects samples sequentially from the full conditionals, while VB derives a more manageable and time-saving way to exploit them through calculation of their expected log values (Ormerod and Wand [107]).

$$q^{\star}(\boldsymbol{\theta}_j) = \exp\left\{\mathbb{E}_{\boldsymbol{\theta}_{/j}}\left[\log P(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{/j}, \boldsymbol{y})\right]\right\} + \text{constant}. \tag{2.13}$$

An equivalent alternative result arises if we picture the Directed Acyclic Graph (DAG) of the Bayesian model. It follows that instead of using the full conditional of $\boldsymbol{\theta}_j$, we can alternatively benefit from the distribution of the parameter given the Markov Blanket and apply VB (Dechter and Pearl [31]), where Markov Blanket is the set of parents, co-parents and children of $\boldsymbol{\theta}_j$ node in the DAG. In the equations below, we denote

Markov Blanket as MB_j. Then,

$$P(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{/j}, \boldsymbol{y}) = P(\boldsymbol{\theta}_j \mid \mathrm{MB}_j). \tag{2.14}$$

Hence,

$$q^\star(\boldsymbol{\theta}_j) = \exp\left\{\mathbb{E}_{\boldsymbol{\theta}_{/j}}\left[\log P(\boldsymbol{\theta}_j \mid \mathrm{MB}_j)\right]\right\} + \text{constant}. \tag{2.15}$$

The constant term on the right hand side of equation (2.15) refers to the normalization constant of $q^\star(\boldsymbol{\theta}_j)$.

Concerning the coordinate ascent algorithm, it converges to at least a local maxima and convexity properties can be used to guarantee it (Boyd and Vandenberghe [17]). The variational scheme is accomplished iteratively by updating the variational parameters at each iteration according to the lower bound. The ELBO value levels-up while proceeding, and therefore the algorithm convergences when the increase in $\mathcal{L}_{\mathrm{MF}}(\boldsymbol{y}; q)$ is negligible. More precisely, the algorithm stops when the ELBO value in the current iteration ($\mathcal{L}^i_{\mathrm{MF}}(\boldsymbol{y}; q)$) is equal or almost equal to the ELBO value in the previous iteration ($\mathcal{L}^{i-1}_{\mathrm{MF}}(\boldsymbol{y}; q)$), indicating that the parameter estimates have already reached their final value. Therefore, the stopping criterion is the insignificant difference of the current and previous ELBO (we choose 10^{-6}, however any value $> 10^{-5}$ retrieves the same result). In the following, to simplify the notation, $\mathcal{L}_{\mathrm{MF}}(\boldsymbol{y}; q)$ will be referred as $\mathcal{L}(\boldsymbol{y}; q)$ and the optimal Mean Field $q^\star(\cdot)$ as $q(\cdot)$.

2.4 Annealing

Association of Variational Inference with poor local optimas is a common argument regarding algorithm's drawbacks. It holds that it is affected by the initialization of the variational parameters leading to non-accurate estimates in cases of poorly selected initial values (Mandt et al. [84]). Nevertheless, one possible way to deal with this disadvantage is through "annealing". Mandt et al. [84] bring forth a strategy to smooth out non-convex objectives responsible for trapping the variational algorithm into poor local optimas. They suggest a complex randomized algorithm called Variational Tempering, which introduces a temperature latent variable in the model that automatically adjusts convexity at each iteration. However, in Variational Tempering we have to approximate by Monte-Carlo a multi-dimensional normalizing constant in order to make inference, adding extra complexity to the model. Therefore, we figure out instead that simple deterministic annealing (addition of a constant term at each iteration) can do the job well enough without having to go through any random tempering variable.

The whole idea lies on the "annealing" of solely the data likelihood and not both priors and likelihood (unormalized posterior) (Neal [102]). The reason is for preventing the optimization algorithm getting stuck due to any skewed priors (Mandt et al. [84]).

The Bayes theorem in equation (2.1) is differentiated to the annealed version by the introduction of a vector $\boldsymbol{T}$ directly applied to the likelihood

$$P(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{P(\boldsymbol{y} \mid \boldsymbol{\theta})^{\mathbf{1/T}} P(\boldsymbol{\theta})}{P(\boldsymbol{y})}, \tag{2.16}$$

where $\boldsymbol{T}$ in particular is a sequence of positive real values starting from value 1 ($\{T \in \mathbb{R}^+ \mid T \geq 1\}$), whilst the length of this sequence depends on the user's choice - how slow or fast she desires the annealing to be (Mandt et al. [84]). For instance, if $\boldsymbol{T}$ is 100 elements long the variational algorithm will be initially running for 100 iterations so as to temper the ELBO and escape convergence to poor local optimas. After the end of the annealing process, the ELBO retrieves its original form and the variational algorithm carries on the iterations until it converges to better local optimas thanks to the optimization guidance from the annealing. Annealing iterations may slightly delay the convergence due to the extra iteration steps on smoothing the objective function. As previously explained, the variational algorithm has to complete the annealing iterations first and then move towards convergence.

In coordinate ascent settings, there is no learning rate as in stochastic gradient algorithms and therefore, the slower we anneal the better in order to achieve good optimal values (Mandt et al. [84]). A candidate $\boldsymbol{T}$ would be $\boldsymbol{T} = \{T_i\}_{i=1}^{I}$ with $T_i \in \{1, ..., 100\}$ and $i = 1, ..., 100$. High temperatures, as $T_i = 100, 99, 98, ...$, result in little likelihood impact on the posterior (likelihood in the power of $1/T_i$), whereas lower temperatures increase the influence, indicating an inverse relationship. Eventually, $T_{100} = 1$ retrieves the original likelihood form. This is a useful approach in cases likelihood is non-convex, such as in mixture models, where multi-modalities are present. The annealing on these scenarios tries to burnish the non-convex effect of the likelihood for a few variational iterations (in the upper example the annealed iterations are 100), so as to lead the optimization into the right direction in search of the global optima. An explanatory scheme is shown below, where the original ELBO function in equation (2.6) is transformed to its annealed version in equation (2.17), defined as $\mathcal{L}^{\mathrm{A}}(\boldsymbol{y}; q)$. For a given T_i, the non-concave term (likelihood non-convex, hence log-likelihood non-concave) is highlighted in (2.17)

$$\mathcal{L}^{\mathrm{A}}(\boldsymbol{y}; q) = \frac{1}{T_i} \cdot \mathbb{E}_{\boldsymbol{\theta}}[\underbrace{\log P(\boldsymbol{y} \mid \boldsymbol{\theta})}_{\text{non-concave}}] + \mathbb{E}_{\boldsymbol{\theta}}[\log P(\boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{\theta}}[q(\boldsymbol{\theta})]. \tag{2.17}$$

During the first iteration, $\mathcal{L}^{\mathrm{A}}(\boldsymbol{y}; q)$ is almost concave due to the low weight of the non-concave term $\log P(\boldsymbol{y} \mid \boldsymbol{\theta})$. Variational Inference starts optimizing a smooth function for a few iterations.

First annealing iterations $\frac{1}{T_i} = \frac{1}{100}, \frac{1}{99}, \ldots$:

$$\frac{1}{T_i} \cdot \mathbb{E}_{\boldsymbol{\theta}}\left[\log P(\boldsymbol{y} \mid \boldsymbol{\theta})\right] << \mathbb{E}_{\boldsymbol{\theta}}\left[\log P(\boldsymbol{y} \mid \boldsymbol{\theta})\right], \text{ thus } \mathcal{L}^{\mathrm{A}}(\boldsymbol{y}; q) \text{ almost concave.}$$

As the algorithm proceeds, the weight of the non-concave term is slowly retrieved and eventually $\mathcal{L}^{\mathrm{A}}(\boldsymbol{y}; q)$ returns to its original form when $1/T_{100} = 1$. Annealing is over and the variational algorithm carries on into the right direction until it reaches convergence.

Last annealing iterations $\frac{1}{T_i} = \ldots, \frac{1}{3}, \frac{1}{2}, 1$:

$$\frac{1}{T_i} \cdot \mathbb{E}_{\boldsymbol{\theta}}\left[\log P(\boldsymbol{y} \mid \boldsymbol{\theta})\right] \rightarrow \mathbb{E}_{\boldsymbol{\theta}}\left[\log P(\boldsymbol{y} \mid \boldsymbol{\theta})\right], \text{ thus } \mathcal{L}^{\mathrm{A}}(\boldsymbol{y}; q) \rightarrow \mathcal{L}(\boldsymbol{y}; q) \text{ (annealing stops).}$$

2.5 Variational Regression Models

The applicability of Variational Bayes is evident in the class of regression models. The Mean Field derivation of the simple linear model (single-response and multi-response case (Brown et al. [18], Bottolo et al. [16])), the linear mixed model (Zhou and Stephens [158]), the probit (Albert and Chib [2]) and the probit mixed model (Baragatti [5]) yields closed form updates rendering Variational Bayes an attractive algorithm for rapid regression inference. In the this section, we provide the variational densities and the Coordinate Ascent pseudocodes for the aforementioned models, as illustrative examples, in preparation for more complex mixture models presented in Chapter 3.

The notation for the regression analysis is mainly borrowed by Ormerod and Wand [107], owing to the clarity of it. For convenience, we also use the same notation $\mathcal{L}(\boldsymbol{y}; q)$ for the lower bound in each model. However, each ELBO function is different and available in the Appendix A. Vectors and matrices are indicated in bold while the design matrix $\boldsymbol{X}$ has $N \times p$ dimensions, with N the number of observations and p the number of predictors including the intercept. Lastly, all the hyperparameters are fixed unless it is stated differently.

2.5.1 Linear Regression Model

The linear regression model is mostly known for its simplicity in terms of interpretation and applicability. For that reason, we are interested in demonstrating the variational single-response regression model as well as its multi-response version (Brown et al. [18], Bottolo et al. [16]).

Bayesian Single-response Linear Model: The likelihood and the conjugate prior densities for each parameter are

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_N), \tag{2.18}$$

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \tag{2.19}$$

$$\sigma^2 \sim \mathcal{IG}(A, B), \tag{2.20}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector of response variables that follows an $N-$dimensional Gaussian, $\boldsymbol{X}$ is the corresponding design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients distributed as a $p-$dimensional Gaussian and $\sigma^2 \boldsymbol{I}_N$ the diagonal covariance matrix with σ^2 following an Inverse-Gamma distribution.

The Mean Field approximation of the true posterior is the product of the individual parameter variational distributions due to $\boldsymbol{\beta}$ and σ^2 which already constitute disjoint groups in the parameter space (see equation (2.21)). Based on equation (2.15), the optimal variational densities of $\boldsymbol{\beta}$ and σ^2 are derived in equation (2.22) and (2.23) respectively. Specifically, the variational density of $\boldsymbol{\beta}$ is a p-dimensional Gaussian with variational parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$. The subscript $q(\boldsymbol{\beta})$ denotes that the parameter is a variational parameter of $q(\boldsymbol{\beta})$, while the subscript $\boldsymbol{\beta}$ implies the hypeparameter of the prior $P(\boldsymbol{\beta})$. This notation is followed accordingly in all the subsequent regression models. As for the variational density of σ^2, this is an Inverse-Gamma with shape and scale parameters $A_{q(\sigma^2)}$ and $B_{q(\sigma^2)}$ respectively. Note that both variational densities resemble the usual Gibbs sampling update in an MCMC scheme, with the full conditional distribution being a conjugate form between the likelihood and the prior.

$$q(\boldsymbol{\beta}, \sigma^2) = q(\boldsymbol{\beta}) q(\sigma^2), \tag{2.21}$$

$$q(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}), \tag{2.22}$$

$$q(\sigma^2) = \mathcal{IG}\left(A_{q(\sigma^2)}, B_{q(\sigma^2)}\right) \quad \text{with fixed } A_{q(\sigma^2)} = A + \frac{N}{2}. \tag{2.23}$$

Algorithm 1 Coordinate Ascent for the Variational Single-response Linear Model

Initialize: $B_{q(\sigma^2)} \in \mathbb{R}^+$

Repeat:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} = \left\{ \left(\frac{A + \frac{N}{2}}{B_{q(\sigma^2)}} \right) \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta})} = \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \left\{ \left(\frac{A + \frac{N}{2}}{B_{q(\sigma^2)}} \right) \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} \right\}$$

$$B_{q(\sigma^2)} = B + \frac{1}{2} \left\{ \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \right)^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \right) + \operatorname{tr}\left(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \right) \right\}$$

Stop:

$\mathcal{L}(\boldsymbol{y}; q)^{\text{current}} - \mathcal{L}(\boldsymbol{y}; q)^{\text{previous}} \leq \epsilon$, where $\epsilon = 10^{-6}$

The closed form variational equations for each variational parameter are given in Algorithm 1. Particularly, this pseudocode is the variational scheme for the single-response linear regression model *via* Coordinate Ascent, beginning with the initialization of the common terms in all the equations. Here, this term is the positive scale parameter $B_{q(\sigma^2)}$, which is initialized with a positive value (the specific value is user's choice according to her prior belief). Then, at each iteration the variational parameters are updated until the difference between the previous and current ELBO value is close to zero (our stopping criterion is less than 10^{-6}).

Bayesian Multi-response Linear Model (with multiple predictors): The multi-response linear model predicts more than one responses (usually correlated), introducing a matrix of quantitative responses and not a vector as in the singe-response linear model. This matrix is distributed by the Matrix-Gaussian density (Gupta and Nagar [52]), which is the generalization of the Multivariate Gaussian distribution to matrix-valued random variables. The hierarchical Bayesian model is described in a synthetic manner in Denison et al. [34], while its variational Mean Field scheme, to be presented shortly, is exclusively product of our work.

The likelihood and priors are

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathcal{MN}_{Nq}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_N, \boldsymbol{\Sigma}), \tag{2.24}$$

$$\boldsymbol{\beta} \sim \mathcal{MN}_{pq}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \tag{2.25}$$

$$\boldsymbol{\Sigma} \sim \mathcal{IW}_q\left(\nu_{\boldsymbol{\Sigma}}, \boldsymbol{Q}_{\boldsymbol{\Sigma}}\right), \tag{2.26}$$

where $\boldsymbol{y}$ is a $N \times q$ matrix of multidimensional response variables in (2.24), $\boldsymbol{X}$ is the corresponding design matrix, $\boldsymbol{\Sigma}$ the covariance matrix of the responses and $\boldsymbol{I}_N$ the identity covariance matrix between the observations which are assumed independent. The prior for $\boldsymbol{\Sigma}$ is an Inverse-Wishart with degrees of freedom $\nu_{\boldsymbol{\Sigma}}$ and scale matrix $\boldsymbol{Q}_{\boldsymbol{\Sigma}}$ (equation (2.26)). The variable $\boldsymbol{\beta}$ is a $p \times q$ matrix of regression coefficients distributed as a Matrix-Gaussian in (2.25), with expected matrix $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, covariance matrix between the predictors $\boldsymbol{V}_{\boldsymbol{\beta}}$ and covariance matrix between the responses $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$.

$$q(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = q(\boldsymbol{\beta})q(\boldsymbol{\Sigma}), \tag{2.27}$$

$$q(\boldsymbol{\beta}) = \mathcal{MN}_{pq}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{V}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}), \tag{2.28}$$

$$q(\boldsymbol{\Sigma}) = \mathcal{IW}_q\left(\nu_{q(\boldsymbol{\Sigma})}, \boldsymbol{Q}_{q(\boldsymbol{\Sigma})}\right) \text{ with fixed } \nu_{q(\boldsymbol{\Sigma})} = \nu_{\boldsymbol{\Sigma}} + N. \tag{2.29}$$

Regarding the variational product distribution, this is analogous to the single-response case in equation (2.21). The joint variational density is given in (2.27), where $q(\boldsymbol{\beta})$ is a Matrix-Gaussian with $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$, $\boldsymbol{V}_{q(\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$ variational parameters (equation (2.28)). The Mean Field approximated density for $\boldsymbol{\Sigma}$ is the Inverse-Wishart of (2.29), with parameters $\nu_{q(\boldsymbol{\Sigma})}$ and $\boldsymbol{Q}_{q(\boldsymbol{\Sigma})}$ denoting the degrees of freedom and the scale matrix accordingly.

Algorithm 2 Coordinate Ascent for the variational Multi-response Linear Model

Initialize: $\boldsymbol{Q}_{q(\boldsymbol{\Sigma})}$ $p \times p$ positive definite matrix

Repeat:

$\boldsymbol{\Sigma}_{q(\beta)} = \{(\nu_{\boldsymbol{\Sigma}} + N)\boldsymbol{Q}_{\boldsymbol{\Sigma}} + \boldsymbol{\Sigma}_{\beta}\}^{-1}$

$\boldsymbol{V}_{q(\beta)} = \left[\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}_{\beta}^{-1}\right]^{-1}$

$\boldsymbol{\mu}_{q(\beta)} = \boldsymbol{V}_{q(\beta)}\left[\boldsymbol{X}^T\boldsymbol{y}(\nu_{\boldsymbol{\Sigma}} + N)\boldsymbol{Q}_{q(\boldsymbol{\Sigma})} + \boldsymbol{V}_{\beta}^{-1}\boldsymbol{\mu}_{\beta}\boldsymbol{\Sigma}_{\beta}^{-1}\right]\boldsymbol{\Sigma}_{q(\beta)}$

$\boldsymbol{Q}_{q(\boldsymbol{\Sigma})} = \boldsymbol{Q}_{\boldsymbol{\Sigma}} + \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}\right) + \operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Sigma}_{q(\beta)}\right)$

Stop:

$\mathcal{L}(\boldsymbol{y}; q)^{\text{current}} - \mathcal{L}(\boldsymbol{y}; q)^{\text{previous}} \leq \epsilon$, where $\epsilon = 10^{-6}$

Algorithm 2 is the iterative scheme for updating the variational parameters of the multi-response linear model. The common parameter in all the closed form equations is the variational scale matrix of $q(\boldsymbol{\Sigma})$, $\boldsymbol{Q}_{q(\boldsymbol{\Sigma})}$, and hence the only one to be initialized by a $p \times p$ positive definite matrix given it is a covariance matrix.

2.5.2 Linear Mixed Regression Model

The linear mixed model is an extension of the single-response linear model that permits random effects. It is connected to variance components because it allows to have different variance for each individual, while standard models have the same variance for all the individuals. The linear mixed model is a method for analyzing dependent data, such as repeated measurements and longitudinal data (Nelder and Baker [104]). For instance, they can be used in clinical trials when we test the same subject in different time points (longitudinal data) or same subject with different treatment regimes (repeated measurements). Generally, we are mostly interested in mixed models when analyzing different groups of observations (hospital 1, hospital 2 etc.) or related individuals (family 1, family 2 etc.) with random intercepts and/or slopes (group specific).

Bayesian Linear Mixed Regression Model: We specify that we use the same hierarchical Variance Component model and notation as in Ormerod and Wand [107]. However, several variations can be found in McCullagh and Nelder [89]. The hierarchical model can be described by a set of equations

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma_{\epsilon}^2 \sim \mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}), \quad \text{or } \boldsymbol{y} \mid \boldsymbol{B}, \sigma_{\epsilon}^2 \sim \mathcal{N}_N(\boldsymbol{C}\boldsymbol{B}^T, \boldsymbol{R}), \tag{2.30}$$

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{0}, \sigma_{\beta}^2\boldsymbol{I}_p), \tag{2.31}$$

$$\boldsymbol{u} \mid \boldsymbol{G} \sim \mathcal{N}_{\boldsymbol{K}}(\boldsymbol{0}, \boldsymbol{G}), \tag{2.32}$$

$$\sigma_{\epsilon}^2 \sim \mathcal{IG}(A_{\epsilon}, B_{\epsilon}), \tag{2.33}$$

$$\sigma_{\boldsymbol{u}_l}^2 \sim \mathcal{IG}(A_{\boldsymbol{u}_l}, B_{\boldsymbol{u}_l}), \tag{2.34}$$

where $\boldsymbol{y}$ in equation (2.30) is an $N \times 1$ vector of response variables that is distributed

by an N-dimensional Gaussian, with $\boldsymbol{X}$ the design matrix of the fixed effects $\boldsymbol{\beta}$, $\boldsymbol{Z}$ the design matrix of the random effects $\boldsymbol{u}$ and $\boldsymbol{R}$ the covariance matrix of the model equal to $\sigma_\epsilon^2 \boldsymbol{I}_N$. The $\boldsymbol{C}$ and $\boldsymbol{B}$ matrices are created for simplification of the model: $\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}]$ and $\boldsymbol{B} = [\boldsymbol{\beta}^T\ \boldsymbol{u}^T]$. In equation (2.31), $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed regression coefficients and $\boldsymbol{u}$ in (2.32) is a $K \times 1$ vector of random effects with $K = \sum_l K_l$ and $l = 1, ..., r$. In equation (2.32), $\boldsymbol{G}$ is the covariance matrix of the random effects equivalent to blockdiag$(\sigma_{\boldsymbol{u}_1}^2 \boldsymbol{I}_{K_1}, ..., \sigma_{\boldsymbol{u}_r}^2 \boldsymbol{I}_{K_r})$.

To understand better the structure of this random effects model, we consider the example with the group of hospitals. $\boldsymbol{u}$ is a $K \times 1$ random effect vector comprised of r sub-vectors ($\boldsymbol{u} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_r]$). Each sub-vector, *i.e.*, $\boldsymbol{u}_1 = [u_{11}, u_{12}, ..., u_{K1}]$ corresponds to hospital 1, with elements $K1$ random effects (random intercept and/or random slope). Consequently, the design matrix $\boldsymbol{Z}$ is a $N \times K$ blockdiagonal with each block representing a hospital. The same block structure appears in the $\boldsymbol{G}$ covariance matrix of $\boldsymbol{u}$, as shown before.

With regards to the variational approximation of the linear mixed model, the product distribution that results in a tractable solution is

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_{\boldsymbol{u}_1}^2, ..., \sigma_{\boldsymbol{u}_r}^2, \sigma_\epsilon^2) = q(\boldsymbol{\beta}, \boldsymbol{u}) q(\sigma_{\boldsymbol{u}_1}^2) ... q(\sigma_{\boldsymbol{u}_r}^2) q(\sigma_\epsilon^2), \tag{2.35}$$

with

$$q(\boldsymbol{\beta}, \boldsymbol{u}) = \mathcal{N}_{p+K}(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}), \tag{2.36}$$

$$q(\sigma_{\boldsymbol{u}_l}^2) = \mathcal{IG}\left(A_{q(\sigma_{\boldsymbol{u}_l}^2)}, B_{q(\sigma_{\boldsymbol{u}_l}^2)}\right), \text{ with fixed } A_{q(\sigma_{\boldsymbol{u}_l}^2)} = A_{\boldsymbol{u}_l} + \frac{K_l}{2} \text{ and } 1 \le l \le r, \tag{2.37}$$

$$q(\sigma_\epsilon^2) = \mathcal{IG}\left(A_{q(\sigma_\epsilon^2)}, B_{q(\sigma_\epsilon^2)}\right), \text{ with fixed } A_{q(\sigma_\epsilon^2)} = A_\epsilon + \frac{N}{2}. \tag{2.38}$$

Algorithm 3 Coordinate Ascent for the variational Linear Mixed Regression Model

Initialize: $B_{q(\sigma_\epsilon^2)}$ and $B_{q(\sigma_{\boldsymbol{u}_l}^2)} \in \mathbb{R}^+$ with $l = 1, ..., r$

Repeat:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} = \left\{ \frac{A_\epsilon + \frac{N}{2}}{B_{q(\sigma_\epsilon^2)}} \boldsymbol{C}^T \boldsymbol{C} + \text{blockdiag}\left(\sigma_\beta^{-2} \boldsymbol{I}_p, \frac{A_{\boldsymbol{u}_1} + \frac{K_1}{2}}{B_{q(\sigma_{\boldsymbol{u}_1}^2)}} \boldsymbol{I}_{K_1}, ..., \frac{A_{\boldsymbol{u}_r} + \frac{K_r}{2}}{B_{q(\sigma_{\boldsymbol{u}_r}^2)}} \boldsymbol{I}_{K_r} \right) \right\}^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} = \left(\frac{A_\epsilon + \frac{N}{2}}{B_{q(\sigma_\epsilon^2)}} \right) \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^T \boldsymbol{y}$$

$$B_{q(\sigma_\epsilon^2)} = B_\epsilon + \frac{1}{2} \left\{ \left(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \right)^T \left(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \right) + \text{tr}\left(\boldsymbol{C}^T \boldsymbol{C} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \right) \right\}$$

$$B_{q(\sigma_{\boldsymbol{u}_l}^2)} = B_{\boldsymbol{u}_l} + \frac{1}{2} \left\{ \boldsymbol{\mu}_{q(\boldsymbol{u}_l)}^T \boldsymbol{\mu}_{q(\boldsymbol{u}_l)} + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_l)} \right) \right\} \text{ for } l = 1, ..., r$$

Stop:

$\mathcal{L}(\boldsymbol{y}; q)^{\text{current}} - \mathcal{L}(\boldsymbol{y}; q)^{\text{previous}} \le \epsilon$, where $\epsilon = 10^{-6}$

The variational joint density of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ in equation (2.36) is a $(p+K)$-dimensional Gaussian (since $\boldsymbol{\beta}$ is p-dimensional and $\boldsymbol{u}$ is K-dimensional) with mean vector $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and covariance matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$. The variational densities for the random effect variance $\sigma^2_{\boldsymbol{u}_l}$ and the fixed effect variance σ^2_ϵ are both Inverse-Gamma (equations (2.37) and (2.38)).

Algorithm 3 presents the variational updates for the linear mixed model, with initialization required for the scale variational matrices of $q(\sigma^2_\epsilon)$ and $q(\sigma^2_{\boldsymbol{u}_l})$.

2.5.3 Probit Regression Model

The probit regression model is suitable for the classification of binary data based on their predicted probability. It belongs to the family of the generalized linear models (Nelder and Baker [104], McCullagh and Nelder [89]) with probit link function the cumulative distribution of a standardized Normal, denoted as $\Phi(\cdot)$. The variational inference for the probit regression model is applied with the use of auxiliary variables (Holmes and Held [60]), since Gibbs sampling becomes tractable when these latent variables are incorporated (Albert and Chib [2]).

Bayesian Probit Regression Model: The probit regression likelihood is alternatively written as in equation (2.39) after the introduction of the auxiliary variables in (2.41). It is shown that each binary observation is dependent on the outcome of a continuous latent variable, which makes Bayesian inference feasible.

$$P(\boldsymbol{y} \mid \boldsymbol{z}) = \; I(\boldsymbol{z} \geq 0)^{\boldsymbol{y}} I(\boldsymbol{z} < 0)^{\mathbf{1}_N - \boldsymbol{y}}, \tag{2.39}$$

where $\boldsymbol{y}$ is a $N \times 1$ vector of independent response variables of either 1 when $z_n > 0$ or 0 when $z_n \leq 0$, as shown in equation (2.40). The auxiliary variable $\boldsymbol{z}$ in (2.41) is a normally distributed vector of independent variables denoted as z_n, with $1 \leq n \leq N$, fixed covariance matrix $\boldsymbol{I}_N$ and mean vector dependent of $\boldsymbol{\beta}$. In equation (2.42), $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients which follows a p-dimensional Gaussian with hyperparameters $\boldsymbol{\mu}_\beta$, $\boldsymbol{\Sigma}_\beta$.

$$y_n \mid z_n = \begin{cases} 1 & z_n > 0 \\ 0 & z_n \leq 0 \end{cases}, \tag{2.40}$$

$$\boldsymbol{z} \mid \boldsymbol{\beta} \sim \mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_N), \tag{2.41}$$

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \tag{2.42}$$

Note that in (2.39), $I(\cdot)$ is the index function and $\mathbf{1}_N$ the $N \times 1$ column vector with all entries equal to 1.

The factorization of the variational densities that leads to a tractable solution is

$$q(\boldsymbol{\beta}, \boldsymbol{z}) = q\boldsymbol{\beta})q(\boldsymbol{z}), \tag{2.43}$$

with

$$q(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{\Sigma}_{q(\beta)}), \tag{2.44}$$

$$q(\boldsymbol{z}) = \left[\frac{I(\boldsymbol{z} \geq \mathbf{0})}{\Phi(\boldsymbol{X}\boldsymbol{\mu}_{q(\beta)})}\right]^{\boldsymbol{y}} \left[\frac{I(\boldsymbol{z} < \mathbf{0})}{\mathbf{1}_N - \Phi(\boldsymbol{X}\boldsymbol{\mu}_{q(\beta)})}\right]^{\mathbf{1}_N - \boldsymbol{y}} \times \mathcal{N}_N(\boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}, \boldsymbol{I}_N). \tag{2.45}$$

In equation (2.44), the variational posterior of $\boldsymbol{\beta}$ is a p-dimensional Gaussian with mean vector $\boldsymbol{\mu}_{q(\beta)}$ and covariance matrix $\boldsymbol{\Sigma}_{q(\beta)}$. Regarding the variational posterior of z in (2.45), it is now a truncated N-dimensional Gaussian,. For the Gaussian density part, the variational parameters are the mean linear predictor $\boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}$ and the identity covariance matrix (independence across z_n's).

Algorithm 4 gives all the closed form variational equations for the distributions' parameters in (2.44) and (2.45). $\boldsymbol{\mu}_{q(z)}$ is the truncated Gaussian variational parameter. $\phi(\cdot)$ in $\boldsymbol{\mu}_{q(z)}$ denotes the probability distribution function of the standard Gaussian.

Algorithm 4 Coordinate Ascent for the variational Probit Regression Model

Initialize: $\boldsymbol{\mu}_{q(z)}$

Repeat:

$$\boldsymbol{\mu}_{q(\beta)} = \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_\beta^{-1}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{\mu}_{q(z)} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right)$$

$$\boldsymbol{\mu}_{q(z)} = \boldsymbol{X}\boldsymbol{\mu}_{q(\beta)} + \frac{\phi\left(\boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}\right)}{\left\{\Phi(\boldsymbol{X}\boldsymbol{\mu}_{q(\beta)})\right\}^{\boldsymbol{y}}\left\{\mathbf{1}_N - \Phi(\boldsymbol{X}\boldsymbol{\mu}_{q(\beta)})\right\}^{\mathbf{1}_N - \boldsymbol{y}}}$$

Stop:

$\mathcal{L}(\boldsymbol{y}; q)^{\text{current}} - \mathcal{L}(\boldsymbol{y}; q)^{\text{previous}} \leq \epsilon$, where $\epsilon = 10^{-6}$

2.5.4 Probit Mixed Regression Model

The probit mixed regression effects model can be used in cases where the covariates are grouped to one or more classification factors. It performs as a multi-level generalized model and is suitable for binary response variables (Baragatti [5]). As the name indicates, it combines the probit regression and the random effects models, hence we borrow the notation from both of them and derive the model from scratch.

Bayesian Probit Mixed Regression Model: The probit mixed likelihood in equation (2.46) introduces, as in the probit model, the auxiliary variables $\boldsymbol{z}$ shown in (2.47) which are also dependent on the random effects $\boldsymbol{u}$ specified in equation (2.49), apart from the fixed effects $\boldsymbol{\beta}$ shown in equation (2.48).

$$P(\boldsymbol{y} \mid \boldsymbol{z}) = I(\boldsymbol{z} \geq \mathbf{0})^{\boldsymbol{y}} I(\boldsymbol{z} < \mathbf{0})^{\mathbf{1}_N - \boldsymbol{y}}, \tag{2.46}$$

$$\boldsymbol{z} \mid \boldsymbol{\beta}, \boldsymbol{u} \sim \mathcal{N}_N(\boldsymbol{C}\boldsymbol{B}^T, \boldsymbol{I}_N). \tag{2.47}$$

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \tag{2.48}$$

$$\boldsymbol{u} \mid \boldsymbol{G} \sim \mathcal{N}_{\boldsymbol{K}}(\boldsymbol{0}, \boldsymbol{G}), \tag{2.49}$$

$$\sigma^2_{\boldsymbol{u}_l} \sim \mathcal{IG}(A_{\boldsymbol{u}_l}, B_{\boldsymbol{u}_l}). \tag{2.50}$$

Regarding the variational product distribution, this is a combination of the variational posteriors of the probit and the linear mixed regression model, with the only difference appearing on the absence of the residual variance σ^2_ϵ because the data are Bernoulli and not normally distributed.

$$q(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_{\boldsymbol{u}_1}, ..., \sigma^2_{\boldsymbol{u}_r}, \boldsymbol{z}) = q(\boldsymbol{\beta}, \boldsymbol{u})q(\sigma^2_{\boldsymbol{u}_1})...q(\sigma^2_{\boldsymbol{u}_r})q(\boldsymbol{z}), \tag{2.51}$$

with

$$q(\boldsymbol{\beta}, \boldsymbol{u}) = \mathcal{N}_{p+K}(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}), \tag{2.52}$$

$$q(\sigma^2_{\boldsymbol{u}_l}) = \mathcal{IG}\left(A_{q(\sigma^2_{\boldsymbol{u}_l})}, B_{q(\sigma^2_{\boldsymbol{u}_l})}\right), \text{ with } 1 \leq l \leq r, \tag{2.53}$$

$$q(\boldsymbol{z}) = \left[\frac{I(\boldsymbol{z} \geq \boldsymbol{0})}{\Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})}\right]^{\boldsymbol{y}} \left[\frac{I(\boldsymbol{z} < \boldsymbol{0})}{\boldsymbol{1}_N - \Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})}\right]^{\boldsymbol{1}_N - \boldsymbol{y}} \times \mathcal{N}_N(\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{I}_N). \tag{2.54}$$

The joint approximated density of $(\boldsymbol{\beta}, \boldsymbol{u})$ in equation (2.52) is a multivariate Gaussian with p+K dimensions (since $\boldsymbol{\beta}$ is a $p\times1$ vector and $\boldsymbol{u}$ a $K\times1$), with variational parameters $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$. In equation (2.53), the scalar $\sigma^2_{\boldsymbol{u}_l}$, related to the variance of the random effects, is approximated by an Inverse-Gamma with parameters $A_{q(\sigma^2_{\boldsymbol{u}_l})}$, $B_{q(\sigma^2_{\boldsymbol{u}_l})}$. The variational density of the auxiliary variables $\boldsymbol{z}$ in equation (2.54) is a truncated N-dimensional Gaussian (independence across the dimensions) with variational mean vector $\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, where $\boldsymbol{C}$ encompasses both fixed and random effects design matrices $\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}]$.

Algorithm 5 Coordinate Ascent for the variational Probit Mixed Regression Model

Initialize: $\boldsymbol{\mu}_{q(\boldsymbol{z})}$ and $B_{q(\sigma^2_{\boldsymbol{u}_l})} \in \mathbb{R}^+$ with $l = 1, ..., r$

Repeat:

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} = \left\{\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}\left(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}}, \frac{A_{\boldsymbol{u}_1} + \frac{K_1}{2}}{B_{q(\sigma^2_{\boldsymbol{u}_1})}}\boldsymbol{I}_{K_1}, ..., \frac{A_{\boldsymbol{u}_r} + \frac{K_r}{2}}{B_{q(\sigma^2_{\boldsymbol{u}_r})}}\boldsymbol{I}_{K_r}\right)\right\}^{-1} \times$$
$$\left\{\boldsymbol{C}^T\boldsymbol{\mu}_{q(\boldsymbol{z})} + \text{blockdiag}\left(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}}, \frac{A_{\boldsymbol{u}_1} + \frac{K_1}{2}}{B_{q(\sigma^2_{\boldsymbol{u}_1})}}\boldsymbol{I}_{K_1}, ..., \frac{A_{\boldsymbol{u}_r} + \frac{K_r}{2}}{B_{q(\sigma^2_{\boldsymbol{u}_r})}}\boldsymbol{I}_{K_r}\right)\boldsymbol{M}^T\right\}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{z})} = \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \frac{\phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})}{\left\{\Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right\}^{\boldsymbol{y}}\left\{\boldsymbol{1}_N - \Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})})\right\}^{\boldsymbol{1}_N - \boldsymbol{y}}}$$

$$B_{q(\sigma^2_{\boldsymbol{u}_l})} = B_{\boldsymbol{u}_l} + \frac{1}{2}\left\{\boldsymbol{\mu}^T_{q(\boldsymbol{u}_l)}\boldsymbol{\mu}_{q(\boldsymbol{u}_l)} + \text{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_l)}\right)\right\}, \text{ for } l = 1, ..., r$$

Stop:

$\mathcal{L}(\boldsymbol{y}; q)^{\text{current}} - \mathcal{L}(\boldsymbol{y}; q)^{\text{previous}} \leq \epsilon$, where $\epsilon = 10^{-6}$

Algorithm 5 presents the Coordinate Ascent scheme for the update of the variational parameters in equations (2.52)-(2.54). The initialization concerns the mean variational vector of $\boldsymbol{z}$, $\boldsymbol{\mu}_{q(\boldsymbol{z})}$, which can take any values in the real space (we usually set it equal to $\mathbf{0}$ to express our agnostic opinion) and the scale variational parameter of the random effect variance, $B_{q(\sigma^2_{\boldsymbol{u}_l})}$, which can be any non-negative value.

2.5.5 Summary on Variational Regression Models

In this section, we provided the variational iterative schemes for a variety of regression models. Specifically, we presented the closed form equations for updating the variational parameters of the linear regression model (single and multi-response), the linear mixed, the probit regression and the probit mixed model. The Evidence Lower Bounds are also calculated for each model and supplied in Appendix A.

As a general conclusion, Variational Bayes can be easily applied to conjugate regression models and it is a fast alternative to MCMC, which samples iteratively from the full conditionals, whereas the variational algorithm updates closed form variational parameters until convergence of the ELBO is reached. In cases of non-conjugate models, further approaches like Taylor approximations can be exploited to achieve conjugacy, with a successful example being the Beta mixtures in Chapter 3, Subsection 3.2.1.

2.6 Mixture Models

Mixture models are a useful parametric tool for unsupervised clustering (Fraley and Raftery [46]), with unsupervised referring to algorithms that try to reveal any hidden group structures in the data. Mixture models are based on probabilistic principles while they admit flexibility in choosing the sub-populations' distribution. A detailed survey on the theory and applications of mixture models can be found in Titterington et al. [138] and McLachlan et al. [92]. On the other hand, K-means (MacQueen et al. [82]) and Hierarchical clustering (Ward Jr [143]) are non-probabilistic and thus unable to assume a component distribution. Moreover, a mixture model, in contrast to K-means and Hierarchical, does not exclusively allocate with probability one an observation into a group (hard clustering). Specifically, for each datapoint, a mixture model returns a vector of allocation probabilities, called responsibilities, introducing the level of confidence in assigning this point into each one of the components. For further information refer to Titterington et al. [138], Lindsay [78] and McLachlan and Basford [90].

With regards to the number of components in a mixture model, there are cases where the data model may bear infinite in magnitude sub-distributions instead of finite.

These are the so-called Dirichlet Process mixture models that live in the large family of Bayesian non-parametric models, where their infinite dimensional parameter space can grow with the sample size (Gershman and Blei [49]). For the Dirichlet Process scenario, schemes like the stick-breaking point or the Chinese restaurant process are utilized (Teh et al. [133]) to facilitate inference.

The reason we meticulously study mixture models is because they offer an elegant and applicable way to making inference on DNA datasets regarding the number and heterogeneity of the hidden groups. For instance, individuals can be allocated into groups based on their rate of DNA methylation (the rates refer to beta-intensities, described in Chapter 1, Subsection 1.5.2). In this framework, a mixture model of Beta distributions could have been an appropriate approach to cluster the subjects, whereas in cases where methylation counts are recorded, as in Chapter 1, Subsection 1.5.1, a Poisson mixture model would be more suitable. This flexible choice of parametric densities distinguishes them from the non-probabilistic K-means and Hierarchical clustering.

2.6.1 Finite Mixture Models

Mixture models have been attracting the interest of researchers in statistics and machine learning as clustering tools in unsupervised settings, since McLachlan and Basford [90] firstly introduced such models as a simple way in determining the hidden number of groups on a dataset. For an up-to-date work on the theory and methodological development of mixture models see McLachlan et al. [92].

As regards the literature, the most common type of mixture model for clustering is the mixture of finite numbers of Gaussian densities. One main reason is that most features seem to follow a normal shape (height, weight etc.) and therefore, an assumption of a Gaussian density might not deviate much from the reality. Another complement reason is the unbounded support range of the distribution, which offers non-restrictions on the area the observation can live. In addition, Gaussian enjoys a plethora of important mathematical properties that ease the inference and conclude to intelligible outcomes.

The formulation of the Bayesian multivariate Gaussian mixture model with fixed number of components can be found in Bishop [11], who introduces the latent allocation parameter $\boldsymbol{z}$ to facilitate the mathematical derivation.

$$\boldsymbol{y}_n \mid \boldsymbol{z}_n, \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m \sim \prod_{m=1}^{M} \mathcal{N}_D(\boldsymbol{y}_n \mid \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)^{z_{nm}}, \tag{2.55}$$

$$(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m) \sim \mathcal{NW}_D(\boldsymbol{\mu}_{0m}, \boldsymbol{\beta}_{0m}, \boldsymbol{W}_{0m}, \boldsymbol{\nu}_{0m}), \tag{2.56}$$

$$\boldsymbol{z}_n \mid \boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi}), \tag{2.57}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\phi}_0), \tag{2.58}$$

where $\boldsymbol{y}_n$ is the n^{th} D-dimensional datapoint distributed as a mixture of M multivariate Gaussian densities (M fixed), with component specific mean vector $\boldsymbol{\mu}_m$ and precision matrix $\boldsymbol{\Lambda}_m$. The joint prior density of the m^{th} component's parameters $(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is a Normal-Wishart, with D dimensions and hyperparameters $\boldsymbol{\mu}_{0m}$ (mean vector), $\boldsymbol{\beta}_{0m}$ (real valued coefficient of the precision $\boldsymbol{\Lambda}_m$ that links it to $\boldsymbol{\mu}_m$), scale covariance matrix $\boldsymbol{W}_{0m}$ and degrees of freedom $\boldsymbol{\nu}_{0m}$. For each $\boldsymbol{y}_n$, a latent variable $\boldsymbol{z}_n$ exists comprised of $M \times 1$ elements, with the $M-1$ values being 0 and the one corresponding to $\boldsymbol{y}_n$'s cluster being equal to 1. Hence, the latent allocation parameter $\boldsymbol{z}_n$ can follow *a priori* a Categorical distribution with parameter vector $\boldsymbol{\pi} = [\pi_1, ..., \pi_M]$, while the component weights $\boldsymbol{\pi}$ are Dirichlet distributed with concentration parameter $\boldsymbol{\phi}_0$.

Regardless the flexibility of the Finite mixtures in clustering, the assumption of the presence of a fixed number of unobserved groups restricts model-determination. One way to determine the best fitting model could be to compare the performance for differing numbers of clusters. However, this is time-consuming, especially when the inferential algorithm requires long time to converge (*i.e.*, in high-dimensional data structures).

2.6.2 Dirichlet Process

To overcome the problem of model-determination in Finite mixtures with fixed M, a stochastic process called Dirichlet process (DP) is utilized to introduce the Bayesian non-parametric Dirichlet Process mixtures, allowing flexibility with respect to the unknown number of components. An in depth material presentation regarding Dirichlet Process can be found in Teh [132].

In general, a Dirichlet Process is a distribution over distributions, meaning that every draw from such a process is a probability distribution. It is named after the Dirichlet distributed finite dimensional marginal distributions and it is a popular way in clustering procedures while simultaneously determining the number of components. Elicited from El-Arini [38], the Dirichlet Process is described as follows.

Let G be DP distributed

$$G \sim \text{DP}(\phi, G_0), \tag{2.59}$$

where G_0 is a base distribution and ϕ a positive definite scaling parameter. G is a random probability measure that has the same domain as G_0.

In Figure 2.1, we consider a continuous base distribution G_0, *i.e.*, a Gaussian (red curve), while the sampled distribution G is discrete, constructed out of countably

infinite number of point masses (blue vertical lines). A set is proved to be countably infinite when it has the same cardinality as the natural numbers $\mathbb{N}$. We claim that G is DP distributed over B, with parameters ϕ, G_0, if for any finite set of partitions $S_1 \cup S_2 \cup ... \cup S_i \in B$

$$(G(S_1), ..., G(S_i)) \sim \text{Dirichlet}\left(\phi G_0(S_1), ..., \phi G_0(S_i)\right). \tag{2.60}$$

At this point, we explain that G_0 is the mean distribution of the Dirichlet Process. For example, for any measurable subset of B, here S_1, $\mathbb{E}[G(S_1)] = G_0(S_1)$. In regard to the scale parameter ϕ, it is associated with the DP variance: $\mathbb{V}\text{ar}[G(S_1)] = G_0(S_1)(1 - G_0(S_1))/(\phi + 1)$, where high values of ϕ imply low variance and consequently higher concentration of the G sample densities around G_0. Concerning G, there are constructive ways to build its form such as the Chinese restaurant process, and the stick-breaking point which exploits the discreteness of G by composing a weighted sum of points masses (Teh [132], Sethuraman [123]). In this thesis, we focus only on the latter one.

For further ways of G construction refer to Teh [132].

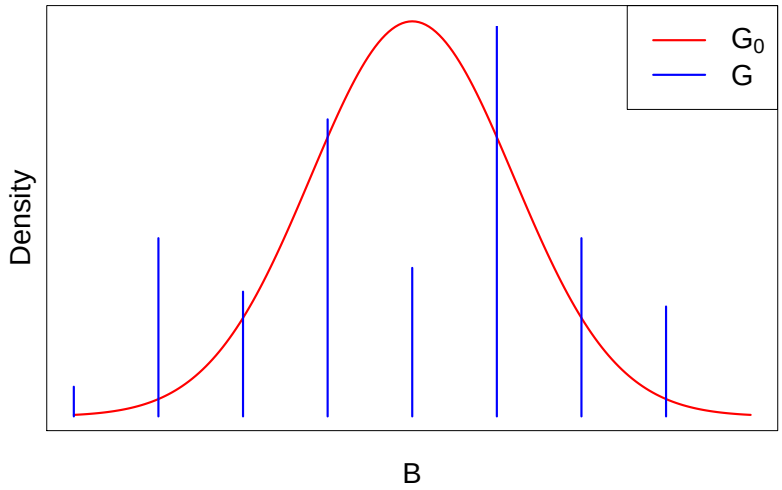


Figure 2.1 Example of a random sample distribution G from a Dirichlet Process, when G_0 is a univariate Gaussian. G is a discrete random draw from a Dirichlet Process (blue point masses), while G_0 is the Gaussian base distribution of this Dirichlet Process (red density). x-axis represents the sample space of G, denoted as B.

2.6.3 Stick-breaking Point Representation

The stick-breaking point is a method for constructing the form of the discrete Dirichlet Process distributed G. In particular, infinite number of point masses are produced through the stick-breaking scheme to form the G distribution. The point masses of G sum up to one - requirement for G to be a proper probability function - and hence each one can work as a mixing weight to a mixture model. Specifically, the m^{th} point mass

can correspond to the m^{th} component's mixing weight π_m, linking Dirichlet Process to the infinite mixture models (infinite number of components). However, it is impossible to represent fully an infinite model on computers and therefore, a high truncation level M is set instead. Consequently, M components are estimated.

Algorithm 6 Constructive Scheme of $\boldsymbol{\pi}$

1. Draw b_1 from G_0
2. Draw w_1 from Beta$(1, \phi)$
3. $\pi_1 = w_1$
4. Draw b_2 from G_0
5. Draw w_2 from Beta$(1, \phi)$
6. $\pi_2 = w_2(1 - w_1)$
7. ...
8. Draw b_M from G_0
9. Draw w_M from Beta$(1, \phi)$
10. $\pi_M = w_M(1 - w_{m-1})(1 - w_{m-2})...(1 - w_1)$

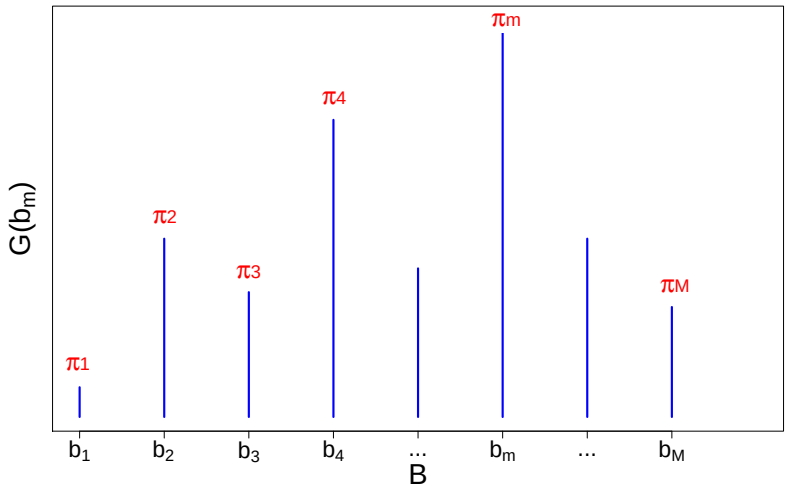


Figure 2.2 The resulted discrete form of a Dirichlet Process distribution, denoted as G, after the stick-breaking point implementation. The points on the x-axis have been sampled from Beta$(1, \phi)$ and belong to the support range B of the $G(\cdot)$ distribution. The point masses represent the mixing weights of the Dirichlet Process mixture model.

In Algorithm 6, the constructive stick-breaking point scheme of the mixing weights $\boldsymbol{\pi}$ is provided, with π_m being the probability mass of the m^{th} component. The goal is to build the simplex $\boldsymbol{\pi} = [\pi_1, ..., \pi_M]$ by breaking the $[0, 1]$ interval into M sub-intervals. These sub-intervals are not necessarily equally spaced since their cutting length is based upon independently sampling w_m from Beta$(1, \phi)$. Therefore, the variable w_m represents the cut length of the m^{th} sub-interval. In parallel, we draw samples b_m from G_0 (base distribution), since it shares the same support range with G. These b_m

samples will be the G inputs with mass equal to π_m. Figure 2.2 displays an example form of the G distribution, with the height of the point masses being equal to the corresponding component weight π_m.

The result of this operation is a discrete distribution with respect to the mixing weights π_m and the G_0 samples b_m

$$G = \sum_{m=1}^{M} \pi_m(\boldsymbol{w})\delta_{b_m}, \quad \text{with} \quad \pi_m(\boldsymbol{w}) = w_m \prod_{k=1}^{m-1}(1 - w_k), \tag{2.61}$$

where $\delta_{b_m} = 1$ at b_m and 0 elsewhere. The distribution over $\boldsymbol{\pi}$ is often called GEM, $\boldsymbol{\pi} \sim \text{GEM}(\phi)$, with the initials standing for Griffiths, Engen and McCloskey (Pitman [111]).

2.6.4 Dirichlet Process Mixture Model

The general hierarchical Dirichlet Process mixture model can now be defined under the stick-breaking point representation

$$\boldsymbol{y}_n \mid \boldsymbol{z}_n, \boldsymbol{\theta}_m \sim \prod_{m=1}^{M} f_D(\boldsymbol{y}_n \mid \boldsymbol{\theta}_m)^{z_{nm}}, \tag{2.62}$$

$$\boldsymbol{\theta}_m \sim G_0, \tag{2.63}$$

$$\boldsymbol{z}_n \mid \boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi}), \tag{2.64}$$

$$\boldsymbol{\pi} \sim \text{GEM}(\phi), \tag{2.65}$$

where $\boldsymbol{y}_n$ in equation (2.62) is the n^{th} D-dimensional datapoint distributed as a mixture of M $f(\cdot)$ discrete or continuous distributions, $\boldsymbol{\theta}_m$ are the model parameters for the m^{th} component that follow the DP base distribution $G_0(\cdot)$ (see equation (2.63)), $\boldsymbol{z}_n$ in equation (2.64) is the latent allocation vector for the n^{th} datapoint distributed as a Categorical and $\boldsymbol{\pi}$ is the GEM distributed vector of mixing weights (see equation (2.65)).

2.7 Feature Selection

The next step after the definition of the mixture model is to apply an inferential algorithm in order to retrieve the true number of components, the mixing weights and the component variational distributions. Variational Bayes is one of the candidate algorithms that we successfully employ for this complex task. Given its output, we propose an extra yet informative step concerning a feature selection scheme *per* cluster.

Based on Lin et al. [76] and Lin [75], we suggest exploiting the fitted variational distributions as means to calculate the posterior discriminative accuracy measure. This measure defines those features that significantly contribute in composing each cluster.

To apply this feature selection setting, the mixture model with M components and component specific parameters $\boldsymbol{\theta}_m$, for $m = \{1, ..., M\}$ is first estimated and then presented in the form of

$$g(\boldsymbol{y} \mid \boldsymbol{\theta}) = \pi_1 f_1(\boldsymbol{y} \mid \boldsymbol{\theta}_1) + ... + \pi_M f_M(\boldsymbol{y} \mid \boldsymbol{\theta}_M), \tag{2.66}$$

with $\boldsymbol{y}$ being the $N \times D$ data matrix, π_m the m^{th} variational mixing weight and $f_m(\cdot)$ the D-dimensional variational distribution of the m^{th} component. For notation simplicity, the dependence on the parameters in $f_m(\cdot)$ will be implicit this point onward, *i.e.*, $f_1(\boldsymbol{y} \mid \boldsymbol{\theta}_1)$ will be presented as $f_1(\boldsymbol{y})$. As $f_{/m}$ we represent the conditional mixture, which corresponds to all the remaining components (along with their weights) apart from the $\pi_m f_m$ term. For instance, if $m = 1$

$$f_{/1}(\boldsymbol{y}) \equiv \pi_2 f_2(\boldsymbol{y}) + ... + \pi_M f_M(\boldsymbol{y}) = g(\boldsymbol{y}) - \pi_1 f_1(\boldsymbol{y}), \tag{2.67}$$

with $g(\boldsymbol{y})$ being the full mixture model.

In addition, after considering equations (2.66) and (2.67), we define a few useful quantities which facilitate the calculation of the final discriminative measure for the m^{th} cluster

$$\delta_m = \int_{\boldsymbol{y}} f_m(\boldsymbol{t}) f_{/m}(\boldsymbol{t})\, d\boldsymbol{t}, \tag{2.68}$$

$$\Delta_m = \int_{\boldsymbol{y}} f_m(\boldsymbol{t}) g(\boldsymbol{t})\, d\boldsymbol{t}, \tag{2.69}$$

$$d_m = \frac{\delta_m}{\Delta_m}, \tag{2.70}$$

and

$$\tilde{\pi}_m(\boldsymbol{y}) = \frac{\pi_m f_m(\boldsymbol{y})}{g(\boldsymbol{y})}. \tag{2.71}$$

Equations (2.68) and (2.69) remind of a concordance index which naturally measures the agreement/overlapping of two densities. Scott and Szewczyk [122] discussed about the closeness (concordance) of two densities based on similarity distances. In particular, δ_m and Δ_m play the role of similarity measures, with values near 0 indicating non-agreement of the two densities, whereas higher ones better agreement. Regarding d_m in equation (2.70), it is an index between $[0, 1]$ which defines the level of discrimination of group m from the rest components, with low levels implying good discrimination and high the opposite. As for $\tilde{\pi}_m(\boldsymbol{y})$ in equation (2.71), it represents the probability of correctly classifying a data-vector $\boldsymbol{y}$ into the m^{th} component.

Having defined the aforementioned quantities, we move on to calculating the expected true-positive π_{m^+} and false-positive π_{m^-} classification rates

$$\pi_{m^+} = \mathbb{E}[\tilde{\pi}_m(\boldsymbol{y}) \mid \boldsymbol{y} \sim f_m], \tag{2.72}$$

$$\pi_{m^-} = \mathbb{E}[\tilde{\pi}_m(\boldsymbol{y}) \mid \boldsymbol{y} \sim f_{/m}]. \tag{2.73}$$

Substituting equation (2.71) into π_{m^+} and π_{m^-} and then using the first-order approximation of the ratio of the two expectations (proofs in the supplementary material of Lin et al. [76]), the result is

$$\pi_{m^+} \approx \frac{\mathbb{E}[\pi_m f_m(\boldsymbol{y}) \mid \boldsymbol{y} \sim f_m]}{\mathbb{E}[g(\boldsymbol{y}) \mid \boldsymbol{y} \sim f_m]} = \tau_{m^+}, \tag{2.74}$$

$$\pi_{m^-} \approx \frac{\mathbb{E}[\pi_m f_m(\boldsymbol{y}) \mid \boldsymbol{y} \sim f_{/m}]}{\mathbb{E}[g(\boldsymbol{y}) \mid \boldsymbol{y} \sim f_{/m}]} = \tau_{m^-}. \tag{2.75}$$

High values of τ_{m^+} and low of τ_{m^-} denote good discrimination of the m^{th} component from the rest $M-1$ in the mixture. Hence, they are respectively called true- and false- positive discriminative threshold probabilities for assigning data-points into the corresponding cluster. Using trivial algebra, we prove that

$$\tau_{m^+} = 1 - d_m, \tag{2.76}$$

$$\tau_{m^-} = \frac{\pi_m \delta_m}{\int\limits_{\boldsymbol{y}} g^2(\boldsymbol{t}) d\boldsymbol{t} - \pi_m \Delta_m}. \tag{2.77}$$

In cases of intractable integrals in equations (2.76) and (2.77), we use numerical calculation over the observed data points $\boldsymbol{y}$. Note that these τ_m measures can be computed for each feature dimension separately (or subsets of features). As an example, $\tau_{m^+}(1)$ would indicate the calculation of the τ_{m^+} probability based on the first feature, while $\tau_{m^+}([1,4,5])$ based on the first, fourth and fifth feature. Therefore, a general notation is $\tau_{m^+}(h)$ and $\tau_{m^-}(h)$, with $h \subseteq \{1:D\}$ the subset of features.

By taking advantage of $\tau_{m^+}(h)$ and $\tau_{m^-}(h)$, we can compute the weighted discriminative threshold probability for classification into the m^{th} component, referred as the aggregate discriminative accuracy measure

$$A_m(h) = \pi_m \tau_{m^+}(h) + (1 - \pi_m)\tau_{m^-}(h). \tag{2.78}$$

In equation (2.78), $A_m(h)$ is a rate that shows in what extend subset h characterizes the m^{th} component. Values close to 1 manifest that features h are the only necessary ones in creating component m. On the contrary, values near zero show lack of h contribution.

In practice, to find the optimal set of features that discriminate the m^{th} component from the rest, we have to compute $A_m(h)$ for all the possible $2^D - 1$ subsets (D the number of features) and select the set that maximizes $A_m(\cdot)$. In small case scenarios where $D = 2$ or 3 or 4, it is feasible to enumerate all the subsets and compute the measure for each one. On the other hand, when D is large the situation turns difficult. To give an

instance, if $D = 10$ all the possible calculations are $1,024 - 1 = 1,023$, let alone for higher feature dimensions ($1,048,575$ subsets for $D = 20$). On account of that, we alternatively build a forward selection algorithm that adds sequentially features. In particular, the algorithm starts at the first iteration by searching the feature that maximizes $A_m(\cdot)$. At the second iteration, it keeps the selected feature from the previous iteration and searches for the second feature that along with the previous feature maximize $A_m(\cdot)$. This procedures carries on until the algorithm converges (Algorithm 7). At this point we highlight two things: a) the maximum $A_m(h)$ of the current iteration may be higher, equal or even lower than the value in the previous iteration but at convergence the maximum $A_m(h)$ needs to be higher than all the previous iterations, and b) the maximum $A_m(h)$ at the convergence point may not necessarily reach the value 1 (100% discriminative accuracy), but it needs to be the highest possible. Alternative methods such as Stochastic selection algorithms used for evaluating subsets in regression (Hans et al. [55]) can be also investigated, when forward selection has slow progression caused by the high amount of features. However, we exclusively work with forward selection due to its efficient execution in our applications.

Algorithm 7 Forward Selection of Discriminative Features for the m^{th} Component

Fix:

$m \in \{1, 2, ..., M\}$

Initialize:

$h = \{1 : D\}$ and $k = \{\varnothing\}$

Repeat:

1) For l in h: compute $A_m([l, k])$ and select $l^\star = \arg\max_{l \in h} A_m([l, k])$

2) Update $h = \{h \neq l^\star\}$

3) Update $k = [k, l^\star]$

4) Return $A_m(k)^{\text{current}}$

Stop:

$|A_m(k)^{\text{current}} - A_m(k)^{\text{previous}}| \leq \epsilon$, where $\epsilon = 10^{-3}$

The forward selection scheme is presented in Algorithm 7. Specifically, we start by fixing the component label m for which the discriminative features need to be found. At next, we initialize the set of features h to consist of all the D features, and finally we create an empty set k that will store the selected features at each iteration. The algorithm computes the discriminative accuracy measure $A_m([l, k])$ for each element of h sequentially and separately (note that the set h is smaller than D after the first iteration), denoted as l, along with the k set at the current iteration. It then selects that l element, $l^\star$, that maximizes the measure. The set h is then updated to contain all the D features except $l^\star$. Regarding k, it uploads $l^\star$ while retaining its previous values. The algorithm stops when the absolute difference between the current maximum

$A_m(k)$ is lower than the previous by a negligible value, like $\epsilon = 10^{-3}$ (convergence). We choose as stopping criterion the 10^{-3} difference to avoid overcrowding the optimal set of features with features that add less than 0.001 improvement in the discriminative accuracy. The algorithm's output is the selected k set of features that maximizes the discriminative accuracy measure for the m^{th} component.

2.8 Summary

In Chapter 2, we presented the theory of Variational Inference with focus on the Mean Field approximation, also known as Variational Bayes. The optimal general variational distribution was formed and then derived for popular regression models, revealing the applicability of this inferential method and preparing the ground for applications in complex mixture scenarios in Chapter 3. Additionally, we introduced a simple approach to dealing with poor variational initialization by smoothing out non-convex lower bounds for a few iterations. Furthermore, we discussed about Finite mixture models and their utility as model-based clustering tools, which although lack model-determination. To overcome this issue, we presented the Dirichlet Process, opening the path to the Dirichlet Process mixture models. Finally, we provided a measure for component discrimination that exploits the fitted variational distributions and returns those features that discriminate each component from the rest.

Chapter 3

Variational Mixture Models

3.1 Overview

Having defined Variational Bayes, derived the Mean Field algorithm for a variety of regression models in Chapter 2 and described the principles of the Finite and Dirichlet Process mixtures, we have all the necessary tools and knowledge to elaborate the full Mean Field methodology for notable discrete and continuous mixture models.

Our goal in this chapter is to make available the mathematical implementation for a wide range of fast model-based clustering tools, due to the demand in determining the hidden groups of non-normally distributed DNA methylation data, such as beta-intensities derived from array-based platforms (see Chapter 1, Subsection 1.5.2), or methylated counts from Bisulfite Sequencing techniques (see Chapter 1, Subsection 1.5.1). Given this chapter, the user will be able to choose the tool that suits better her data type and straightforwardly program it in a language of her preference. The mathematical derivations concern both Finite and Dirichlet Process mixtures, however emphasis is given on the latter due to its ability in automatically determining the number of clusters.

In general, we analyze models where the likelihood is a mixture of

- Gaussian densities, ideal for data with unrestricted support range
- Beta densities, ideal for data with bounded support range
- Bernoulli/Binomial distributions (with or without confounding parameters such as sex, age, ethnicity and other demographic factors), ideal for binary data or counts with known number of independent trials/experiments
- Poisson distributions (with or without confounding parameters), ideal for counts with unknown/non-fixed number of trials

Regarding the input data in the aforementioned variational mixture algorithms, these are an $N \times D$ matrix, with N being the number of samples and D the number of features. Each row of the input data matrix corresponds to a sample, while each column to a feature for the specific sample, with all features measured in the same units, *i.e.*, the level of DNA methylation is recorded for D differentially methylated genomic regions in N individuals. We also assume that the observations between samples are independent, as well as the measurements within each sample (features) and therefore, our variational mixture models are built accordingly. Generally, this is a very strong assumption, especially when we consider CpGs, and we recognize that it may not be completely true. However, we do it for computational reasons (Zhang et al. [156]). Moreover, in our real examples we are working with differentially methylated regions (DMRs) and CpGs within genes that span the entire genome, hence the assumption of independence may not be violated. To illustrate our independence assumption, we use the Beta mixture model. In this model, each sub-population of samples is distributed as a Beta density with component specific parameters that also vary across the feature dimensions.

Mixtures	**Finite**		**Dirichlet Process**	
	with covariates	**without covariates**	**with covariates**	**without covariates**
Beta	-	Easy to derive	-	Main text (3.2.1)
Gaussian	-	Easy to derive	-	Appendix (B.3.1)
Bernoulli/Binomial	Main text (3.3.2)	Main text (3.3.1)	Easy to derive	Appendix (B.2.2)
Poisson	Easy to derive	Appendix (B.1.1)	Main text (3.3.3)	Appendix (B.2.1)

Table 3.1 Mixture models for which the variational derivation is provided either on the main text or the Appendix B, or it can be straightforwardly derived based on the provided material. Dash lines imply non supply of the mathematical procedure for the corresponding model.

In Table 3.1, we arrange all the analyzed mixture models into two main categories: 1) Finite and 2) Dirichlet Process, which further split in two sub-categories: a) with covariates and b) without covariates. "With covariates" are those models that take into consideration the occurrence of confounding parameters, *i.e.*, sex, age etc., that may produce spurious clustering, thus we remove their impact. The "without covariates" models assume no presence of external factors that can distort the clustering process. Taking into account covariates is a great novelty and only few algorithms are constructed to do it (Carvalho et al. [20]).

In this chapter, we provide the complete variational derivation of only a representative segment of the aforementioned models in order to avoid excessive amount of technicalities. Regarding "annealing", the temperature addition that accounts for poor initialization in Chapter 2, we introduce its easy implementation in only one of the presented models to circumvent repeating the same procedure (multiplying

log-likelihood by a constant) and also to maintain clarity on the variational steps in the original Bayesian model.

Generally, one model from each genre is presented so as to cover all the different mathematical approaches. Those models are indicated as "Main text" in Table 3.1, whilst in the Appendix B are stored those with the "Appendix" specification. In our real application analysis in Chapter 5, we focus on Dirichlet Process mixtures and not on Finite, since only the former provide model determination. Hence, we mostly omit the Finite mixture model derivation here and supply it in the Appendix B, although the variational algorithm can be comfortably derived in accordance with the existing material. The easily derived variational models, given the knowledge of Chapter 3, are denoted as "Easy to derive". Lastly, the dash lines represent continuous models for which we have not proceeded with the variational implementation, and these concern cases of existence of covariates. The reason for skipping their derivation is because there are alternative ways to deal with such scenarios, like clustering the residuals of an appropriate regression model with predictors the confounding parameters. The residuals can work as a clear representation of the original data, since they are free from factors that distort the clustering outcome.

To summarize this chapter, we deliver the variational Finite mixture of Binomial densities (with "annealing"), the variational Dirichlet Process Poisson mixture when covariates exist, as well as the Finite Bernoulli mixture with covariates, while we get started with a detailed derivation of the variational Dirichlet Process mixture of Beta densities. With regards to the Gaussian mixtures, we choose not to present the mathematical procedure on the main text because of two reasons: a) it is an extensively and explicitly discussed model in the literature, especially the multi-variate case (with dependent features) as presented in Bishop [11], Chapter 10 and b) our main interest is in providing tools for non-normally distributed data such as DNA methylation values derived from different DNA profiling techniques. Nonetheless, Gaussian mixtures may be useful in cases of beta-intensities data that are influenced by covariates (to be discussed in Section 3.2.2) and therefore, we also present our code version of the variational Gaussian mixture model with independent features in Appendix B.

3.2 Mixture Models for Continuous Random Variables

3.2.1 Variational Dirichlet Process Beta Mixture

The Dirichlet Process Beta mixture model is ideal for determining the hidden clusters of data with bounded support range (Ma and Leijon [80], Lai et al. [71]), *e.g.* proportions.

For example, data drawn from array-based DNA profiling platforms, such as 450K and EPIC, concern proportion of DNA methylation at each CpG. This type of data, called beta-intensities (Chapter 1, Subsection 1.5.2), is confined to live in the $[0, 1]$ interval and therefore, Dirichlet Process Beta mixture would have been the first thought of a suitable model-based clustering method. However, this model assumes independence between the feature dimensions, an assumption we cannot make for the beta-intensities on the CpG, because adjacent CpGs show signs of correlated methylation levels (Eckhardt et al. [37], Maksimovic et al. [83]). Consequently, we suggest implementing the Dirichlet Process Beta mixture model not directly on methylation levels of individual CpGs, but on aggregated methylation rates (median beta-intensity) of differentially methylated regions (DMRs). In this way, the correlation is absorbed within the differentially methylated region after the aggregation of the correlated beta-intensities, resulting in relaxation of the association between the DMRs.

In this section, we start by displaying the hierarchical model of the Dirichlet Process Beta mixture while we carry on with the implementation of the variational inference procedure which concerns the derivation of all the variational distributions.

With regards to the hierarchical structure, the likelihood and the priors are

$$\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z} \sim \prod_{n=1}^{N} \prod_{m=1}^{M} \prod_{d=1}^{D} \text{Beta}\left(y_{nd} \mid u_{dm}, v_{dm}\right)^{z_{nm}}, \tag{3.1}$$

$$\boldsymbol{v} \sim \prod_{m=1}^{M} \prod_{d=1}^{D} \text{Gamma}\left(v_{dm} \mid \mu_{0dm}, \eta_{0dm}\right), \tag{3.2}$$

$$\boldsymbol{u} \sim \prod_{m=1}^{M} \prod_{d=1}^{D} \text{Gamma}\left(u_{dm} \mid \alpha_{0dm}, \beta_{0dm}\right), \tag{3.3}$$

$$\boldsymbol{z} \mid \boldsymbol{w} \sim \prod_{n=1}^{N} \text{Categorical}\left(\boldsymbol{z}_n \mid \boldsymbol{w}\right), \tag{3.4}$$

$$\boldsymbol{w} \sim \prod_{m=1}^{M} \text{Beta}\left(w_m \mid 1, \phi_{0m}\right), \tag{3.5}$$

where N is the number of samples, M the initial number of components (M a high integer as defined in the stick-breaking point process, Chapter 2) and D the features dimension of the dataset. In equation (3.1), the random data $\boldsymbol{y}$ is an $(N \times D)$-dimensional matrix, with N corresponding to the number of independent samples and D to the number of independent features. The double independence leads to the factorization of the likelihood, and given the allocations $\boldsymbol{z}$ the components are independent too. Regarding the vector of observations $\boldsymbol{y}$, this is distributed as a mixture of M independent Beta densities with component specific shape parameters (u_{dm} and v_{dm}) that vary across the features dimension D due to the assumption of independence between the features, with subscript m indicating the component index and d the specific feature dimension. The parameters $\boldsymbol{u}$, $\boldsymbol{v}$ of the Dirichlet Process Beta mixture are $D \times M$ matrices, with

each element u_{dm} and v_{dm} following *a priori* a Gamma density with hyperparameters $(\alpha_{0dm}, \beta_{0dm})$ in equation (3.3) and (μ_{0dm}, η_{0dm}) in (3.2). The latent allocation variable $\boldsymbol{z}$ is an $N \times M$ matrix, with each row denoting an M vector for the n^{th} sample, where its $M-1$ values are 0 and the one corresponding to the cluster of the $\boldsymbol{y}_n$ sample is equal to 1. Thus, $\boldsymbol{z}$ is distributed as a Categorical distribution in equation (3.4), with parameter $\boldsymbol{w}$. In respect of $\boldsymbol{w}$, this is the stick-breaking point vector that consists of M elements each one following a Beta density with shape hyperparameter ϕ_{0m} ($\boldsymbol{\phi}_0 = [\phi_{01}, ..., \phi_{0M}]$) and is related to the mixing weights of the model - $\boldsymbol{\pi}^T = [\pi_1, ..., \pi_M]$ - *via* the following equation: $\pi_m = w_m \prod_{j=1}^{m-1} (1 - w_j)$ (Chapter 2, Subsection 2.6.3).

A helpful additional way to understand the conditional dependencies of the Dirichlet Process Beta mixture parameters is the Directed Acyclic Graph in Figure 3.1. The white nodes correspond to the latent allocation z_{nm}, the stick-breaking point w_m and the component specific parameters (u_{dm}, v_{dm}), with the subscript m signifying the m^{th} component, n the n^{th} sample and d the d^{th} feature dimension. Each parameter node is located within a box that bears the parameter's dimensions, *i.e.*, z_{nm} is the (n, m) element of the $N \times M$ $\boldsymbol{z}$ matrix. The light grey node corresponds to the observation y_{nd}. As for the directed arrows, these show the conditional independencies/dependencies between the variables. In particular, the random data variable $\boldsymbol{y}$ (the $N \times D$ data matrix with elements y_{nd}) depends on the Beta component specific parameters $\boldsymbol{u}$ and $\boldsymbol{v}$, as well as on the latent variable $\boldsymbol{z}$ (arrow edges point on $\boldsymbol{y}$). On the other hand, $\boldsymbol{y}$ is conditionally independent of the stick breaking point variable $\boldsymbol{w}$ given the latent variable $\boldsymbol{z}$.

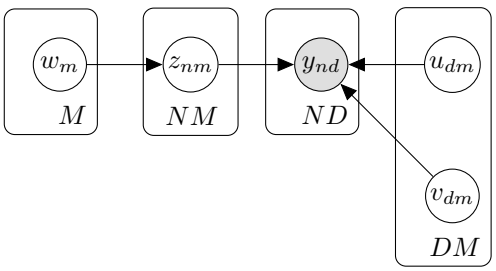


Figure 3.1 Directed Acyclic Graph of the Dirichlet Process Beta mixture model. The nodes represent the random variables, the directed edges the conditional dependence and the boxes the dimensionality of each parameter. The light grey node corresponds to the variable datapoint y_{nd}.

Mean Field approximation

After the introduction of the hierarchical Dirichlet Process Beta mixture, we proceed to the mathematical derivation of the Mean Field approximation. The aim is to produce closed form equations for the variational posterior parameters of $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}$ and $\boldsymbol{w}$. The decomposition of the joint approximated posterior density that results in tractable solutions is

$$q(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}) = q_z(\boldsymbol{z})\, q_w(\boldsymbol{w})\, q_u(\boldsymbol{u})\, q_u(\boldsymbol{v}), \tag{3.6}$$

where each variable $\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}$ bears its own variational distribution. The distribution indices are omitted for simplicity reasons from this point onward, *i.e.*, $q_z(\boldsymbol{z})$ will be reported as $q(\boldsymbol{z})$.

Variational distribution of $\boldsymbol{z}$

We first consider the derivation of the $q(\boldsymbol{z})$ distribution. Based on the Mean Field approximation presented in Chapter 2 and specifically equation (2.13), the optimal approximated distribution, in its logarithmic form, is given by the expected log-full conditional of $\boldsymbol{z}$ (equation (3.7)). The expectation is with respect to $\boldsymbol{w}, \boldsymbol{u}$ and $\boldsymbol{v}$, where each parameter follows its variational posterior, *i.e.*, $\boldsymbol{w} \sim q(\boldsymbol{w})$, $\boldsymbol{u} \sim q(\boldsymbol{u})$ and $\boldsymbol{v} \sim q(\boldsymbol{v})$. Therefore, the optimal solution for $q(\boldsymbol{z})$ depends on moments evaluated with respect to the variational distributions of the rest variables. The same holds for the optimal $q(\boldsymbol{w}), q(\boldsymbol{u})$ and $q(\boldsymbol{v})$ solutions, leading to the conclusion that the variational update equations are interlinked and must be solved iteratively (Bishop [11]).

$$\log q(\boldsymbol{z}) \propto \mathbb{E}_{\boldsymbol{w},\boldsymbol{u},\boldsymbol{v}}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v})\Big]. \tag{3.7}$$

The proportionality in equation (3.7) concerns $\boldsymbol{z}$, thus any terms that do not depend on $\boldsymbol{z}$ can be absorbed into the normalizing constant. Also, by making use of the dependencies in the Directed Acyclic Graph in Figure 3.1 ($\boldsymbol{z}$ dependent on $\boldsymbol{w}$, and $\boldsymbol{y}$ dependent on $\boldsymbol{z}$ as well as on $\boldsymbol{u}, \boldsymbol{v}$), equation (3.7) decomposes into

$$\log q(\boldsymbol{z}) \propto \mathbb{E}_{\boldsymbol{w}}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{w})\Big] + \mathbb{E}_{\boldsymbol{u},\boldsymbol{v}}\Big[\log P(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{v})\Big]. \tag{3.8}$$

The next step is to calculate the expected values in equation (3.8). The first expectation with respect to $\boldsymbol{w}$, shown in equation (3.9), refers to the log-prior of $\boldsymbol{z}$, which is a Categorical distribution with parameter vector $\boldsymbol{w}$ (equation (3.4)). This expectation is equal to

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w}}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{w})\Big] &= \mathbb{E}_{\boldsymbol{w}}\left[\log \prod_{n=1}^{N}\prod_{m=1}^{M}\left\{w_m \prod_{j=1}^{m-1}(1-w_j)\right\}^{z_{nm}}\right] \\
&= \mathbb{E}_{\boldsymbol{w}}\left[\sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\log w_m + \sum_{j=1}^{m-1}\log(1-w_j)\right\}\right] \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\mathbb{E}_{w_m}[\log w_m] + \sum_{j=1}^{m-1}\mathbb{E}_{w_m}[\log(1-w_j)]\right\}.
\end{aligned} \tag{3.9}$$

The second expectation with respect to $\boldsymbol{u}, \boldsymbol{v}$ in equation (3.8) corresponds to the logarithmic likelihood, which is equal to

$$\mathbb{E}_{\boldsymbol{u},\boldsymbol{v}}\Big[\log P(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{v})\Big] = \mathbb{E}_{\boldsymbol{u},\boldsymbol{v}}\left[\log \prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{d=1}^{D}\left\{\frac{\Gamma(u_{dm}+v_{dm})}{\Gamma(u_{dm})\Gamma(v_{dm})} y_{nd}^{u_{dm}-1}(1-y_{nd})^{v_{dm}-1}\right\}^{z_{nm}}\right]$$

$$
\begin{aligned}
=&\mathbb{E}_{\boldsymbol{u},\boldsymbol{v}}\Bigg[\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{d=1}^{D} z_{nm}\Bigg\{\log\frac{\Gamma(u_{dm}+v_{dm})}{\Gamma(u_{dm})\Gamma(v_{dm})} \\
&+(u_{dm}-1)\log y_{nd}+(v_{dm}-1)\log(1-y_{nd})\Bigg\}\Bigg] \\
=&\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{d=1}^{D} z_{nm}\Bigg\{\mathbb{E}_{u_{dm},v_{dm}}\left[\log\frac{\Gamma(u_{dm}+v_{dm})}{\Gamma(u_{dm})\Gamma(v_{dm})}\right] \\
&+(\mathbb{E}_{u_{dm}}[u_{dm}]-1)\log y_{nd}+(\mathbb{E}_{v_{dm}}[v_{dm}]-1)\log(1-y_{nd})\Bigg]\Bigg\}.
\end{aligned}
\tag{3.10}
$$

Subsequently, we can derive the unormalized optimal $\log q(\boldsymbol{z})$ form by substituting the expectations (3.9) and (3.10) on the right-hand side of equation (3.8). Any terms independent of $\boldsymbol{z}$ are discarded, since we initially care to find the posterior without the normalizing constant, leaving the definition with the normalizing constant at the end. The resulted log-$q(\boldsymbol{z})$ is proportional to

$$
\log q(\boldsymbol{z}) \propto \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log\rho_{nm}, \tag{3.11}
$$

where we define

$$
\begin{aligned}
\log\rho_{nm}=\sum_{d=1}^{D}\Bigg\{&\mathbb{E}_{u_{dm},v_{dm}}\left[\log\frac{\Gamma(u_{dm}+v_{dm})}{\Gamma(u_{dm})\Gamma(v_{dm})}\right]+(\mathbb{E}_{u_{dm}}[u_{dm}]-1)\log y_{nd} \\
&+(\mathbb{E}_{v_{dm}}[v_{dm}]-1)\log(1-y_{nd})\Bigg\}+\mathbb{E}_{w_m}[\log w_m]+\sum_{j=1}^{m-1}\mathbb{E}_{w_m}[\log(1-w_j)].
\end{aligned}
\tag{3.12}
$$

In order to de-logarithmize the unormalized $\log q(\boldsymbol{z})$, we have to apply the exponential function on both sides of (3.11). Thus,

$$
q(\boldsymbol{z}) \propto \prod_{n=1}^{N}\prod_{m=1}^{M}\rho_{nm}^{z_{nm}}. \tag{3.13}
$$

This is the point where the normalization of $q(\boldsymbol{z})$ in equation (3.13) is required. Considering the definition of z_{nm} (distributed as a Categorical), where for each n value the quantities z_{nm} are binary and sum to 1 over all m values, we obtain

$$
q(\boldsymbol{z}) = \prod_{n=1}^{N}\prod_{m=1}^{M} r_{nm}^{z_{nm}}, \tag{3.14}
$$

with

$$
r_{nm} = \rho_{nm}\Big/\sum_{j=1}^{M}\rho_{nj}. \tag{3.15}
$$

We observe in equation (3.14) that the functional form of $q(\boldsymbol{z})$ is a product of N Categorical distributions, same as the prior $P(\boldsymbol{z}\mid\boldsymbol{w})$, with variational expected value for the n^{th} observation and m^{th} component equal to

$$
\mathbb{E}_{z_{nm}}[z_{nm}] = r_{nm}. \tag{3.16}
$$

The expected value of the latent allocation z_{nm} in equation (3.16) introduces the responsibilities, variables that we referred to in Chapter 2 as one of the mixture models advantages. The responsibility r_{nm} reveals what is the probability the n^{th} sample to belong in cluster m (soft clustering), in contrast to K-means and Hierarchical clustering which allocate with probability 1 the n^{th} observation into a component (hard clustering).

Variational density of $\boldsymbol{w}$

The following mathematical derivation concerns the Mean Field density of $\boldsymbol{w}$, $q(\boldsymbol{w})$. Based again on the general equation for the optimal variational form in equation (2.13), Chapter 2, we have that $\log q(\boldsymbol{w})$ is proportional to the expected log-full conditional of $\boldsymbol{w}$ with respect to $\boldsymbol{z}, \boldsymbol{u}$ and $\boldsymbol{v}$. However, given Figure 3.1, $\boldsymbol{w}$ solely depends on $\boldsymbol{z}$ and therefore, the expectation simplifies to be with respect to $\boldsymbol{z}$. A more manageable form is given in equation (3.17) below

$$\begin{aligned}
\log q(\boldsymbol{w}) &\propto \mathbb{E}_{\boldsymbol{z},\boldsymbol{u},\boldsymbol{v}}\Big[\log P(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v})\Big] \\
&\propto \mathbb{E}_{\boldsymbol{z}}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{w}) + \log P(\boldsymbol{w} \mid \boldsymbol{\phi_0})\Big] \\
&= \mathbb{E}_{\boldsymbol{z}}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{w})\Big] + \log P(\boldsymbol{w} \mid \boldsymbol{\phi_0}).
\end{aligned} \tag{3.17}$$

The subsequent step is to calculate the expectation in equation (3.17), which includes the log-prior of $\boldsymbol{z}$. We then replace the log-prior of $\boldsymbol{w}$, to finally give rise to the unormalized $\log q(\boldsymbol{w})$ in equation (3.18)

$$\begin{aligned}
\log q(\boldsymbol{w}) \propto\ & \mathbb{E}_{\boldsymbol{z}}\left[\sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\log w_m + \sum_{j=1}^{m-1}\log(1-w_j)\right\}\right] + \sum_{m=1}^{M}\Big[(\phi_{0m}-1)\log(1-w_m)\Big] \\
&= \sum_{m=1}^{M}\left[\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{z}_{nm}}[z_{nm}]\left\{\log w_m + \sum_{j=1}^{m-1}\log(1-w_j) + (\phi_{0m}-1)\log(1-w_m)\right\}\right] \\
&= \sum_{m=1}^{M}\left[\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{z}_{nm}}[z_{nm}]\log w_m + \left\{\sum_{n=1}^{N}\sum_{j=m+1}^{M}\mathbb{E}_{\boldsymbol{z}_{nj}}[z_{nj}] + \phi_{0m} - 1\right\}\log(1-w_m)\right].
\end{aligned} \tag{3.18}$$

To annihilate the logarithm in equation (3.18) we apply the exponential on both sides to obtain

$$q(\boldsymbol{w}) \propto \prod_{m=1}^{M} w_m^{\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{z}_{nm}}[z_{nm}]} \times (1-w_m)^{\sum_{n=1}^{N}\sum_{j=m+1}^{M}\mathbb{E}_{\boldsymbol{z}_{nj}}[z_{nj}]+\phi_{0m}-1}. \tag{3.19}$$

In equation (3.19), $q(\boldsymbol{w})$ is proportional to a product of M Beta densities with

$$\begin{aligned}
\delta_m &= \sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{z}_{nm}}[z_{nm}] + 1, \\
\phi_m &= \sum_{n=1}^{N}\sum_{j=m+1}^{M}\mathbb{E}_{\boldsymbol{z}_{nj}}[z_{nj}] + \phi_{0m}
\end{aligned} \tag{3.20}$$

the defined variational parameters. $\mathbb{E}_{\boldsymbol{z}_{nm}}[z_{nm}]$ is calculated in equation (3.16) and corresponds to the responsibilities.

Variational density of $\boldsymbol{u}$ and $\boldsymbol{v}$

Having defined the closed form equations for the variational parameters of $q(\boldsymbol{z})$ and $q(\boldsymbol{w})$ in equations (3.15) and (3.20) respectively, it remains to derive the variational density $q(\boldsymbol{u})$ and $q(\boldsymbol{v})$, where both are Beta shape parameters of the Dirichlet Process Beta mixture.

The optimal variational density for $\boldsymbol{u}$ is proportional to the expected log-full conditional of $\boldsymbol{u}$, where the expectation is with respect to $\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{w}$. However, due to the conditional dependencies in Figure 3.1, $\boldsymbol{u}$ is associated only with $\boldsymbol{v}$ and $\boldsymbol{z}$ as co-parents of the data node $\boldsymbol{y}$. Similarly, $\boldsymbol{v}$ is connected with $\boldsymbol{z}$ and $\boldsymbol{u}$ through $\boldsymbol{y}$. Therefore, the unormalized variational densities of $\boldsymbol{u}$ and $\boldsymbol{v}$ are given in equation (3.21) and (3.22)

$$q(\boldsymbol{u}) \propto \exp\left\{\mathbb{E}_{\boldsymbol{v},\boldsymbol{z}}\left[\log P(\boldsymbol{u} \mid \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) + \log P(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z})\right]\right\}, \tag{3.21}$$

$$q(\boldsymbol{v}) \propto \exp\left\{\mathbb{E}_{\boldsymbol{u},\boldsymbol{z}}\left[\log P(\boldsymbol{v} \mid \boldsymbol{\mu}_0, \boldsymbol{\eta}_0) + \log P(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z})\right]\right\}. \tag{3.22}$$

Following that, we expand the variational density of $\boldsymbol{u}$ by replacing the log-prior of $\boldsymbol{u}$ and the log-likelihood. Similar procedure applies to $q(\boldsymbol{v})$ and therefore, we omit its derivation details to straightforwardly present the result later on.

The $\log q(\boldsymbol{u})$ is then obtained in equation (3.23), deprived however from its normalizing constant

$$\begin{aligned}\log q(\boldsymbol{u}) \propto &\sum_{m=1}^{M}\sum_{d=1}^{D}\left[(\alpha_{0dm}-1)\log u_{dm} - \beta_{0dm}u_{dm}\right] \\ &+ \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}_{z_{nm}}[z_{nm}]\sum_{d=1}^{D}\left\{\mathbb{E}_{v_{dm}}\left[\log\frac{\Gamma(u_{dm}+v_{dm})}{\Gamma(u_{dm})+\Gamma(v_{dm})}\right] + (u_{dm}-1)\log y_{nd}\right\}.\end{aligned} \tag{3.23}$$

At this point, we observe that equation (3.23) does not remind any of the known kernels due to the non tractable expectation term with respect to v_{dm}. Lai et al. [71] prove that this quantity can be alternatively approximated by a first order Taylor polynomial. In particular, $\mathbb{E}_{v_{dm}}\big[\log[\Gamma(u_{dm}+v_{dm})/\Gamma(u_{dm})+\Gamma(v_{dm})]\big]$, called for simplicity $Q(u_{dm})$, has a lower bound $\tilde{Q}(u_{dm})$ defined as in equation (3.25)

$$Q(u_{dm}) = \mathbb{E}_{v_{dm}}\left[\log\frac{\Gamma(u_{dm}+v_{dm})}{\Gamma(u_{dm})+\Gamma(v_{dm})}\right] \geq \tilde{Q}(u_{dm}), \tag{3.24}$$

with

$$\tilde{Q}(u_{dm}) = \log u_{dm}\left[\Psi(\mathbb{E}_{u_{dm}}[u_{dm}] + \mathbb{E}_{v_{dm}}[v_{dm}]) - \Psi(\mathbb{E}_{u_{dm}}[u_{dm}])\right]\mathbb{E}_{u_{dm}}[u_{dm}]. \tag{3.25}$$

This $\tilde{Q}(u_{dm})$ approximation is then replaced into the expected term with respect to u_{dm} in equation (3.23) and the approximated unormalized logarithmic $q(\boldsymbol{u})$ is

$$\log q(\boldsymbol{u}) \propto \sum_{m=1}^{M}\sum_{d=1}^{D}\Bigg[\Bigg\{\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\Bigg[\Psi(\mathbb{E}_{u_{dm}}[u_{dm}]+\mathbb{E}_{v_{dm}}[v_{dm}])-\Psi(\mathbb{E}_{u_{dm}}[u_{dm}])\Bigg] \times \mathbb{E}_{u_{dm}}[u_{dm}]+a_{dm}-1\Bigg\}\log u_{dm}-\Bigg\{\beta_{dm}-\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\log y_{nd}\Bigg\}u_{dm}\Bigg]. \tag{3.26}$$

The log-kernel of $q(\boldsymbol{u})$ now resembles a product of $M \times D$ Gamma densities with variational parameters for each Gamma defined as

$$\begin{aligned}\alpha_{dm} &= \alpha_{0dm}+\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\Bigg[\Psi\big(\mathbb{E}_{u_{dm}}[u_{dm}]+\mathbb{E}_{v_{dm}}[v_{dm}]\big)-\Psi\big(\mathbb{E}_{u_{dm}}[u_{dm}]\big)\Bigg]\mathbb{E}_{u_{dm}}[u_{dm}],\\ \beta_{dm} &= \beta_{0dm}-\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\log y_{nd}.\end{aligned} \tag{3.27}$$

A similar result is obtained for $q(\boldsymbol{v})$, where $\boldsymbol{v}$ is approximated by a product of Gamma densities too, with Gamma specific variational parameters

$$\begin{aligned}\mu_{dm} &= \mu_{0dm}+\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\Bigg[\Psi\big(\mathbb{E}_{u_{dm}}[u_{dm}]+\mathbb{E}_{v_{dm}}[v_{dm}]\big)-\Psi\big(\mathbb{E}_{v_{dm}}[v_{dm}]\big)\Bigg]\mathbb{E}_{v_{dm}}[v_{dm}],\\ \eta_{dm} &= \eta_{0dm}-\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\log(1-y_{nd}).\end{aligned} \tag{3.28}$$

Variational Expectations

Thus far, we have derived all the variational parameters of the Dirichlet Process Beta mixture model in equations (3.15), (3.20), (3.27) and (3.28). The next stage is to calculate the expectations found within these variational equations. Those mean quantities are with respect to the corresponding variational densities, as we discussed in the beginning of the variational derivations section, and are easily attained in (3.29).

For example, to compute $\mathbb{E}_{u_{dm}}[u_{dm}]$ we exploit the fact that

$$u_{dm} \sim q(u_{dm}) = \text{Gamma}(u_{dm} \mid \alpha_{dm}, \beta_{dm}).$$

The expected value of u_{dm} is then equal to the integral $\int_{u_{dm}} u_{dm} q(u_{dm}) du_{dm}$, which results in the fraction of the shape variational parameters $\mathbb{E}_{u_{dm}}[u_{dm}] = \alpha_{dm}/\beta_{dm}$. The rest expectations can be calculated likewise.

$$\begin{aligned}\mathbb{E}_{z_{nm}}[z_{nm}] &= r_{nm},\\ \mathbb{E}_{u_{dm}}[u_{dm}] &= \frac{\alpha_{dm}}{\beta_{dm}},\\ \mathbb{E}_{v_{dm}}[v_{dm}] &= \frac{\mu_{dm}}{\eta_{dm}},\\ \mathbb{E}_{u_{dm}}[\log u_{dm}] &= \Psi(\alpha_{dm})-\log\beta_{dm},\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{v_{dm}}\left[\log v_{dm}\right] &= \Psi(\mu_{dm}) - \log \eta_{dm}, \\
\mathbb{E}_{w_m}\left[\log w_m\right] &= \Psi(\delta_m) - \Psi(\delta_m + \phi_m), \\
\mathbb{E}_{w_m}\left[\log(1 - w_m)\right] &= \Psi(\phi_m) - \Psi(\delta_m + \phi_m), \\
\mathbb{E}_{w_m}\left[w_m\right] &= \frac{\delta_m}{\phi_m + \delta_m}.
\end{aligned} \tag{3.29}$$

As for $\mathbb{E}_{u_{dm},v_{dm}}\big[\log[\Gamma(u_{dm} + v_{dm})/\Gamma(u_{dm}) + \Gamma(v_{dm})]\big]$ in equation (3.12), this is an intractable integral with respect to (u_{dm}, v_{dm}) and hence, a closed form solution is not provided. To deal with this obstacle, Lai et al. [71] deploy again a first order Taylor polynomial, denoted as $\tilde{R}_{dm}$ in equation (3.31), to approximate the term. $\tilde{R}_{dm}$ specifically works as a lower bound for the expectation, as indicated in the inequality (3.30)

$$\mathbb{E}_{u_{dm},v_{dm}}\left[\log \frac{\Gamma(u_{dm} + v_{dm})}{\Gamma(u_{dm})\Gamma(v_{dm})}\right] \geq \tilde{R}_{dm}, \tag{3.30}$$

where

$$\begin{aligned}
\tilde{R}_{dm} = {} & \frac{\Gamma(\mathbb{E}_{u_{dm}}\left[u_{dm}\right] + \mathbb{E}_{v_{dm}}\left[v_{dm}\right])}{\Gamma(\mathbb{E}_{u_{dm}}\left[u_{dm}\right])\Gamma(\mathbb{E}_{v_{dm}}\left[v_{dm}\right])} \\
& + \left[\Psi\big(\mathbb{E}_{u_{dm}}\left[u_{dm}\right] + \mathbb{E}_{v_{dm}}\left[v_{dm}\right]\big) - \Psi\big(\mathbb{E}_{u_{dm}}\left[u_{dm}\right]\big)\right]\mathbb{E}_{u_{dm}}\left[u_{dm}\right] \\
& \times \left[\mathbb{E}_{u_{dm}}\left[\log u_{dm}\right] - \log \mathbb{E}_{u_{dm}}\left[u_{dm}\right]\right] \\
& + \left[\Psi\big(\mathbb{E}_{v_{dm}}\left[u_{dm}\right] + \mathbb{E}_{v_{dm}}\left[v_{dm}\right]\big) - \Psi\big(\mathbb{E}_{v_{dm}}\left[v_{dm}\right]\big)\right]\mathbb{E}_{v_{dm}}\left[v_{dm}\right] \\
& \times \left[\mathbb{E}_{v_{dm}}\left[\log v_{dm}\right] - \log \mathbb{E}_{v_{dm}}\left[v_{dm}\right]\right].
\end{aligned} \tag{3.31}$$

Variational Lower Bound

All the necessary variational equations and expectations have now been derived and we can finally complete the Variational Bayes procedure with the calculation of the Evidence Lower Bound. The ELBO is the objective function that works as the criterion for the optimal variational parameters selection. Particularly, it is a function of the variational parameters that is calculated at each iteration based on the current input values. When the current ELBO value is negligibly different to the previous ($\epsilon = 10^{-6}$), the variational algorithm converges to the optimal variational estimates.

Regarding the general form of the Mean Field lower bound, this is given in equation (2.9), Chapter 2, where $\boldsymbol{\theta}$ now contains the Dirichlet Process Beta mixture parameters $\{\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}\}$. More precisely, the ELBO is equal to the subtraction of the expected log-joint distribution (including the observations $\boldsymbol{y}$) and the expected log-variational joint distribution, where both expectations are with respect to the variational parameters. These two quantities can be further expanded leading to equation (3.32).

$$\begin{aligned}\mathcal{L}(\boldsymbol{y};q) &= \mathbb{E}_{z,w,u,v}[\log P(\boldsymbol{y},\boldsymbol{z},\boldsymbol{w},\boldsymbol{u},\boldsymbol{v})] - \mathbb{E}_{z,w,u,v}[\log q(\boldsymbol{z},\boldsymbol{w},\boldsymbol{u},\boldsymbol{v})] \\ &= \mathbb{E}_{z,u,v}[\log P(\boldsymbol{y}\mid\boldsymbol{z},\boldsymbol{u},\boldsymbol{v})] + \mathbb{E}_{z,w}[\log P(\boldsymbol{z})] + \mathbb{E}_{w}[\log P(\boldsymbol{w})] \\ &\quad + \mathbb{E}_{u}[\log P(\boldsymbol{u})] + \mathbb{E}_{v}[\log P(\boldsymbol{v})] - \mathbb{E}_{z}[\log q(\boldsymbol{z})] - \mathbb{E}_{w}[\log q(\boldsymbol{w})] \\ &\quad - \mathbb{E}_{u}[\log q(\boldsymbol{u})] - \mathbb{E}_{v}[\log q(\boldsymbol{v})].\end{aligned} \tag{3.32}$$

Finally, the explicit ELBO form, after substituting the likelihood, the priors and the variational distributions of the Dirichlet Process Beta mixture model is

$$\begin{aligned}\mathcal{L}(\boldsymbol{y};q) &= \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}_{z_{nm}}[z_{nm}]\log\rho_{nm} - \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}_{z_{nm}}[z_{nm}]\log\mathbb{E}_{z_{nm}}[z_{nm}] \\ &\quad + \sum_{m=1}^{M}\Big[\log\phi_{0m} + (\phi_{0m}-1)\mathbb{E}_{w_m}[\log(1-w_m)]\Big] \\ &\quad - \sum_{m=1}^{M}\Big[\Gamma(\delta_m+\phi_m) - \Gamma(\delta_m) - \Gamma(\phi_m) + (\delta_m-1)\mathbb{E}_{w_m}[\log w_m] \\ &\quad + (\phi_m-1)\mathbb{E}_{w_m}[\log(1-w_m)]\Big] \\ &\quad + \sum_{m=1}^{M}\sum_{d=1}^{D}\Big[a_{0dm}\log\beta_{0dm} - \log\Gamma(a_{0dm}) - a_{dm}\log\beta_{dm} + \log\Gamma(a_{dm}) \\ &\quad + (a_{0dm}-a_{dm})\mathbb{E}_{u_{dm}}[\log u_{dm}] - (\beta_{0dm}-\beta_{dm})\mathbb{E}_{u_{dm}}[u_{dm}]\Big] \\ &\quad + \sum_{m=1}^{M}\sum_{d=1}^{D}\Big[\mu_{0dm}\log\eta_{0dm} - \log\Gamma(\mu_{0dm}) - \mu_{dm}\log\eta_{dm} + \log\Gamma(\mu_{dm}) \\ &\quad + (\mu_{0dm}-\mu_{dm})\mathbb{E}_{v_{dm}}[\log v_{dm}] - (\eta_{0dm}-\eta_{dm})\mathbb{E}_{v_{dm}}[v_{dm}]\Big].\end{aligned} \tag{3.33}$$

Algorithm 8 Updating Scheme of the Variational Dirichlet Process Beta Mixture

1: Initialize:

1. Choose the initial number of components M
2. Choose initial values for the variational parameters:
   $\delta_m, \phi_m, \alpha_{dm}, \mu_{dm}, \beta_{dm}$ and η_{dm}
3. Choose values for the hyperparameters:
   $\phi_{0m}, \alpha_{0dm}, \mu_{0dm}, \beta_{0dm}$ and η_{0dm}

2: Repeat:

1. Calculate the expected values in equations (3.29) and (3.31)
2. Update the variational parameters in equations (3.12), (3.15), (3.20), (3.27) and (3.28)

3: Stop:

$\mathcal{L}(\boldsymbol{y};q)^{\text{current}} - \mathcal{L}(\boldsymbol{y};q)^{\text{previous}} \leq \epsilon$, where $\epsilon = 10^{-6}$

4: Pre-Final step:

Calculate the posterior mixing weight π_m as $\pi_m = \mathbb{E}_{w_m}[w_m]\prod_{j=1}^{m-1}\big(1-\mathbb{E}_{w_j}[w_j]\big)$

5: Final step:

Discard components with almost zero weight

To summarize the updating scheme for the variational Dirichlet Process Beta mixture model, we supply Algorithm 8. The algorithm begins with step 1 where we initialize the number of components M, the variational parameters and the prior hyperparameters presented in equations (3.1)-(3.5). We then iteratively update the expectations in equations (3.29) and (3.31), and the variational parameters in (3.12), (3.15), (3.20), (3.27) and (3.28), until the algorithm converges in step 3 (difference between current ELBO value and previous value lower than $\epsilon = 10^{-6}$). The pre-final step concerns the calculation of the variational weights of the components, π_m, after substituting the variational expected value of the stick-breaking point parameter (step 4). At the final stage, we retain the components with non-zero mixing weight, featuring the ability of Dirichlet Process in determining the number of clusters in the Beta mixture model.

3.2.2 Bounded Data with Confounding Parameters

In bounded data cases where confounding factors exist, such as beta-intensities measured in individuals of different group age and sex, the danger of a distorted clustering outcome lurks. The Dirichlet Process Beta mixture model should then be avoided because it does not take into consideration the confounding effects.

One alternative approach, that we suggest, is fitting a Beta regression model (Ferrari and Cribari-Neto [45]) to each feature (in total D) *via* the `betareg` package of Zeileis et al. [154], with covariates the confounding factors (*i.e.*, feature ~ sex + age). Espinheira et al. [39] recommend using the standardized weighted residuals (coded as "sweighted2" in R), which are distributed as a standard Gaussian. As a result, instead of clustering the original beta-intensities we could exploit the "sweighted2" residuals and apply the Dirichlet Process Gaussian mixtures on them with dependent features (Bishop [11], Chapter 10) if features are individual CpG sites, or independent features (Appendix B, B.3.1) if we consider aggregated beta-intensities such as median CpG methylation in differentially methylated regions. This way, the influence from any confounding parameter is vanished and the credibility of the inferential application is ensured.

3.3 Mixture Models for Discrete Random Variables

3.3.1 Variational Finite Binomial Mixture

The Finite mixture of Binomial distributions is a hierarchical model with fixed number of components and a Beta prior imposed upon the probability parameter of each sub-population's Binomial distribution. It is a rational model choice when the aim is to cluster count data for which the number of trials/experiments is known and the trials are independent. For example, a suitable application involves data produced by

Bisulfite Sequencing DNA mapping techniques such as whole genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS). Both methods provide the number of methylated reads (counts) *per* tested CpG site, along with the site's read depth (trials) (Chapter 1, Section 1.5.1). Nonetheless, as it has already been discussed in the previous section, due to the model structure of independence between the features in all the studied models in this thesis, we recommend applying the Binomial mixture on non-adjacent CpG sites (less correlated methylated counts) or on differentially methylated regions (DMR) which aggregate the correlated methylation values, reducing this way the dependence between the DMR counts. On a different note, we stress that the Finite Binomial mixture can successfully apply to binary data of analogous structure, since the Bernoulli probability function is a sub-case of the Binomial distribution for number of trials equal to one.

The structure of the current analysis starts with the listing of the likelihood and priors and progress with the Mean Field derivation of the variational densities as well as the calculation of the Evidence Lower Bound. We also introduce the way "annealing" applies to this mixture model, which is a consistent procedure to all the models in this chapter and the reason we skip introducing it again in the rest of the hierarchical mixtures. Moreover, the variational expectations involved into the approximated densities are summarized at the end, in similar manner to the variational Dirichlet Process Beta mixture framework presented earlier.

In order to begin, we introduce the hierarchical model

$$\boldsymbol{y} \mid \boldsymbol{p}, \boldsymbol{z} \sim \prod_{n=1}^{N} \prod_{d=1}^{D} \prod_{m=1}^{M} [\text{Binomial}(y_{nd} \mid s_{nd}, p_{dm})]^{z_{nm}/T_i}, \tag{3.34}$$

$$\boldsymbol{p} \sim \prod_{d=1}^{D} \prod_{m=1}^{M} \text{Beta}(p_{dm} \mid a_{0dm}, b_{0dm}), \tag{3.35}$$

$$\boldsymbol{z} \mid \boldsymbol{\pi} \sim \prod_{n=1}^{N} [\text{Categorical}(\boldsymbol{z}_n \mid \boldsymbol{\pi})]^{1/T_i}, \tag{3.36}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\pi} \mid \boldsymbol{\phi}_0), \tag{3.37}$$

where $\boldsymbol{y}$ is the $N \times D$ data matrix, with N denoting the total number of samples ($n \in \{1, 2, ..., N\}$) and D the total number of features ($d \in \{1, 2, ..., D\}$). The samples (rows of $\boldsymbol{y}$) are independent as well as the features (columns of $\boldsymbol{y}$), thus the distribution of $\boldsymbol{y}$ in equation (3.34) is such that the observations from the same group are distributed as a Binomial distribution with component specific parameters (s_{nd}, p_{dm}) that vary across the feature dimensions. With regards to index M, this is the fixed number of components, with $m \in \{1, 2, ..., M\}$. This is different from the Dirichlet Process mixtures, where the parameter M is the number of components we initialize the model with in order to eventually do cluster determination (usually the estimated number of components is $<< M$). Hence, Finite and Dirichlet Process definitions of M should

not be confused. In regard to s_{nd}, this is the number of non-random trials for the n^{th} sample and d^{th} feature, while p_{dm} is the random Binomial probability parameter of the d^{th} feature in the m^{th} cluster. Specifically, p_{dm} is an element of the $\boldsymbol{p}$ matrix with $D \times M$ dimensions, where p_{dm} follows *a priori* a Beta density with hyperparameters (a_{0dm}, b_{0dm}) (equation (3.35)). The latent variable $\boldsymbol{z}$, as in the Dirichlet Process Beta mixture, is modelled by a product of N Categorical distributions. However, this time the parameter vector is $\boldsymbol{\pi}$ instead of $\boldsymbol{w}$ (equation (3.36)). This happens by reason of directly modelling the mixing weights $\boldsymbol{\pi}$, whereas in the Dirichlet Process mixture we have to pass through the stick-breaking point representation to eventually derive $\boldsymbol{\pi}$. In equation (3.37), a Dirichlet prior with concentration parameter $\boldsymbol{\phi}_0$ is imposed upon $\boldsymbol{\pi}$ due to the simplex nature of the variable ($\boldsymbol{\pi} = [\pi_1, ..., \pi_M]$ with $\sum_{m=1}^{M} \pi_m = 1$). Finally, the annealing is achieved by simply raising the full likelihood (likelihood and prior of $\boldsymbol{z}$) in the power of $1/T_i$, where T_i is the temperature constant at the i^{th} iteration of the Mean Field algorithm (Chapter 2, Section 2.4).

To show how annealing interferes in practice, we write explicitly the forms of the log-likelihood and the log-prior of $\boldsymbol{z}$ in equations (3.38) and (3.39).

$$\log P(\boldsymbol{y} \mid \boldsymbol{p}, \boldsymbol{z}) = \sum_{n=1}^{N} \sum_{d=1}^{D} \sum_{m=1}^{M} \frac{z_{nm}}{T_i} \left[\log \binom{s_{nd}}{y_{nd}} + y_{nd} \log p_{dm} + (s_{nd} - y_{nd}) \log(1 - p_{dm}) \right], \tag{3.38}$$

$$\log P(\boldsymbol{z} \mid \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{z_{nm}}{T_i} \log \pi_m, \tag{3.39}$$

where the constant temperature T_i simply multiplies both probability functions, acting as a concavity regulator of the non-concave full likelihood (product of equation (3.38) and (3.39)).

Regarding the conditional dependencies of the model parameters, these can be clearly illustrated in a Directed Acyclic Graph, as in the Dirichlet Process Beta mixture. In Figure 3.2, the data variable $\boldsymbol{y}$ depends on the Binomial parameter $\boldsymbol{p}$ and the latent variable $\boldsymbol{z}$, whilst it is conditionally independent of the mixing weights $\boldsymbol{\pi}$ given $\boldsymbol{z}$.

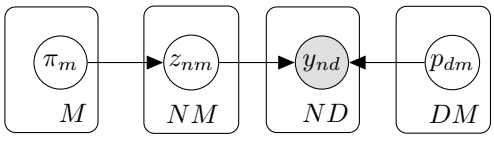


Figure 3.2 Directed Acyclic Graph of the Finite Binomial mixture model. The nodes represent the random variables, the directed edges the conditional dependence and the boxes the dimensionality of each parameter. The light grey node corresponds to the variable datapoint y_{nd}.

Mean Field approximation

In regards to the inference of the Finite Binomial mixture, we are able to proceed with

the Mean Field derivation of the parameter distributions having already defined the hierarchical model. The joint approximated posterior distribution takes a tractable form when it factorizes into $q(\boldsymbol{\pi})q(\boldsymbol{z})q(\boldsymbol{p})$.

Variational distribution of $\boldsymbol{z}$

The log-variational distribution of the latent allocation $\boldsymbol{z}$, $q(\boldsymbol{z})$, is proportional to the expected log-full conditional of $\boldsymbol{z}$ with respect to the remaining variables. However, based on the dependencies in Figure 3.2, the full-conditional simplifies to the product of likelihood and prior of $\boldsymbol{z}$ and hence, $\log q(\boldsymbol{z})$ is

$$\begin{aligned}
\log q(\boldsymbol{z}) &\propto \mathbb{E}_{/\boldsymbol{z}}\left[\log P(\boldsymbol{y} \mid \boldsymbol{p}, \boldsymbol{z}) + \log P(\boldsymbol{z} \mid \boldsymbol{\pi})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{z}}\Bigg[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M}\frac{z_{nm}}{T_i}\left\{y_{nd}\log p_{dm} + (s_{nd} - y_{nd})\log(1-p_{dm})\right\} \\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M}\frac{z_{nm}}{T_i}\log\pi_m\Bigg] \\
&= \sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M}\frac{z_{nm}}{T_i}\left\{y_{nd}\mathbb{E}_{p_{dm}}\left[\log p_{dm}\right] + (s_{nd} - y_{nd})\mathbb{E}_{p_{dm}}\left[\log(1-p_{dm})\right]\right\} \\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M}\frac{z_{nm}}{T_i}\mathbb{E}_{\pi_m}\left[\log\pi_m\right] \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M}\frac{z_{nm}}{T_i}\Bigg\{\sum_{d=1}^{D}y_{nd}\mathbb{E}_{p_{dm}}\left[\log p_{dm}\right] + \sum_{d=1}^{D}(s_{nd} - y_{nd})\mathbb{E}_{p_{dm}}\left[\log(1-p_{dm})\right] \\
&\quad + \mathbb{E}_{\pi_m}\left[\log\pi_m\right]\Bigg\}.
\end{aligned} \tag{3.40}$$

In equation (3.40), we present the unormalized $\log q(\boldsymbol{z})$. If we now define as $\log\rho_{nm}$ the expression inside the brackets on the right-hand side, we obtain

$$\log\rho_{nm} = \sum_{d=1}^{D}y_{nd}\mathbb{E}_{p_{dm}}\left[\log p_{dm}\right] + \sum_{d=1}^{D}(s_{nd} - y_{nd})\mathbb{E}_{p_{dm}}\left[\log(1-p_{dm})\right] + \mathbb{E}_{\pi_m}\left[\log\pi_m\right], \tag{3.41}$$

concluding with

$$\log q(\boldsymbol{z}) \propto \sum_{n=1}^{N}\sum_{m=1}^{M}\frac{z_{nm}}{T_i}\log\rho_{nm}. \tag{3.42}$$

In equation (3.42), after the definition of ρ_{nm}, the unormalized $q(\boldsymbol{z})$ reminds the kernel of a product of N Categorical distributions. In order for $q(\boldsymbol{z})$ to be a proper product of Categorical probability functions, the variational parameter of $\boldsymbol{z}$, also known as responsibility (here will be denoted as $\boldsymbol{r}$, like on the Dirichlet Process Beta mixture), has to be constrained into the $[0, 1]$ interval and have $\sum_{m=1}^{M} r_{nm} = 1$. Moreover, each responsibility r_{nm} should be raised in the power of $1/T_i$ due to the log property. Hence, $r_{nm} = \left(\rho_{nm}/\sum_{j=1}^{M}\rho_{nj}\right)^{1/T_i}$.

The normalized variational log-distribution of $\boldsymbol{z}$ is

$$\log q(\boldsymbol{z}) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm} \log r_{nm}, \tag{3.43}$$

revealing that $\boldsymbol{z}$ is approximated by a Categorical product with variational parameter $\boldsymbol{r}$ ($N \times M$ matrix)

$$\boldsymbol{z} \mid \boldsymbol{\pi} \sim \prod_{n=1}^{N} \text{Categorical}(\boldsymbol{z}_n \mid \boldsymbol{r}). \tag{3.44}$$

Variational distribution of $\boldsymbol{\pi}$

Regarding the approximated posterior distribution of the mixing weights, this is proportional to the expected log-full conditional of $\boldsymbol{\pi}$ with respect to the remaining parameters, which expands to the product of $\boldsymbol{z}$ prior and prior of $\boldsymbol{\pi}$ based on the dependencies in the Directed Acyclic Graph shown in Figure 3.2. Therefore, the unormalized $\log q(\boldsymbol{\pi})$ can be as

$$\begin{aligned}
\log q(\boldsymbol{\pi}) &\propto \mathbb{E}_{/\boldsymbol{\pi}}\left[\log P(\boldsymbol{z} \mid \boldsymbol{\pi}) + \log P(\boldsymbol{\pi})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{\pi}}\left[\sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm} \log \pi_m + \sum_{m=1}^{M} (\phi_{0m} - 1) \log \pi_m\right], \\
&= \sum_{m=1}^{M}\left(\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] + \phi_{0m} - 1\right) \log \pi_m.
\end{aligned} \tag{3.45}$$

In equation (3.45), $\log q(\boldsymbol{z})$ resembles the log-kernel of a Dirichlet distribution if we define as ϕ_m the expression inside the parenthesis

$$\phi_m = \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] + \phi_{0m}. \tag{3.46}$$

In particular, the variational distribution of $\boldsymbol{\pi}$ is a Dirichlet with parameter vector $\boldsymbol{\phi}$ (M-dimensional), where its elements are formalised in equation (3.46)

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\pi} \mid \boldsymbol{\phi}). \tag{3.47}$$

Variational distribution of $\boldsymbol{p}$

Thus far, we have retrieved the closed form variational parameters for the random variables $\boldsymbol{z}$ and $\boldsymbol{\pi}$. The last derivation refers to the Binomial probability parameter $\boldsymbol{p}$. Particularly, the Mean Field $\log q(\boldsymbol{p})$ is proportional to the expected annealed log-likelihood and log-prior of $\boldsymbol{p}$, where the expectation is with respect to $\boldsymbol{z}$. By substituting the necessary distributions, we attain the approximated posterior of $\boldsymbol{p}$ deprived from its normalizing constant

$$\begin{aligned}
\log q(\boldsymbol{p}) &\propto \mathbb{E}_{/\boldsymbol{p}}\left[\log P(\boldsymbol{y} \mid \boldsymbol{p}, \boldsymbol{z}) + \log P(\boldsymbol{p})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{p}}\Bigg[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} \frac{z_{nm}}{T_i} \left\{y_{nd} \log p_{dm} + (s_{nd} - y_{nd}) \log(1 - p_{dm})\right\} \\
&\quad + \sum_{d=1}^{D}\sum_{m=1}^{M}\Big\{(a_{0dm} - 1) \log p_{dm} + (b_{0dm} - 1) \log(1 - p_{dm})\Big\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{m=1}^{M}\sum_{d=1}^{D}\Bigg\{\sum_{n=1}^{N}\frac{\mathbb{E}_{z_{nm}}[z_{nm}]}{T_i}y_{nd}\log p_{dm} + \sum_{n=1}^{N}\frac{\mathbb{E}_{z_{nm}}[z_{nm}]}{T_i}(s_{nd}-y_{nd})\log(1-p_{dm}) \\
&\quad + (a_{0dm}-1)\log p_{dm} + (b_{0dm}-1)\log(1-p_{dm})\Bigg\} \\
&= \sum_{m=1}^{M}\sum_{d=1}^{D}\Bigg\{\Bigg(\sum_{n=1}^{N}\frac{\mathbb{E}_{z_{nm}}[z_{nm}]}{T_i}y_{nd} + a_{0dm} - 1\Bigg)\log p_{dm} \\
&\quad + \Bigg(\sum_{n=1}^{N}\frac{\mathbb{E}_{z_{nm}}[z_{nm}]}{T_i}(s_{nd}-y_{nd}) + b_{0dm} - 1\Bigg)\log(1-p_{dm})\Bigg\}.
\end{aligned}
\tag{3.48}
$$

On the right-hand side of equation (3.48), if we set the expressions in the two parentheses as

$$
\begin{aligned}
a_{dm} &= a_{0dm} + \sum_{n=1}^{N}\frac{\mathbb{E}_{z_{nm}}[z_{nm}]}{T_i}y_{nd}, \\
b_{dm} &= b_{0dm} + \sum_{n=1}^{N}\frac{\mathbb{E}_{z_{nm}}[z_{nm}]}{T_i}(s_{nd}-y_{nd}),
\end{aligned}
\tag{3.49}
$$

we obtain the kernel of a log-product of $D \times M$ Beta densities. Thus, we conclude that the normalized variational distribution of $\boldsymbol{p}$ is a Beta product with $\boldsymbol{a}$ and $\boldsymbol{b}$ parameters ($D \times M$ matrices), whose elements are defined in equation (3.49)

$$
\boldsymbol{p} \sim \prod_{d=1}^{D}\prod_{m=1}^{M}\text{Beta}(p_{dm} \mid a_{dm}, b_{dm}). \tag{3.50}
$$

Variational Expectations

The next step after the derivation of the Mean Field equations in (3.41), (3.46) and (3.49) is to calculate the variational expectations involved in them. We then provide in equation (3.53) the approximated posterior estimates of the mixing weights for the Finite Binomial mixture model. To give an instance of the calculations regarding the expectations, the expected value of $\log p_{dm}$ (p_{dm} is the Binomial probability parameter of the d^{th} feature in the m^{th} component) is found by solving the following tractable integral

$$
\mathbb{E}_{p_{dm}}[\log p_{dm}] = \int_{p_{dm}} \log p_{dm} q(p_{dm}) dp_{dm} = \Psi(a_{dm}) - \Psi(a_{dm} + b_{dm}), \tag{3.51}
$$

where $q(p_{dm}) \sim \text{Beta}(p_{dm} \mid a_{dm}, b_{dm})$ and $\Psi(\cdot)$ the digamma function. The remaining expectations are computed accordingly with respect to the corresponding variational distributions.

$$
\begin{aligned}
\mathbb{E}_{z_{nm}}[z_{nm}] &= r_{nm}, \\
\mathbb{E}_{\pi_m}[\log \pi_m] &= \Psi(\phi_m) - \Psi\left(\sum_{m=1}^{M}\phi_m\right), \\
\mathbb{E}_{p_{dm}}[\log p_{dm}] &= \Psi(a_{dm}) - \Psi(a_{dm} + b_{dm}), \\
\mathbb{E}_{p_{dm}}[(1-p_{dm})] &= \Psi(b_{dm}) - \Psi(a_{dm} + b_{dm})
\end{aligned}
\tag{3.52}
$$

and

$$\pi_m = \frac{\phi_{0m} + \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}]}{M\phi_{0m} + N}. \tag{3.53}$$

Variational Lower Bound

Having produced all the necessary closed form equations (variational parameters and variational expectations), we move forward to finding the explicit form of the Evidence Lower Bound. The Mean Field ELBO for the Finite Binomial mixture can be computed by subtracting the expected log-variational distributions from the expected log-joint distribution (including the data variable $\boldsymbol{y}$), where the expectations are with respect to the variational parameters (see equation (3.54) below).

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}; q) &= \mathbb{E}_{\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{p}}[\log P(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{p}, \boldsymbol{\pi})] - \mathbb{E}_{\boldsymbol{z}}[\log q(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{\pi})] - \mathbb{E}_{\boldsymbol{p}}[\log q(\boldsymbol{p})] \\ &= \mathbb{E}_{\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{p}}[\log P(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{p}) + \log P(\boldsymbol{z} \mid \boldsymbol{\pi}) + \log P(\boldsymbol{\pi}) + \log P(\boldsymbol{p})] \\ &\quad - \mathbb{E}_{\boldsymbol{z}}[\log q(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{\pi})] - \mathbb{E}_{\boldsymbol{p}}[\log q(\boldsymbol{p})].\end{aligned} \tag{3.54}$$

The ELBO is derived explicitly in equation (3.55), after replacing the Finite mixture likelihood, the priors and the variational distributions.

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}; q) = &\sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{E}_{z_{nm}}[z_{nm}] \log \rho_{nm} - \sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{E}_{z_{nm}}[z_{nm}] \log r_{nm} \\ &+ \log C(\boldsymbol{\phi}) - \log C(\boldsymbol{\phi_0}) + \sum_{m=1}^{M} (\phi_{0m} - \phi_m)\mathbb{E}_{\pi_m}[\log \pi_m] \\ &+ \sum_{d=1}^{D}\sum_{m=1}^{M} \Big\{ \log\Gamma(a_{0dm} + b_{0dm}) - \log\Gamma(a_{dm} + b_{dm}) - \log\Gamma(a_{0dm}) - \log\Gamma(b_{0dm}) \\ &+ \log\Gamma(a_{0dm}) + \log\Gamma(b_{0dm}) + (a_{0dm} - a_{dm})\mathbb{E}_{\boldsymbol{p}}[\log p_{dm}] \\ &+ (b_{0dm} - b_{dm})\mathbb{E}_{\boldsymbol{p}}[\log(1 - p_{dm})] \Big\},\end{aligned} \tag{3.55}$$

with $C(\cdot)$ being the inverse multivariate beta function that works as a normalizing constant for the Dirichlet distribution.

To summarize the Variational Bayes process for the Finite Binomial mixture, we provide Algorithm 9. We start by fixing the number of components, initializing the variational parameters and setting the hyperparameters according to our prior belief. We afterwards let the algorithm calculate the expectation terms in (3.52) in order to update the variational parameters in equations (3.41), (3.46) and (3.49). The Mean Field scheme stops when the ELBO value at the current iteration is negligibly different to the previous ($< 10^{-6}$). The final step concerns the estimation of the approximated posterior weights for the M components, a step that comes in contrast

to the Dirichlet Process mixture where the algorithm discards the components with almost zero estimated weight (Algorithm 8).

Algorithm 9 Updating Scheme of the Variational Finite Binomial Mixture

1: Initialize:

1. Pre-Fix the number of components M to some value (usually a high one, *i.e* 100 or lower when we have reason to believe the true number of components should be considerably less - this way the algorithm's convergence speed is increased)
2. Choose initial values for the variational parameters:
   ϕ_m, a_{dm} and b_{dm}
3. Choose values for the hyperparameters:
   ϕ_{0_m}, a_{0dm} and b_{0dm}

2: Repeat:

1. Calculate the expected values in equations (3.52)
2. Update the variational parameters in equations (3.41), (3.46) and (3.49)

3: Stop:

$\mathcal{L}(\boldsymbol{y}; q)^{\text{current}} - \mathcal{L}(\boldsymbol{y}; q)^{\text{previous}} \leq \epsilon$, where $\epsilon = 10^{-6}$

4: Final step:

Calculate the posterior mixing weight π_m as in equation (3.53)

3.3.2 Variational Finite Bernoulli Mixture with Covariates

The Bernoulli mixture is a special case of the Binomial mixture when the number of trials is one and therefore, an ideal model for grouping binary data. A possible application concerns DNA methylation data extracted from EPIC and 450K platform arrays (beta-intensities *per* CpG site). In particular, the beta-intensities are aggregated within known from the literature differentially methylated regions (DMRs) and subsequently transformed into binary (through simple data transformations[1]), with 0 corresponding to non-significantly methylated DMR and 1 to significantly.

In this section, we upgrade this model-based clustering of dichotomous variables to Finite Bernoulli mixture with covariates. This is a technique that takes into consideration the presence of confounding parameters, when we have reasons to believe that factors taint the outcome (samples of different sex, age group etc.). In a nutshell, we perform regression and clustering together which has not been attempted before according to

[1]Consider for simplicity a case/control experiment. One way to aggregate is to calculate the median of the methylation levels across CpGs for each sample and each DMR. Then, for each DMR, if the median methylation level of the sample is outside of the confidence healthy individuals interval, which is created by looking at the median methylation level across the healthy individuals, then the specific DMR takes value 1 otherwise it takes 0 value. In practice, 0 and 1 correspond to the outcome of a non-parametric test of the median methylation level for each case versus all the controls.

our knowledge. The knowledge we have is that for continuous random variables we can use the Beta regression residuals, which are normally distributed, to clear the confounding effects (Subsection 3.2.2). However, for discrete random variables, this is more difficult since there is not a unique definition of residuals and their distribution. Therefore, we need a model with both regression and clustering skills. For the Binomial version with covariates, we simply set the number of trials (> 1), while we multiply the binomial coefficient to the likelihood in equation (3.56).

The analysis begins with the strategy that eases the Bayesian inference (Pólya-Gamma augmentation, Polson et al. [113]) and proceeds with the exposition of the hierarchical model (likelihood and priors). The final part concerns the Mean Field derivation.

The likelihood of the Finite Bernoulli mixture without covariates is

$$\boldsymbol{y} \mid \boldsymbol{p}, \boldsymbol{z} \sim \prod_{n=1}^{N} \prod_{m=1}^{M} \prod_{d=1}^{D} \text{Bernouli}[(y_{nd} \mid 1, p_{ndm})]^{znm}, \tag{3.56}$$

where $\boldsymbol{y}$, similarly to the Finite Binomial mixture, is the $N \times D$ data matrix and $\boldsymbol{z}$ the $N \times M$ latent allocation. Regarding the Bernoulli probability parameter $\boldsymbol{p}$, this is an array of $N \times D \times M$ dimensions, allowing each sample to have also unique probability parameter. This construction facilitates the introduction of covariates in the model and we shortly explain the reason.

In the case where covariates exist, the expected value of the d^{th} feature for the n^{th} sample, which is equal to the probability parameter $p_{ndm} = \mathbb{E}(y_{nd})$, depends on the $\boldsymbol{x}_n$ covariates. $\boldsymbol{x}$ is an $N \times L$ matrix (L the number of covariates/confounding effects) and $\boldsymbol{\beta}$ is the $L \times D \times M$ covariate coefficients array with $\boldsymbol{\beta}_{dm}$ elements. This dependence on the covariates can be expressed through the logit function as

$$\text{logit}(p_{ndm}) = \psi_{ndm} \leftrightarrow p_{ndm} = \frac{\exp(\psi_{ndm})}{1 + \exp(\psi_{ndm})}, \text{ where } \psi_{ndm} = \boldsymbol{x}_n^T \boldsymbol{\beta}_{dm}. \tag{3.57}$$

Hence, the mean value p_{ndm} is now a function of the linear predictor ψ_{ndm} in equation (3.57). We point out that the new variable ψ_{ndm} varies across the samples due to the sample specific covariate values ($\boldsymbol{x}_n$). Therefore, the Bernoulli probability parameter $\boldsymbol{p}$, which is a function of $\boldsymbol{\psi}$ ($N \times D \times M$ matrix), also has to be sample specific by definition (equation (3.56)) when confounding parameters are considered.

The new likelihood with covariates is given in equation (3.58) below.

$$\begin{aligned} P(\boldsymbol{y} \mid \boldsymbol{p}, \boldsymbol{z}) &= \prod_{n=1}^{N} \prod_{m=1}^{M} \prod_{d=1}^{D} \left\{ \left[\frac{\exp(\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm})}{1 + \exp(\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm})} \right]^{y_{nd}} \left[1 - \frac{\exp(\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm})}{1 + \exp(\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm})} \right]^{y_{nd}-1} \right\}^{z_{nm}} \\ &= \prod_{n=1}^{N} \prod_{m=1}^{M} \prod_{d=1}^{D} \left\{ \frac{[\exp(\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm})]^{y_{nd}}}{[1 + \exp(\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm})]} \right\}^{z_{nm}}. \end{aligned} \tag{3.58}$$

However, the functional form of equation (3.58) does not permit Bayesian inference with respect to $\boldsymbol{\beta}_{dm}$ (Bishop [11]). Specifically, when multiplying the likelihood by the prior of $\boldsymbol{\beta}_{dm}$ (*i.e.*, a multivariate Gaussian) in order to determine the posterior of $\boldsymbol{\beta}_{dm}$ the resulted kernel does not remind of any known distribution and therefore, we are required to find the normalizing constant (computation of a complex integral).

To tackle this issue, Polson et al. [113] prove that quantities similar to the fraction in equation (3.58) are equivalent to a manageable equation that leads to a tractable Bayesian solution. The key is the introduction of a new variable whose role is to work as intermediary step to facilitate conjugate inference (closed form for the posterior of the regression coefficients $\boldsymbol{\beta}$). This augmented variable is defined as $\boldsymbol{\omega}$ and has the same dimensions as $\boldsymbol{y}$ ($N \times D$) (see equation (3.64) and its explanation below for details).

$$\frac{[\exp(\psi)]^a}{[1+\exp(\psi)]^b} = 2^{-b}\exp(\kappa\psi)\int_0^\infty \exp(-\omega\psi^2/2)P(\omega)d\omega, \tag{3.59}$$

where $\kappa = a - b/2$, ψ a random variable (which coincides to the linear predictor in our case) and $P(\omega) = PG(\omega \mid b, 0)$ the Pólya-Gamma prior distribution of ω (Polson et al. [113]). Moreover, they showed that if ω follows *a priori* a $PG(\omega \mid b, 0)$ then its posterior is $PG(\omega \mid b, \psi)$ with mean value

$$\mathbb{E}_\omega[\omega] = \frac{b}{2\psi}\tanh(\psi/2) = \frac{b}{2\psi}\left(\frac{\exp(\psi)-1}{1+\exp(\psi)}\right). \tag{3.60}$$

The likelihood of the Finite Bernoulli mixture model with covariates can now be written equivalently as in equation (3.64), after defining the following values

$$b = 1, \tag{3.61}$$

$$a_{nd} = y_{nd}, \tag{3.62}$$

$$\kappa_{nd} = y_{nd} - 1/2. \tag{3.63}$$

Then,

$$\begin{aligned} P(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{z}) = \prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{d=1}^{D}\Bigg\{ & \frac{1}{2}\exp\left(\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm}\kappa_{nd}\right) \\ & \times \int_0^\infty \exp(-\frac{1}{2}\omega_{nd}(\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm})^2)PG(\omega_{nd} \mid 1, 0)d\omega_{nd}\Bigg\}^{z_{nm}}, \end{aligned} \tag{3.64}$$

where z_{nm} the latent component allocation variable of the n^{th} sample into the m^{th} group ($z_{nm} = 0$ or 1). Regarding the augmented variable $\boldsymbol{\omega}$, this is an $N \times D$ matrix whose elements ω_{nd} are independent across rows and columns, as in $\boldsymbol{y}$. $\boldsymbol{\omega}$ marginalises out and now the likelihood obtains a tractable form for Bayesian inference on $\boldsymbol{\beta}$.

The complete likelihood of the Finite Bernoulli mixture with covariates (joint distribution of observed $\boldsymbol{y}$ and augmented $\boldsymbol{\omega}$) in equation (3.65) implies that each element

ω_{nd} of $\boldsymbol{\omega}$ is distributed *a priori* as a mixture of Pólya-Gamma densities. However, the Pólya-Gamma parameters for each M component are fixed to $(1, 0)$ and therefore, the mixture for the augmented variable simplifies to a single prior for ω_{nd} in equation (3.66).

$$\begin{aligned} P(\boldsymbol{y}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{z}) = & \prod_{n=1}^{N} \prod_{m=1}^{M} \prod_{d=1}^{D} \left\{ 2^{-1} \exp\left(\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm} \kappa_{nd} - \omega_{nd} [\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm}]^2 / 2 \right) \right\}^{z_{nm}} \\ & \times \prod_{n=1}^{N} \prod_{d=1}^{D} \underbrace{\prod_{m=1}^{M} \{ PG(\omega_{nd} \mid 1, 0) \}^{z_{nm}}}_{\text{Pólya-Gamma mixture}}, \end{aligned} \tag{3.65}$$

$$\begin{aligned} P(\boldsymbol{y}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{z}) = & \prod_{n=1}^{N} \prod_{m=1}^{M} \prod_{d=1}^{D} \left\{ 2^{-1} \exp\left(\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm} \kappa_{nd} - \omega_{nd} [\boldsymbol{x}_n^T \boldsymbol{\beta}_{dm}]^2 / 2 \right) \right\}^{z_{nm}} \\ & \times \prod_{n=1}^{N} \prod_{d=1}^{D} PG(\omega_{nd} \mid 1, 0). \end{aligned} \tag{3.66}$$

Having defined in equation (3.66) the final form of the joint distribution of $(\boldsymbol{y}, \boldsymbol{\omega})$, we proceed with imposing prior distributions on the model parameters $\boldsymbol{\beta}, \boldsymbol{z}$ and $\boldsymbol{\pi}$

$$\boldsymbol{\beta} \sim \prod_{d=1}^{D} \prod_{m=1}^{M} \mathcal{N}_L(\boldsymbol{\beta}_{dm} \mid \boldsymbol{\mu}_{0dm}, \boldsymbol{S}_{0dm}), \tag{3.67}$$

$$\boldsymbol{z} \mid \boldsymbol{\pi} \sim \prod_{n=1}^{N} [\text{Categorical}(\boldsymbol{z}_n \mid \boldsymbol{\pi})], \tag{3.68}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\pi} \mid \boldsymbol{\phi}_0), \tag{3.69}$$

where $\boldsymbol{\beta}$ is an $L \times D \times M$ regression coefficients array, independent across the feature dimensions D and the components M, whereas correlation is allowed between the covariates. Thus, $\boldsymbol{\beta}_{dm}$ is assumed to follow *a priori* an L-dimensional Gaussian with mean array $\boldsymbol{\mu}_{0dm}$ and covariance array $\boldsymbol{S}_{0dm}$. The latent variable $\boldsymbol{z}$ and the mixing weights $\boldsymbol{\pi}$ are distributed as a Categorical and a Dirichlet distribution respectively, similarly to the Finite Binomial mixture.

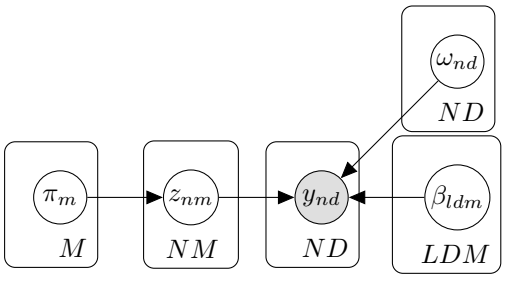


Figure 3.3 Directed Acyclic Graph of the Finite Bernoulli mixture model with covariates. The nodes represent the random variables, the directed edges the conditional dependence and the boxes the dimensionality of each parameter. The light grey node corresponds to the datapoint y_{nd}.

To facilitate the understanding of the connections between the random variables of the Finite Bernoulli mixture with covariates, we illustrate the Directed Acyclic Graph in Figure 3.3. The data variable $\boldsymbol{y}$ is dependent on the augmented variable $\boldsymbol{\omega}$, while $\boldsymbol{\omega}$ is

independent of the rest of the parameters. Based on the figure, we assist the Mean Field derivation of the posterior distributions. The joint approximated posterior is a factorization of the form $q(\boldsymbol{\pi})q(\boldsymbol{z})q(\boldsymbol{\beta})q(\boldsymbol{\omega})$.

Variational distribution of $\boldsymbol{z}$

In the context of the log-variational distribution of $\boldsymbol{z}$, this is proportional to the expected log-joint distribution of $(\boldsymbol{y}, \boldsymbol{\omega})$ and the log-prior of $\boldsymbol{z}$, given the parameter dependencies in Figure 3.3. The expectation is with respect to the remaining parameters.

$$\begin{aligned}
\log q(\boldsymbol{z}) &\propto \mathbb{E}_{/\boldsymbol{z}}\left[\log P(\boldsymbol{y}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{z}) + \log P(\boldsymbol{z} \mid \boldsymbol{\pi})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{z}}\left[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm}\kappa_{nd} - \omega_{nd}(\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm})^2/2\right\} + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log\pi_m\right] \\
&= \sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{\boldsymbol{x}_n^T\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}]\kappa_{nd} - \mathbb{E}_{\omega_{nd}}[\omega_{nd}]\mathbb{E}_{\beta_{dm}}[(\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm})^2]/2\right\} \\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\mathbb{E}_{\pi_m}\left[\log\pi_m\right] \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\Bigg\{\sum_{d=1}^{D}\boldsymbol{x}_n^T\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}]\kappa_{nd} - \sum_{d=1}^{D}\mathbb{E}_{\omega_{nd}}[\omega_{nd}]\mathbb{E}_{\beta_{dm}}[(\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm})^2]/2 \\
&\quad + \mathbb{E}_{\pi_m}\left[\log\pi_m\right]\Bigg\}.
\end{aligned} \tag{3.70}$$

To recognize the kernel of the $\log q(\boldsymbol{z})$ in equation (3.70), we follow the same procedure as in the previously discussed mixture models. We define the expression inside the brackets as $\log\rho_{nm}$, with the unormalized log-$q(\boldsymbol{z})$ reminding now a Categorical product kernel.

$$\log\rho_{nm} = \sum_{d=1}^{D}\boldsymbol{x}_n^T\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}]\kappa_{nd} - \sum_{d=1}^{D}\mathbb{E}_{\omega_{nd}}[\omega_{nd}]\mathbb{E}_{\beta_{dm}}[(\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm})^2]/2 + \mathbb{E}_{\pi_m}\left[\log\pi_m\right]. \tag{3.71}$$

However, in order to have a proper Categorical distribution, we impose the constraint $r_{nm} = \rho_{nm}/\sum_{j=1}^{M}\rho_{nj}$, where r_{nm} is the responsibility variable and therefore, the variational log-Categorical density is

$$\log q(\boldsymbol{z}) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log r_{nm}. \tag{3.72}$$

Variational distribution of $\boldsymbol{\pi}$

At next, we derive the variational mixing weights distribution. The derivation and result is identical to the variational Finite Binomial mixture,

$$\log q(\boldsymbol{\pi}) \propto \mathbb{E}_{/\boldsymbol{\pi}}\left[\log P(\boldsymbol{z} \mid \boldsymbol{\pi}) + \log P(\boldsymbol{\pi})\right]$$

$$\begin{aligned}
&\propto \mathbb{E}_{/\boldsymbol{\pi}}\left[\sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log\pi_m + \sum_{m=1}^{M}(\phi_{0m}-1)\log\pi_m\right] \\
&= \sum_{m=1}^{M}\left(\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}] + \phi_{0m} - 1\right)\log\pi_m,
\end{aligned} \tag{3.73}$$

where $\log q(\boldsymbol{\pi})$ in equation (3.73) is proportional to a product of Dirichlet distributions, with variational parameter vector $\boldsymbol{\phi}$ and ϕ_m elements obtained as

$$\phi_m = \sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}] + \phi_{0m}. \tag{3.74}$$

Variational distribution of $\boldsymbol{\beta}$

The next Mean Field approximated distribution is $q(\boldsymbol{\beta})$. Its unormalized form is proportional to the expected summation of the log-joint of $(\boldsymbol{y}, \boldsymbol{\omega})$ and the log-prior of $\boldsymbol{\beta}$. The expansion of this expectation with respect to the remaining variational parameters leads to a product of $D \times M$ L-dimensional independent Gaussian densities

$$\begin{aligned}
\log q(\boldsymbol{\beta}) &\propto \mathbb{E}_{/\boldsymbol{\beta}}\left[\log P(\boldsymbol{y}, \boldsymbol{\omega} \mid \boldsymbol{\beta}, \boldsymbol{z}) + \log P(\boldsymbol{\beta})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{\beta}}\Bigg[\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{d=1}^{D} z_{nm}\left\{\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm}\kappa_{nd} - \omega_{nd}[\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm}]^2/2\right\} \\
&\quad + \sum_{d=1}^{D}\sum_{m=1}^{M}\left\{-\frac{1}{2}[\boldsymbol{\beta}_{dm} - \boldsymbol{\mu}_{0dm}]^T\boldsymbol{S}_{0dm}^{-1}[\boldsymbol{\beta}_{dm} - \boldsymbol{\mu}_{0dm}]\right\}\Bigg] \\
&\propto \sum_{m=1}^{M}\sum_{d=1}^{D}\Bigg\{-\frac{1}{2}\boldsymbol{\beta}_{dm}^T\left(\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\mathbb{E}_{\omega_{nd}}[\omega_{nd}]\boldsymbol{x}_n^T\boldsymbol{x}_n + \boldsymbol{S}_{0dm}^{-1}\right)\boldsymbol{\beta}_{dm} \\
&\quad + \left(\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\kappa_{nd}\boldsymbol{x}_n + \boldsymbol{\mu}_{0dm}^T\boldsymbol{S}_{0dm}^{-1}\right)\boldsymbol{\beta}_{dm}\Bigg\},
\end{aligned} \tag{3.75}$$

with variational mean array and covariance given in the set of equations (3.76).

$$\begin{aligned}
\boldsymbol{S}_{dm} &= \left\{\boldsymbol{S}_{0dm}^{-1} + \boldsymbol{X}^T\mathbb{E}_{\boldsymbol{z}_m}[\boldsymbol{z}_m^T]\mathbb{E}_{\boldsymbol{\Omega}}[\boldsymbol{\Omega}]\boldsymbol{Y}\right\}^{-1}, \\
\boldsymbol{\mu}_{dm} &= \boldsymbol{S}_{dm}\left\{\boldsymbol{S}_{0dm}^{-1}\boldsymbol{\mu}_{0dm} + \boldsymbol{X}^T\mathbb{E}_{\boldsymbol{z}_m}[\boldsymbol{z}_m^T]\boldsymbol{\kappa}\right\},
\end{aligned} \tag{3.76}$$

with

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}_m}[\boldsymbol{z}_m] &= \mathrm{diag}(\mathbb{E}_{z_{1m}}[z_{1m}], ..., \mathbb{E}_{z_{Nm}}[z_{Nm}]), \\
\mathbb{E}_{\boldsymbol{\Omega}}[\boldsymbol{\Omega}] &= \mathrm{diag}(\mathbb{E}_{\omega_1}[\omega_1], ..., \mathbb{E}_{\omega_N}[\omega_N]).
\end{aligned} \tag{3.77}$$

Variational distribution of $\boldsymbol{\omega}$

As for the latent parameter $\boldsymbol{\omega}$ and based on Figure 3.3, the augmented variable is associated with the data variable $\boldsymbol{y}$ as a parent node. Hence, the log-variational density of $\boldsymbol{\omega}$ is proportional to the expected log-joint of $(\boldsymbol{y}, \boldsymbol{\omega})$ and the log-Pólya-Gamma prior of $\boldsymbol{\omega}$. The explicit form of the normalized $\log q(\boldsymbol{\omega})$ is

$$\log q(\boldsymbol{\omega}) = \sum_{n=1}^{N}\sum_{d=1}^{D}\Bigg\{-\frac{1}{2}\omega_n \sum_{m=1}^{M} \mathbb{E}_{z_{nm}}[z_{nm}]\mathbb{E}_{\beta_m}[(\boldsymbol{x}_n^T\boldsymbol{\beta}_m)^2] + \log \mathrm{PG}(\omega_n \mid 1, 0)\Bigg\}. \tag{3.78}$$

Thus, by applying the exponential function on the $\log q(\boldsymbol{\omega})$ in equation (3.78) we attain that $q(\omega_{nd})$ is a Pólya-Gamma

$$q(\omega) = \prod_{n=1}^{N}\prod_{d=1}^{D} \mathrm{PG}(\omega_{nd} \mid 1, c_{nd}), \tag{3.79}$$

with variational parameter

$$c_{nd} = \sqrt{\sum_{m=1}^{M} \mathbb{E}_{z_{nm}}[z_{nm}]\mathbb{E}_{\beta_{dm}}[(\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm})^2]}, \tag{3.80}$$

where

$$\mathbb{E}_{\beta_{dm}}[(\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm})^2] = (\boldsymbol{x}_n^T\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}])^2 + \boldsymbol{x}_n^T\mathrm{Cov}(\boldsymbol{\beta}_{dm})\boldsymbol{x}_n. \tag{3.81}$$

The variational mean value of ω_{nd}, given equation (3.60), is

$$\mathbb{E}_{\omega_{nd}}[\omega_{nd}] = \frac{1}{2c_{nd}}\tanh(c_{nd}/2). \tag{3.82}$$

Variational Expectations

After the definition of the closed form variational equations of the Finite Bernoulli mixture model with covariates, we calculate the variational expectations and posterior mixing weights by conditioning on the corresponding variational distributions.

$$\begin{aligned}
\mathbb{E}_{z_{nm}}[z_{nm}] &= r_{nm},\\
\mathbb{E}_{\pi_m}[\log \pi_m] &= \Psi(\phi_m) - \Psi\left(\sum_{m=1}^{M}\phi_m\right),\\
\mathbb{E}_{\beta_{dm}}[(\boldsymbol{x}_n^T\boldsymbol{\beta}_{dm})^2] &= (\boldsymbol{y}_n^T\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}])^2 + \boldsymbol{y}_n^T\mathrm{Cov}(\boldsymbol{\beta}_{dm})\boldsymbol{y}_n,\\
\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}] &= \boldsymbol{\mu}_{dm},\\
\mathbb{E}_{\omega_{nd}}[\omega_{nd}] &= \frac{1}{2c_{nd}}\tanh(c_{nd}/2),
\end{aligned} \tag{3.83}$$

and

$$\pi_m = \frac{\phi_{0m} + \sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]}{M\phi_{0m} + N}. \tag{3.84}$$

Variational Lower Bound

The final step concerns the Evidence Lower Bound calculation. The ELBO function for the Bernoulli mixture model with covariates can be found by subtracting the expected

log-variational distributions from the expected log-joint distribution of $(\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\pi})$.

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}; q) &= \mathbb{E}_{z,\pi,\beta,\omega}[\log P(\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\pi})] - \mathbb{E}_z[\log q(\boldsymbol{z})] - \mathbb{E}_\pi[\log q(\boldsymbol{\pi})] \\ &\quad - \mathbb{E}_\beta[\log q(\boldsymbol{\beta})] - \mathbb{E}_\omega[\log q(\boldsymbol{\omega})] \\ &= \mathbb{E}_{z,\pi,\beta,\omega}[\log P(\boldsymbol{y}, \boldsymbol{\omega} \mid \boldsymbol{z}, \boldsymbol{\beta}) + \log P(\boldsymbol{z} \mid \boldsymbol{\pi}) + \log P(\boldsymbol{\pi}) + \log P(\boldsymbol{p})] \\ &\quad - \mathbb{E}_z[\log q(\boldsymbol{z})] - \mathbb{E}_\pi[\log q(\boldsymbol{\pi})] - \mathbb{E}_\beta[\log q(\boldsymbol{\beta})] - \mathbb{E}_\omega[\log q(\boldsymbol{\omega})].\end{aligned} \tag{3.85}$$

The explicit form after the substitution of the distributions is

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}; q) &= \sum_{n=1}^{N}\sum_{m=1}^{M}\left\{\sum_{d=1}^{D}\mathbb{E}_{z_{nm}}[z_{nm}]\kappa_{nd}\boldsymbol{x}_n^T\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}] + \mathbb{E}_{z_{nm}}[z_{nm}]\mathbb{E}_{\pi_m}[\log \pi_m]\right\} \\ &\quad - \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}_{z_{nm}}[z_{nm}]\log r_{nm} - \sum_{n=1}^{N}\sum_{d=1}^{D}\left\{\log\left[\exp(c_{nd}) + 1\right] + \frac{1}{2}c_{nd}\right\} + \frac{1}{2}lDM \\ &\quad + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}\Big\{-[\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}] - \boldsymbol{\mu}_{0dm}]^T\boldsymbol{S}_{0dm}^{-1}[\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}] - \boldsymbol{\mu}_{0dm}] \\ &\quad + \log|\boldsymbol{S}_{dm}| - \log|\boldsymbol{S}_{0dm}| - \mathrm{tr}\left(\boldsymbol{S}_{0dm}^{-1}\boldsymbol{S}_{dm}\right)\Big\} \\ &\quad + \log C(\boldsymbol{\phi}) - \log C(\boldsymbol{\phi_0}) + \sum_{m=1}^{M}(\phi_{0m} - \phi_m)\mathbb{E}_{\pi_m}[\log \pi_m].\end{aligned} \tag{3.86}$$

Finally, we summarize this model section with Algorithm 10, where we present the Variational Bayes scheme for the Finite Bernoulli mixture with covariates. The steps are similar to Algorithm 9 for the Finite Binomial mixture (without covariates), owing to the Finite components' structure of the model. The variational algorithm stops when the ELBO value has converged.

Algorithm 10 Updating Scheme of the Variational Finite Bernoulli Mixture with Covariates

1: Initialize:

1. Fix the number of components M
2. Choose initial values for the variational parameters:
   $\phi_m, \boldsymbol{\mu}_{dm}$ and $\boldsymbol{S}_{dm}$
3. Choose values for the hyperparameters:
   $\phi_{0m}, \boldsymbol{\mu}_{0dm}$ and $\boldsymbol{S}_{0dm}$

2: Repeat:

1. Calculate the expected values in equations (3.83)
2. Update the variational parameters in equations (3.71), (3.74) and (3.76)

3: Stop:

$\mathcal{L}(\boldsymbol{y}; q)^{\text{current}} - \mathcal{L}(\boldsymbol{y}; q)^{\text{previous}} \leq \epsilon$, where $\epsilon = 10^{-6}$

4: Final step:

Calculate the posterior mixing weight π_m as in equation (3.84)

3.3.3 Variational Dirichlet Process Poisson Mixture with Covariates

There are cases where clustering count data with unknown number of non-independent trials is in demand. In such scenarios, Binomial mixture cannot be used and thus, alternative model-based tools are employed. A suitable mixture model is the Poisson mixture and especially the Dirichlet Process Poisson mixture with covariates. This is a clustering method that automatically determines the number of groups while taking into account the presence of confounding parameters. To give an example, for individuals that differ by sex and age, the number of significantly affected CpG sites - in terms of aberrant methylation - within a differentially methylated region (DMR) is recorded. In this case, the CpG counts could be modelled by a Dirichlet Process Poisson mixture with covariates and not a Dirichlet Process Binomial mixture model with covariates, since the CpG methylation values within a DMR are correlated, violating the Binomial's condition of independence between the experiments (here CpGs within a DMR).

In this section, we begin with the proper construction of the model in order to include the covariates/confounders whilst we carry on with the presentation of the hierarchical structure. It is also worth mentioning that the Dirichlet Process Poisson mixture with covariates coincides with the Dirichlet Process Negative Binomial mixture with covariates when we fix the over-dispersion parameter. The reason is the Gamma prior imposed upon the sub-population's Poisson mean value. For a Negative Binomial model with random over-dispersion, we could utilise the Pólya-Gamma representation presented in the previous section.

The Poisson model parameter λ_{ndm}, which is linked to the mean value of the n^{th} sample for the d^{th} feature dimension in the m^{th} sub-population, can be parametrized by a linear predictor η_{ndm} (equation (3.87)), so as to consider the occurrence of confounding factors (η_{ndm} includes the sample specific covariates along with their component and feature specific coefficients)

$$\lambda_{ndm} = \exp(\eta_{ndm}), \quad \text{with } \eta_{ndm} = \boldsymbol{x}_n^T \boldsymbol{\beta}_{dm}. \tag{3.87}$$

The Dirichlet Process Poisson mixture likelihood given the confounding variables $\boldsymbol{x}_n$ can be written as

$$\begin{aligned} P(\boldsymbol{y} \mid \boldsymbol{\eta}, \boldsymbol{z}) &= \prod_{n=1}^{N} \prod_{d=1}^{D} \prod_{m=1}^{M} \left[y_{nd}!^{-1} \exp(\eta_{ndm})^{y_{nd}} \exp(-\exp(\eta_{ndm})) \right]^{z_{nm}} \\ &= \prod_{n=1}^{N} \prod_{d=1}^{D} \prod_{m=1}^{M} \left[y_{nd}!^{-1} \exp(\eta_{ndm} y_{nd}) \exp(-\exp(\eta_{ndm})) \right]^{z_{nm}}. \end{aligned} \tag{3.88}$$

However, the likelihood form in equation (3.88) does not resemble a known distribution, complicating this way the Bayesian inference. To overcome this obstacle, we refer to Bartlett and Kendall [6] and Prentice [114] who prove that if the mean Poisson value λ_{ndm} follows a Gamma density (as it would have been originally the case in a Bayesian framework, if covariates were not considered), then the linear predictor η_{ndm} is shown to be distributed as a Gaussian distribution with mean value $\log(y_{nd})$ and scalar variance y_{nd}^{-1}, where y_{nd} is the datapoint of the n^{th} sample at the d^{th} feature (see equation (3.91)).

$$\begin{aligned} &\lambda_{ndm} \sim \text{Gamma}(\lambda_{ndm} \mid a_{ndm}, b_{ndm}), \quad \text{then:} \\ &\eta_{ndm} = \log(\lambda_{ndm}) \sim \mathcal{N}(\eta_{ndm} \mid \log(a_{ndm}) + \log(b_{ndm}), a_{ndm}^{-1}) \quad \text{for large } a_{ndm}. \end{aligned} \tag{3.89}$$

By setting $b_{ndm} = 1$ and $a_{ndm} = y_{nd}$

$$P(\lambda_{ndm} \mid y_{nd}, 1) = \frac{\lambda_{ndm}^{y_{nd}}}{\Gamma(y_{nd})} \exp(-\lambda_{ndm}), \tag{3.90}$$

and with the change of variable formula

$$\begin{aligned} P(\eta_{ndm} \mid y_{nd}, 1) &= P(\lambda_{ndm} = \exp(\eta_{ndm}) \mid y_{nd}, 1) \; \frac{\partial \exp(\eta_{ndm})}{\partial \eta_{ndm}} \\ &= \frac{1}{\Gamma(y_{nd})} \exp(\eta_{ndm} y_{nd}) \exp(-\exp(\eta_{ndm})) \\ &\approx \mathcal{N}(\eta_{ndm} \mid \log(y_{nd}), y_{nd}^{-1}). \end{aligned} \tag{3.91}$$

This Gaussian approximation in (3.91) transforms the Dirichlet Process mixture likelihood into

$$\begin{aligned} P(\boldsymbol{y} \mid \boldsymbol{\eta}, \boldsymbol{z}) &\approx \prod_{n=1}^{N} \prod_{d=1}^{D} \prod_{m=1}^{M} \left[{y_{nd}}^{-1} y_{nd}^{1/2} (2\pi)^{-1/2} \exp\left(-\frac{1}{2} y_{nd} \left[\boldsymbol{x}_n \boldsymbol{\beta}_{dm} - \log y_{nd} \right]^2 \right) \right]^{z_{nm}} \\ &\approx \prod_{n=1}^{N} \prod_{d=1}^{D} \prod_{m=1}^{M} \left[y_{nd}^{-1/2} (2\pi)^{-1/2} \exp\left(-\frac{1}{2} y_{nd} \left[\boldsymbol{x}_n \boldsymbol{\beta}_{dm} - \log y_{nd} \right]^2 \right) \right]^{z_{nm}}. \end{aligned} \tag{3.92}$$

After the definition of the approximated likelihood in equation (3.92) that facilitates the Bayesian inference with respect to the $\boldsymbol{\beta}$ regression coefficients, we present the prior distributions of the model parameters $\boldsymbol{\beta}, \boldsymbol{z}$ and $\boldsymbol{w}$

$$\boldsymbol{\beta} \sim \prod_{d=1}^{D} \prod_{m=1}^{M} \mathcal{N}_L(\boldsymbol{\beta}_{dm} \mid \boldsymbol{\mu}_{0dm}, \boldsymbol{S}_{0dm}), \tag{3.93}$$

$$\boldsymbol{z} \mid \boldsymbol{w} \sim \prod_{n=1}^{N} \text{Categorical}\left(\boldsymbol{z}_n \mid \boldsymbol{w}\right), \tag{3.94}$$

$$\boldsymbol{w} \sim \prod_{m=1}^{M} \text{Beta}\left(w_m \mid 1, \phi_{0m}\right), \tag{3.95}$$

where $\boldsymbol{\beta}$ is the $L \times D \times M$ covariates coefficients array, with L denoting the number of predictors. Each element $\boldsymbol{\beta}_{dm}$ follows *a-priori* an L-dimensional Gaussian, with mean array $\boldsymbol{\mu}_{0dm}$ and covariance array $\boldsymbol{S}_{0dm}$ (similarly to the Finite Bernoulli mixture with covariates). The latent allocation variable $\boldsymbol{z}$ and the stick breaking point variable $\boldsymbol{w}$ are defined equivalently as in the Dirichlet Process Beta mixture model.

In Figure 3.4, we provide the links between the model parameters in a Directed Acylic Graph. Based on the conditional dependencies in this graph, we can derive the Mean Field approximated posterior distribution. This joint variational distribution takes a tractable form when $q(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\beta}) = q(\boldsymbol{z})q(\boldsymbol{w})q(\boldsymbol{\beta})$.

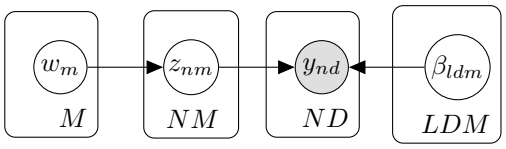


Figure 3.4 Directed Acyclic Graph of the Dirichlet Process Poisson mixture model with covariates. The nodes represent the random variables, the directed edges the conditional dependence and the boxes the dimensionality of each parameter. The light grey node corresponds to the variable datapoint y_{nd}.

Thus far, we have familiarized ourselves with the variational derivation of the latent allocation variable $\boldsymbol{z}$, the stick-breaking point $\boldsymbol{w}$ and the covariates coefficients $\boldsymbol{\beta}$ due to the predecessor mixture models. Therefore, we avoid explaining in detail the Variational Bayes steps in the Dirichlet Process Poisson mixture with covariates and simply display the final form of the variational distributions as well as the Evidence Lower Bound closed form equation.

Variational distribution of $\boldsymbol{z}$

In regard to the variational derivation of $\boldsymbol{z}$,

$$
\begin{aligned}
\log q(\boldsymbol{z}) &\propto \mathbb{E}_{/\boldsymbol{z}}\left[\log P(\boldsymbol{y} \mid \boldsymbol{\eta}, \boldsymbol{z}) + \log P(\boldsymbol{z} \mid \boldsymbol{w})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{z}}\Bigg[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{-\frac{1}{2}y_{nd}\left[(\boldsymbol{x}_n\boldsymbol{\beta}_{dm})^2 - 2\boldsymbol{x}_n\boldsymbol{\beta}_{dm}\log y_{nd}\right]\right\} \\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\log w_m + \sum_{j=1}^{m-1}\log(1-w_j)\right\}\Bigg] \\
&= \mathbb{E}_{/\boldsymbol{z}}\Bigg[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{-\frac{1}{2}y_{nd}\left\{\mathbb{E}_{\boldsymbol{\beta_{dm}}}[(\boldsymbol{x}_n\boldsymbol{\beta}_{dm})^2] - 2\boldsymbol{x}_n\mathbb{E}_{\boldsymbol{\beta_{dm}}}[\boldsymbol{\beta}_{dm}]\log y_{nd}\right\}\right\} \\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\mathbb{E}_{w_m}[\log w_m] + \sum_{j=1}^{m-1}\mathbb{E}_{w_m}[\log(1-w_j)]\right\}\Bigg], \\
&= \mathbb{E}_{/\boldsymbol{z}}\Bigg[\sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\Bigg\{-\frac{1}{2}\sum_{d=1}^{D} y_{nd}\left\{\mathbb{E}_{\boldsymbol{\beta_{dm}}}[(\boldsymbol{x}_n\boldsymbol{\beta}_{dm})^2] - 2\boldsymbol{x}_n\mathbb{E}_{\boldsymbol{\beta_{dm}}}[\boldsymbol{\beta}_{dm}]\log y_{nd}\right\} \\
&\qquad + \mathbb{E}_{w_m}[\log w_m] + \sum_{j=1}^{m-1}\mathbb{E}_{w_m}[\log(1-w_j)]\Bigg\}\Bigg]. \qquad (3.96)
\end{aligned}
$$

The unormalized $\log q(\boldsymbol{z})$ in equation (3.96) is the logarithmic kernel of an N product of Categorical densities. If we define the expression inside the brackets as $\log \rho_{nm}$

$$\begin{aligned}\log \rho_{nm} = &-\frac{1}{2}\sum_{d=1}^{D} y_{nd}\left\{\mathbb{E}_{\boldsymbol{\beta_{dm}}}[(\boldsymbol{x}_n\boldsymbol{\beta}_{dm})^2] - 2\boldsymbol{x}_n\mathbb{E}_{\boldsymbol{\beta_{dm}}}[\boldsymbol{\beta}_{dm}]\log y_{nd}\right\} \\ &+ \mathbb{E}_{w_m}[\log w_m] + \sum_{j=1}^{m-1}\mathbb{E}_{w_m}[\log(1-w_j)],\end{aligned} \tag{3.97}$$

and normalize $\log q(\boldsymbol{z})$ by setting $r_{nm} = \rho_{nm}/\sum_{j=1}^{M}\rho_{nj}$, we obtain the variational distribution of $\boldsymbol{z}$ in equation (3.98), which is a log-product of N Categorical densities

$$\log q(\boldsymbol{z}) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log r_{nm}. \tag{3.98}$$

Variational distribution of $\boldsymbol{w}$

In respect of the variational derivation step for the stick-breaking point $\boldsymbol{w}$, this is equivalent to the corresponding one in the Dirichlet Process Beta mixture (Chapter 3, Subsection 3.2.1). The variational $q(\boldsymbol{w})$ results in a product of Beta densities in equation (3.99), with variational shape parameters given in equation (3.100).

$$\begin{aligned}\log q(\boldsymbol{w}) &\propto \mathbb{E}_{/\boldsymbol{w}}[\log P(\boldsymbol{z}\mid\boldsymbol{w}) + \log P(\boldsymbol{w})] \\ &\propto \mathbb{E}_{/\boldsymbol{w}}\left[\sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\log w_m + \sum_{j=1}^{m-1}\log(1-w_j)\right\} + \sum_{m=1}^{M}(\phi_{0m}-1)\log(1-w_m)\right] \\ &= \sum_{m=1}^{M}\left\{\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\log w_m + \left(\sum_{n=1}^{N}\sum_{j=m+1}^{M}\mathbb{E}_{\boldsymbol{z}_{nj}}[z_{nj}] + \phi_{0m} - 1\right)\log(1-w_m)\right\},\end{aligned} \tag{3.99}$$

$$\begin{aligned}\delta_m &= 1 + \sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}], \\ \phi_m &= \phi_{0m} + \sum_{n=1}^{N}\sum_{j=m+1}^{M}\mathbb{E}_{\boldsymbol{z}_{nj}}[z_{nj}].\end{aligned} \tag{3.100}$$

Variational distribution of $\boldsymbol{\beta}$

The Mean Field posterior density of $\boldsymbol{\beta}$, in its logarithmic form, is proportional to the expected summation of the log-likelihood after the Gaussian approximation (equation (3.92)) and the log-prior of $\boldsymbol{\beta}$

$$\begin{aligned}\log q(\boldsymbol{\beta}) &\propto \mathbb{E}_{/\beta}[\log P(\boldsymbol{y}\mid\boldsymbol{\eta},\boldsymbol{z}) + \log P(\boldsymbol{\beta})] \\ &\propto \mathbb{E}_{/\beta}\Bigg[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{-\frac{1}{2}\log y_{nd} - \frac{1}{2}\log 2\pi - \frac{1}{2}y_{nd}\left[\boldsymbol{x}_n\boldsymbol{\beta}_{dm} - \log y_{nd}\right]^2\right\} \\ &\quad + \sum_{d=1}^{D}\sum_{m=1}^{M}\left\{-\frac{l}{2}\log(2\pi) - \frac{1}{2}\mid\boldsymbol{S}_{0m}\mid - \frac{1}{2}\left[\boldsymbol{\beta}_{dm} - \boldsymbol{\mu}_{0dm}\right]^T\boldsymbol{S}_{0dm}^{-1}\left[\boldsymbol{\beta}_{dm} - \boldsymbol{\mu}_{0dm}\right]\right\}\Bigg]\end{aligned}$$

$$
\begin{aligned}
&\propto \mathbb{E}_{/\beta}\Bigg[\sum_{d=1}^{D}\sum_{m=1}^{M}\Bigg\{\sum_{n=1}^{N} z_{nm}y_{nd}\log y_{nd}\boldsymbol{x}_n\boldsymbol{\beta}_{dm} - \frac{1}{2}\sum_{n=1}^{N} z_{nm}y_{nd}\boldsymbol{\beta}_{dm}{}^{T}\boldsymbol{x}_n{}^{T}\boldsymbol{x}_n\boldsymbol{\beta}_{dm} \\
&\quad - \frac{1}{2}\boldsymbol{\beta}_{dm}{}^{T}\boldsymbol{S}_{0dm}^{-1}\boldsymbol{\beta}_{dm} + \boldsymbol{\mu}_{0dm}^{T}\boldsymbol{S}_{0dm}^{-1}\boldsymbol{\beta}_{dm}\Bigg\}\Bigg] \\
&= \mathbb{E}_{/\beta}\Bigg[\sum_{d=1}^{D}\sum_{m=1}^{M}\Bigg\{\Bigg(\sum_{n=1}^{N} z_{nm}y_{nd}\log y_{nd} + \boldsymbol{\mu}_{0dm}^{T}\boldsymbol{S}_{0dm}^{-1}\Bigg)\boldsymbol{\beta}_{dm} \\
&\quad - \frac{1}{2}\boldsymbol{\beta}_{dm}{}^{T}\Bigg(\sum_{n=1}^{N} z_{nm}y_{nd}\boldsymbol{x}_n{}^{T}\boldsymbol{x}_n + \boldsymbol{S}_{0dm}^{-1}\Bigg)\boldsymbol{\beta}_{dm}\Bigg\}\Bigg] \\
&= \sum_{d=1}^{D}\sum_{m=1}^{M}\Bigg\{\Bigg(\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}]y_{nd}\log y_{nd} + \boldsymbol{\mu}_{0dm}^{T}\boldsymbol{S}_{0dm}^{-1}\Bigg)\boldsymbol{\beta}_{dm} \\
&\quad - \frac{1}{2}\boldsymbol{\beta}_{dm}{}^{T}\Bigg(\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}]y_{nd}\boldsymbol{x}_n{}^{T}\boldsymbol{x}_n + \boldsymbol{S}_{0dm}^{-1}\Bigg)\boldsymbol{\beta}_{dm}\Bigg\}.
\end{aligned}
\tag{3.101}
$$

Equation (3.101) indicates that $\log q(\boldsymbol{\beta})$ is a logarithmic kernel of $D \times M$ multivariate Gaussian distributions with variational covariance and mean array defined as

$$
\begin{aligned}
\boldsymbol{S}_{dm} &= \Bigg\{\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}]y_{nd}\boldsymbol{x}_n{}^{T}\boldsymbol{x}_n + \boldsymbol{S}_{0dm}^{-1}\Bigg\}^{-1}, \\
\boldsymbol{\mu}_{dm} &= \boldsymbol{S}_{dm}\Bigg\{\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}]y_{nd}\log y_{nd} + \boldsymbol{\mu}_{0dm}^{T}\boldsymbol{S}_{0dm}^{-1}\Bigg\}.
\end{aligned}
\tag{3.102}
$$

Variational Expectations

Based on the variational distributions, we compute the expectations within the variational equations and the approximated posterior mixing weights in equation (3.104)

$$
\begin{aligned}
\mathbb{E}_{z_{nm}}[z_{nm}] &= r_{nm}, \\
\mathbb{E}_{w_m}[\log w_m] &= \Psi(\delta_m) - \Psi(\delta_m + \phi_m), \\
\mathbb{E}_{w_m}[\log(1 - w_m)] &= \Psi(\phi_m) - \Psi(\delta_m + \phi_m), \\
\mathbb{E}_{w_m}[w_m] &= \frac{\delta_m}{\phi_m + \delta_m}, \\
\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}] &= \boldsymbol{\mu}_{dm}, \\
\mathbb{E}_{\boldsymbol{\beta}_{\boldsymbol{dm}}}[(\boldsymbol{x}_n\boldsymbol{\beta}_{dm})^2] &= (\boldsymbol{y}_n^{T}\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}])^2 + \boldsymbol{y}_n^{T}\mathrm{Cov}(\boldsymbol{\beta}_{dm})\boldsymbol{y}_n
\end{aligned}
\tag{3.103}
$$

$$
\pi_m = \mathbb{E}_{w_m}[w_m]\prod_{j=1}^{m-1}\big(1 - \mathbb{E}_{w_j}[w_j]\big). \tag{3.104}
$$

Variational Lower Bound

The Evidence Lower Bound is finally computed in closed form and it is a function of the variational parameters of the Dirichlet Process Poisson mixture when covariates exist. The general ELBO function for this model is given in equation (3.105) while the explicit result is obtained in (3.106).

$$\begin{aligned}\mathcal{L}(\boldsymbol{y};q) &= \mathbb{E}_{z,\pi,\beta}[\log P(\boldsymbol{y},\boldsymbol{z},\boldsymbol{\eta},\boldsymbol{\pi})] - \mathbb{E}_z[\log q(\boldsymbol{z})] - \mathbb{E}_\pi[\log q(\boldsymbol{\pi})] - \mathbb{E}_\beta[\log q(\boldsymbol{\beta})] \\ &= \mathbb{E}_{z,\pi,\beta}[\log P(\boldsymbol{y} \mid \boldsymbol{z},\boldsymbol{w}) + \log P(\boldsymbol{z} \mid \boldsymbol{\pi}) + \log P(\boldsymbol{\pi}) + \log P(\boldsymbol{\beta})] \\ &\quad - \mathbb{E}_z[\log q(\boldsymbol{z})] - \mathbb{E}_\pi[\log q(\boldsymbol{\pi})] - \mathbb{E}_\beta[\log q(\boldsymbol{\beta})].\end{aligned} \tag{3.105}$$

The explicit result is

$$\begin{aligned}\mathcal{L}(\boldsymbol{y};q) &= \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}_{z_{nm}}[z_{nm}]\left\{\log\rho_{nm} - \frac{1}{2}\log y_{nd} - \frac{1}{2}y_{nd}(\log y_{nd})^2 - \frac{1}{2}\log 2\pi\right\} \\ &\quad - \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}_{z_{nm}}[z_{nm}]\log r_{nm} + \sum_{m=1}^{M}[\log\phi_{0m} + (\phi_{0m}-1)\log(1-w_m)] \\ &- \sum_{m=1}^{M}\Bigg\{\log\Gamma(\delta_m+\phi_m) - \log\Gamma(\delta_m) - \log\Gamma(\phi_m) + (\delta_m-1)\log w_m \\ &+ (\phi_m-1)\log(1-w_m)\Bigg\} \\ &+ \sum_{d=1}^{D}\sum_{m=1}^{M}\Bigg\{-\frac{1}{2}\log|\boldsymbol{S}_{0m}| - \frac{1}{2}\left[\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}] - \boldsymbol{\mu}_{0dm}\right]^T\boldsymbol{S}_{0dm}^{-1}\left[\mathbb{E}_{\beta_{dm}}[\boldsymbol{\beta}_{dm}] - \boldsymbol{\mu}_{0dm}\right] \\ &+ \frac{1}{2}\log|\boldsymbol{S}_m| + \frac{l}{2} - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{S}_{0m}^{-1}\boldsymbol{S}_m\right)\Bigg\}.\end{aligned} \tag{3.106}$$

Algorithm 11 Updating Scheme of the Variational Dirichlet Process Poisson Mixture with Covariates

1: Initialize:

1. Choose the initial number of components M
2. Choose initial values for the variational parameters:
   $\delta_m, \phi_m, \boldsymbol{\mu}_{dm}$ and $\boldsymbol{S}_{dm}$
3. Choose values for the hyperparameters:
   $\phi_{0m}, \boldsymbol{\mu}_{0dm}$ and $\boldsymbol{S}_{0dm}$

2: Repeat:

1. Calculate the expected values in equations (3.103)
2. Update the variational parameters in equations (3.97), (3.100) and (3.102)

3: Stop:

$\mathcal{L}(\boldsymbol{y};q)^{\text{current}} - \mathcal{L}(\boldsymbol{y};q)^{\text{previous}} \leq \epsilon$, where $\epsilon = 10^{-6}$

4: Pre-Final step:

Calculate the posterior mixing weight π_m as $\pi_m = \mathbb{E}_{w_m}[w_m]\prod_{j=1}^{m-1}\left(1 - \mathbb{E}_{w_j}[w_j]\right)$

5: Final step:

Discard components with almost zero weight

In conclusion, we provide in Algorithm 11 the variational scheme for the Dirichlet Process Poisson mixture model with covariates. This is similar to the updating

procedure of the Variational Dirichlet Process Beta mixture in Algorithm 8, due to the automatic cluster determination structure through Dirichlet Process.

3.4 Summary

In this chapter, we provided the full Mean Field procedure for four mixture models that apply to a plethora of different DNA methylation data types such as beta-intensities extracted from array-based platforms, methylated reads (counts) given by Bisulfite Sequencing techniques, or even binary data that refer to significantly or non significantly methylated DNA regions. Moreover, two of the models that we presented were accordingly structured to consider the existence of factors that distort the clustering outcome (covariates). Regarding the four models, these were the variational Dirichlet Process Beta mixture, the Finite Binomial mixture, the Finite Bernoulli mixture with covariates and the Dirichlet Process Poisson mixture with covariates. In the variational Finite mixture of Binomial distributions, we also introduced the simple implementation of "annealing". Furthermore, due to the non-conjugate nature of most of the presented models, we used the Taylor approximation in the Dirichlet Process Beta mixture to achieve conjugacy, the Pólya-Gamma augmentation in the Finite Bernoulli mixture with covariates and the Gaussian approximation in the Dirichlet Process Poisson mixture with covariates. At this point we warn that the Dirichlet Process Bernoulli mixture with covariates, after the Pólya-Gamma augmentation, will increase the burden of the MCMC algorithm due to the extra parameter $\boldsymbol{\omega}$, since it has the same dimensions as the datapoints $\boldsymbol{y}$. On the other hand, the Variational Inference will manage to handle well computationally-wise the inference for this model due to its scalability skill.

Overall, this chapter contributed in comprehending the inferential procedure of Variational Bayesian mixtures, as well as in straightforwardly facilitating the Mean Field derivation of Finite and Dirichlet Process mixture models for continuous and discrete random variables.

Chapter 4

In Silico Experiments

4.1 Overview

The mixture models are probabilistic clustering techniques that belong to a broader class of clustering tools, called unsupervised learning methods. Unsupervised methods are those algorithms that cluster unlabeled datapoints based on sets of features. With the intention to incorporate the advantages of speed and scalability in the clustering procedure of the mixture models, we introduce the variational mixtures, which are models learned via Variational Inference.

In this chapter, we implement variational mixtures of continuous and discrete random variables in order to test their clustering performance prior to real applications. In particular, we create various synthetic scenarios of bounded continuous data, binary and counts and apply the variational Dirichlet Process Beta mixture model, the variational Dirichlet Process Poisson mixture model and the variational Dirichlet Process Bernoulli mixture model respectively (all these models include the "annealing"). We omit applications using the variational Gaussian mixture (logistic transformed beta-intensities: M-values - values of unrestricted support range, see Chapter 1, section 1.5.2). There reason is that it is a widely discussed model in the literature and therefore, plenty of useful materials are available such as the work of Bishop [11] and Blei and Jordan [13]. Moreover, our focus is on presenting models for non-normally distributed data for the cases logarithmic transformation is not successful to conform them to normality. For instance, there can be scenarios where the log-data are still skewed, or sometimes are even more skewed than the original and thus a Gaussian assumption cannot be the solution (Changyong et al. [23]). In such cases, it is better to retain the original data to avoid any non-relevant inference on the log-transformed ones.

Additionally, in this chapter, we compare the clustering accuracy of the tested variational mixtures with fast state-of-the-art unsupervised clustering tools such as K-means (MacQueen et al. [82]), Hierarchical cluster analysis (Johnson [65]) and Density-Based Spatial Clustering of Applications with noise (DBSCAN) (Ester et al. [41]). In the current analysis, we exclude applications and comparisons to Markov chain Monte Carlo algorithms due to their well-known difficulty in scaling to large datasets, as the ones simulated here. For instance, Zhang et al. [155], similarly with us, propose a Dirichlet Process Beta mixture model for clustering methylation beta-intensities. Although they employ Gibbs sampling to learn the model parameters, their clustering procedure is considerably time-consuming and non-scalable (only a handful of CpGs can we analysed), in contrast to the variational Dirichlet Process Beta mixture model presented here. In general, we mainly wish to compare rapid unsupervised techniques to the Variational Bayes algorithm and particularly algorithms extensively used in genetics due to their simplicity, speed and applicability. In general, we point out that the Markov chain Monte Carlo sampling algorithm in Finite mixture models is known as Reversible Jump MCMC and the reader may refer to Richardson and Green [116].

In summary, this chapter is divided in continuous and discrete synthetic applications along with the feature selection *per* component (Lin et al. [76]) at the end of the chapter. The feature selection step is an *a posteriori* procedure that exploits the variational sub-population distributions of the mixture model to discover the features that discriminate the components, based on the accuracy measure introduced in Chapter 2, equation (2.78). Regarding the applications, the continuous cases concern the variational Dirichlet Process Beta mixture, while the discrete cases the variational Dirichlet Process Poisson mixture and the variational Dirichlet Process Bernoulli mixture. For convenience, this point onward we shall refer to the aforementioned variational mixture models with the acronyms VB-DPBM (Variational Bayes Dirichlet Process Beta mixture), VB-DPPM (Variational Bayes Dirichlet Process Poisson mixture) and VB-DPBerM (Variational Bayes Dirichlet Process Bernoulli mixture). As a further note, Principal component Analysis (PCA) is used for dimensionality reduction purposes that solely aids the graphical representation.

4.2 Unsupervised Clustering

With regards to K-means and Hierarchical cluster analysis, these are two non-probabilistic clustering algorithms applied to unsupervised settings. They are easy in implementation and therefore commonly used in genetics/genomics applications for determining the hidden grouping underlying the examined samples. Despite of their versatility, they carry a serious drawback: the pre-definition of the number of components. One usual approach to deal with this obstacle is to run them for a set of different and specified

number of groups, recording the Sum of Squared Errors (SSE) which measures the squared distance of each observation from the centroid of the cluster. However, this selection measure cannot always ensure straightforward results, especially when the real categories overlap or are hard to distinguish. Alternative objective functions can be reviewed in Kodinariya and Makwana [69], however these do not significantly differentiate from the SSE perspective. Fränti and Sieranoja [47] discuss about the suitability of the Centroid Index over SSE which lacks communication of the result's significance. The Centroid Index denotes how many centroids are incorrectly located. Nonetheless, in any index selection, either SSE or Centroid Index, the issue lies in restarting the K-means algorithm several times to test for the optimal number of clusters. This is a time-consuming and probably non-scalable action on high-dimensional structures. Consequently, the need to employ automatic methods for clustering is deemed essential.

An alternative non-probabilistic method that does not require pre-determination of the number of clusters, as opposed to K-means and Hierarchical clustering, is DBSCAN, which is one of the most cited unsupervised data clustering algorithms, proposed by Ester et al. [41] in 1996. DBSCAN clusters together datapoints that are close to each other (nearby neighbors) and form high-density regions, whereas marks as outliers points that have distant neighbors and mostly lie alone. In order for DBSCAN to accomplish its clustering procedure, it necessitates two parameters: a) the radius of a dense region (definition of neighborhood) and b) the minimum number of datapoints that form this dense region. The minimum number of datapoints can be arbitrarily chosen according to the user's tolerance towards noise, *i.e.*, on large datasets it would be sensible to be more lenient by increasing the minimum number. Regarding the radius parameter, this should not be chosen randomly given that too small values may lead to not assigning most points into groups, due to the strict restriction of a narrow neighborhood, labelling them as noise, while large values may merge together nearby clusters returning considerably less number of clusters than the original. Empirical techniques for choosing a proper radius value are therefore employed, as discussed in Ester et al. [41] and Schubert et al. [121]. For instance, we can set an appropriate radius based on the "knee" in the K-Nearest-Neighbor (KNN) distance plot. DBSCAN overall, despite of automatically determining the number of clusters, is a distribution-free method and thus it lacks providing the sub-population's distribution, which is essential in applying the *a posteriori* discriminative feature selection by Lin et al. [76].

To achieve having the extra information of the estimated sub-population's distribution, while simultaneously requiring no prior knowledge about the true number of components, we employ variational Dirichlet Process mixture models. Mixture models are flexible on choosing the component's distribution with respect to the data type, in contrary to K-means, Hierarchical clustering and DBSCAN which are non-probabilistic and

fit regardless all type of data. As a consequence, the non-probabilistic models return solely cluster assignments for the datapoints and mean cluster vectors, in contrast to the variational Dirichlet Process mixture algorithm where the output is: a) the cluster allocation, b) the responsibilities (how strong is our belief that a given datapoint belongs to a specific cluster), c) the sub-population's distribution and d) the parameter estimates for each one of the sub-populations' distributions.

4.3 Bounded Continuous Synthetic Data

To test the variational Dirichlet Process Beta mixture model performance, we build an R function that simulates bounded continuous data by Beta mixtures with independent observations across the samples and across the features (model definition in Chapter 3, equation (3.5)). The component specific Beta parameters $\boldsymbol{u}$ and $\boldsymbol{v}$ (shape variables) are randomly selected. For example, the u_{dm} element of $\boldsymbol{u}$ is generated by a uniform distribution with parameters $(10, 20)$, similarly to v_{dm}. The reason we choose large $\boldsymbol{u}$ and $\boldsymbol{v}$ parameters for our simulations is to ensure each sub-population is modelled by a convex Beta density, since the non-convex Dirichlet Process Beta mixture likelihood is a combination of convex sub-distributions (Zhang et al. [155]). Moreover, we decide on this random manner of selection and not by manually fixing the model parameters, because both $(\boldsymbol{u}, \boldsymbol{v})$ are matrices of high dimensions. To give an example, for a dataset of three clusters and 200 features, we would have to pre-specify $M \times D = 600$ elements of the $\boldsymbol{u}$ matrix ourselves (M is the number of components and D the number of features), which is a rather needless and time-demanding action. In regard to the input variables of the simulation function, the user is able to set effortlessly the number of samples N, the number of features D and the mixing weights $\boldsymbol{\pi} = [\pi_1, ..., \pi_m, ..., \pi_M]$ with π_m representing the probability of a draw to belong to the m^{th} sub-population/component[1].

4.3.1 Clustering Bounded Continuous Data

Bounded data, as the name states, are the points living in a restricted support range, *i.e.*, $[0, 1]$ or generally $[a, b]$. Beta distribution could easily be treated as an appropriate mechanism to catch the behavior of such data owing to its confined $[0, 1]$ nature, which can be easily generalized to any compact domain. For example, if the random variable y lies between $[-1, 1]$ then the transformed $(y+1)/2$ can be modelled by a Beta density. When it comes to clustering bounded continuous data, a Dirichlet Process mixture of Beta densities is an ideal probabilistic technique, which can be upgraded to variational Dirichlet Process Beta mixture after the addition of Variational Bayes for rapid inference.

[1]The software for each simulation is part of the thesis besides the code for the variational mixture models.

Hyperparameters and Initialization

Regarding Variational Bayes, initialization of the variational parameters as well as specification of the priors' hyperparameters is required. For the variational Dirichlet Process Beta mixture in Chapter 3, Section 3.2.1, the stick-breaking point hyperparameter $\boldsymbol{\phi}_0 = [\phi_{01}, ..., \phi_{0M}]$ is given the same small value 0.01 for each one of its elements in order to offer little prior weight to the existence of each cluster, so that any cluster that survives during the Variational Bayes iterations is supported by the data. With respect to u_{dm} and v_{dm} parameters of the m^{th} Beta density of the d^{th} feature, these are identically distributed as Gamma densities with set of hyperparameters $(\alpha_{0dm}, \beta_{0dm})$ and (μ_{0dm}, η_{0dm}), respectively. To choose values for those Gamma hyperparameters, we depend on Ma and Leijon [80] who set the constraint of $\alpha_{0dm}, \beta_{0dm}, \mu_{0dm}$ and $\eta_{0dm} > 0.6156$ to guarantee β_{0dm} and η_{0dm} are greater than zero, as already holds for α_{0dm} and μ_{0dm}.

In regard to the initialization of the algorithm, according to Corduneanu and Bishop [29] the possibility of being trapped to local optima renders the choice of the initial variational parameters critical. In case the initial variational mean cluster vectors are close to each other, the optimization algorithm may not be able to differentiate between the components, resulting in slow convergence and cancellation of many components. This situation could occur because the mixing weights $[\pi_1, ..., \pi_M]$ are constantly updated (in each variational iteration) and a component that is hard to find its place - in terms of significantly changing weight after each iteration and not keeping a more consistent behavior - may be subsequently discarded. To tackle this issue, Corduneanu and Bishop [29] initialized the cluster means of a Gaussian mixture model through K-means clustering.

In our analysis, we avoid initializing by K-means the cluster means in the Dirichlet Process Beta mixture model as well as in the rest models (Dirichlet Process Poisson and Bernoulli mixture) because of the annealing ploy, discussed in Chapter 2, Section 2.4. Annealing is a solution for overcoming the impact of poor starting values. Therefore, the variational parameters of the Dirichlet Process Beta mixture $(\alpha_{dm}, \beta_{dm}, \mu_{dm}, \eta_{dm}, \phi_m)$ are simply initialized by the corresponding hyperparameters, *i.e.*, α_{dm} is initialized by α_{0dm} etc. For higher safety, we choose to slightly differentiate the stick-breaking point variational parameter *per* component ($\boldsymbol{\phi} = [\phi_1, ..., \phi_M]$) by adding little (*i.e.*, $\epsilon_m < 0.001$) yet different values in each one of the elements. This way we hint the algorithm to start from rather dissimilarly weighted clusters. The initial number of components is usually set to a high number (we recommend $M = 100$) after the stick-breaking point process (Chapter 2, Section 2.6.3). However, for computational and time reasons in the simulation study, or when we have a notion to some degree of the possible number, we set a lower value of 25, 20 or even 10. The choice solely depends on the researcher.

We generally encourage to initialize with a reasonably high integer which although is not meaninglessly increasing the convergence time.

Simulation Scenarios

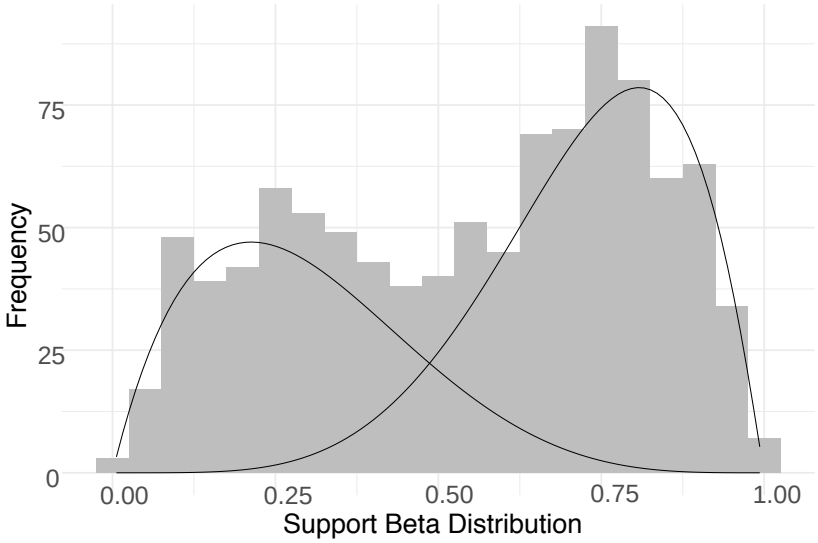


Figure 4.1 Variational density estimation by VB-DPBM for the one dimensional continuous bounded dataset ($D = 1$), with 10K samples ($N = 10\text{K}$) and two clusters ($M = 2$) with mixing weights $[0.4, 0.6]$. The x-axis represents the support range of the Beta distribution. The solid lines denote the two fitted (weighted) components.

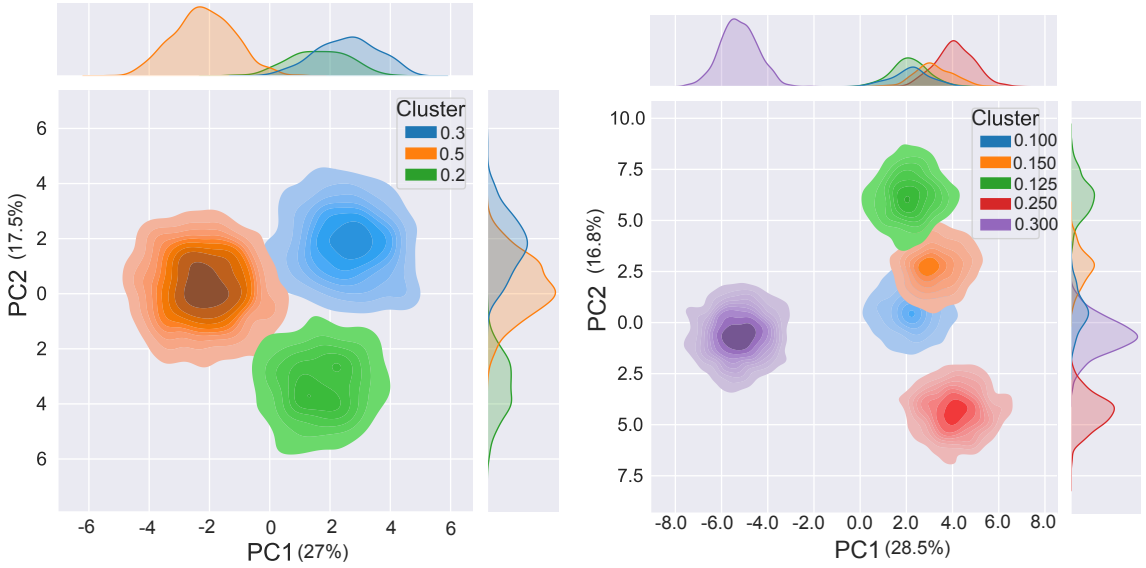
To begin with the simulation study, we first generate a low-dimensional example in terms of feature size, with only one feature ($D = 1$), 10K number of samples ($N = 10\text{K}$) and true number of components equal to two ($M = 2$). The true model parameters can be found in Table 4.1, where $\boldsymbol{\pi}$ is the mixing weights vector and $(\boldsymbol{u}, \boldsymbol{v})$ the component specific shape parameter vectors, with $\boldsymbol{\pi} = [\pi_1, \pi_2]$, $\boldsymbol{u} = [u_1, u_2]$ and $\boldsymbol{v} = [v_1, v_2]$.

	Truth		**VB-DPBM**		**Estimates Divergence**	
$\boldsymbol{\pi}$	0.4	0.6	0.402	0.598	0.002	0.002
$\boldsymbol{u}$	2	5	1.990	5.500	0.010	0.500
$\boldsymbol{v}$	5	2	4.660	2.100	0.340	0.100

Table 4.1 The true and VB-DPBM model parameters of the one dimensional continuous bounded dataset ($D = 1$) with $N = 10K$ and number of components $M = 2$. $\boldsymbol{\pi}$ are the mixing weights, $\boldsymbol{u}$ and $\boldsymbol{v}$ the shape parameters of the Beta mixture. The divergence of the variational estimates from the corresponding true values is also given at the third column.

By applying the VB-DPBM model, we obtain the fitted component specific parameters in Table 4.1 and the fitted component densities in Figure 4.1. In the table, we observe that VB-DPBM manages to correctly determine two clusters with mixing weights

pretty close to the truth. In particular, the estimated weights are 0.402 and 0.598, with the truth being 0.4 and 0.6 respectively. As for the shape parameters $\boldsymbol{u}$ and $\boldsymbol{v}$, the estimated values are considerably close to the true ones with the highest divergence found on $u_2 = 5.5$, which only differentiates by 0.5 units from the true value (see Estimates Divergence column). By using the estimated parameters in Table 4.1, we fit the component densities in the data histogram as displayed in Figure 4.1. Based on the figure, VB-DPBM handles well the bimodality of the simulated dataset by successfully claiming two clusters (two distinct solid lines).



(a) N=1K, D=200, M=3 (b) N=2K, D=200, M=5

Figure 4.2 Two-dimensional variational density plots (Principal Component Analysis is used to present the plots in two dimensions) of two different synthetic continuous bounded datasets: (a) the number of samples is $N = 1K$, the number of features is $D = 200$ and number of components $M = 3$ with mixing weights $[0.5, 0.3, 0.2]$, (b) the number of samples is $N = 2K$, the number of features is $D = 200$ and number of components $M = 5$ with mixing weights $[0.1, 0.15, 0.125, 0.250, 0.375]$. VB-DPBM paints the clusters in different colours, with the corresponding variational mixing weight given on the right of each graph. The marginal component distributions, given the Principal Component, are also displayed at the margins.

The next simulation concerns two bounded continuous datasets of higher feature dimensions. Specifically, the first synthetic dataset has $N = 1$K samples, $D = 200$ features and $M = 3$ number of components with mixing weights $[0.5, 0.3, 0.2]$. The second dataset has $N = 2$K samples, $D = 200$ features and $M = 5$ components with weights $[0.1, 0.15, 0.125, 0.250, 0.375]$. The component specific parameters ($\boldsymbol{u}$ and $\boldsymbol{v}$) are not provided, because these are high-dimensional 200×3 and 200×5 matrices for the first and second scenario respectively, and consequently we avoid to overflow here. However, we mention that each u_{dm} and v_{dm} element of the corresponding matrix is generated randomly from a uniform distribution with hyperparameters $(10, 20)$,

resulting in overlapping clusters of low or medium level. Our anticipation now is to test how well VB-DPBM performs on these two cases. We implement the variational model on each synthetic dataset and depict in 2D plots (Figure 4.2) the clustering allocation based on the full feature dimensions (all $D = 200$ features).

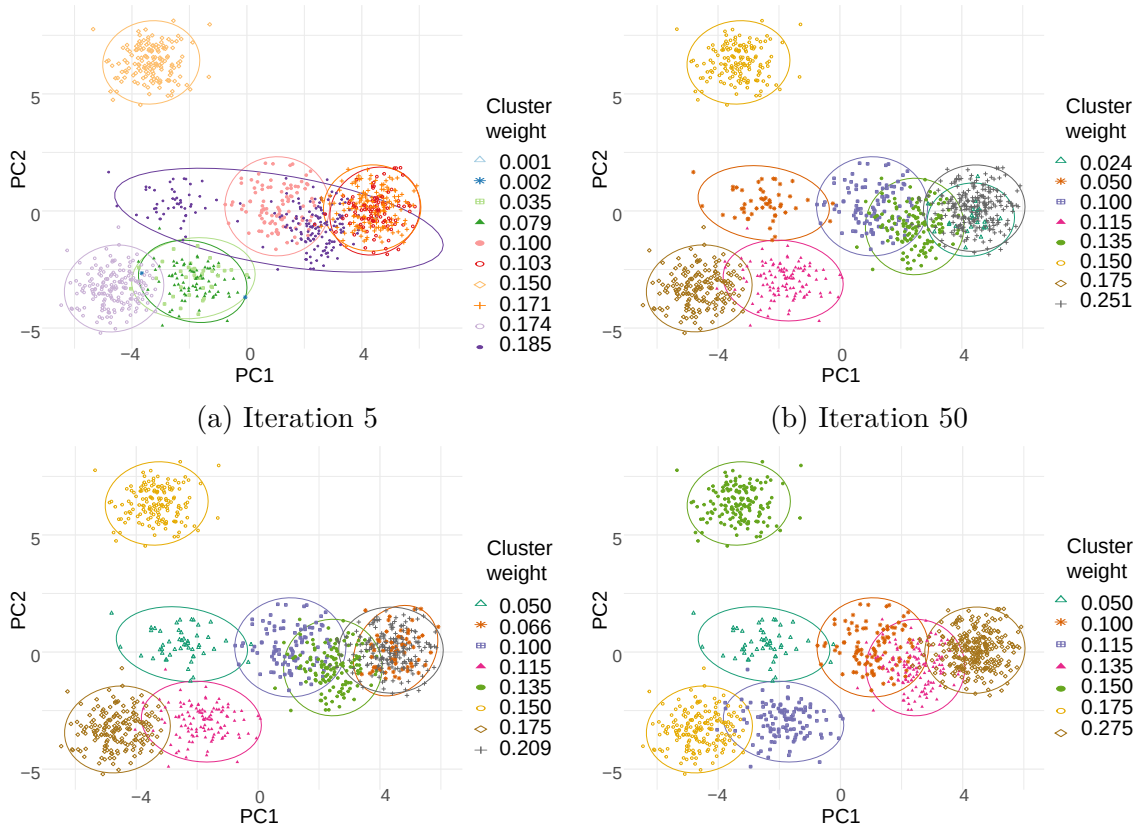
Subfigure 4.2a corresponds to the 2D density plot of the first synthetic dataset of $N = 1K, D = 200$ and $M = 3$, while Subfigure 4.2b to the second, with $N = 2K, D = 200$ and $M = 5$. According to the clustering accuracy measure, the datapoints in both cases are 100% correctly clustered, hence VB-DPBM successfully retrieves the true number of components in both synthetic scenarios, as well as the mixing weight estimates (true density plots same to the variational, therefore omitted). The percentage of variance that PC1 and PC2 explain together in both plots (see labels on x and y axis) has no impact on the clustering results. We remind that the variational algorithms apply to the full datasets and not the first two principal components, thus the 2D graphs serve only as 2D spaces where we colour the clusters estimated on the full feature space.

4.3.2 Mixing Weights Evolution

High interest revolves around the Dirichlet Process and its inclusion in the mixture models, since it helps in the detection of the number of components. In the stick-breaking point implementation, we ought to pre-fix an integer of clusters and then let the variational algorithm conclude to a set of proposed groups. On that account, an interesting question arises about the evolutionary scheme of the mixing weights $\boldsymbol{\pi}$ during the Variational Bayes iterations and especially, about the collapse of clusters with zero importance.

To answer this question, we simulate one bounded continuous dataset of $N = 1$K, $D = 200$ and $M = 7$ components with mixing weights $[0.05, 0.1, 0.115, 0.135, 0.15, 0.175, 0.275]$. The component-specific simulation parameters are randomly selected in accordance with the introductory paragraph of Section 4.3 and the Hyperparameters and Initialization section 4.3.1[2]. We then apply the VB-DPBM algorithm and finally record the clustering evolution at different iterations. This process is analytically depicted in Figure 4.3. In particular, VB-DPBM is initialized with 20 components while at iteration 5 (Subfigure 4.3a) it has already cancelled out half (10 components). At iteration 50 (Subfigure 4.3b) and up until 150 runs (Subfigure 4.3c) the method removes two extra components implying that only a few are left to be discarded, given the slow removal rate. Eventually, the algorithm converges at iteration 200 (Subfigure 4.3d) and to seven clusters with the estimated weights coinciding to the true ones (100% match).

[2]Our aim is to assess the clustering performance (mixing weights and correct allocation of observations) and not the parameters of each component's distribution.



(a) Iteration 5

(b) Iteration 50

(c) Iteration 150

(d) Iteration 200: Convergence

Figure 4.3 Clustering evolution of the VB-DPBM algorithm at different iterations. The synthetic dataset has N = 1K, D = 200 and true M = 7 with mixing weights $[0.05, 0.1, 0.115, 0.135, 0.15, 0.175, 0.275]$. The initial number of components is 20. The clustering results are depicted at: (a) Iteration 5, (b) Iteration 50, (c) Iteration 150 and (d) Iteration 200 (convergence). Each cluster bears its own colour, point shape and estimated mixing weight at each iteration (displayed on the right of each graph).

For further understanding, we provide a second course of action that pictures the evolution of the mixing weight estimates *per* VB-DPBM iteration. For this purpose, we simulate a new bounded continuous dataset with less clusters (for simplicity reasons). Particularly, the synthetic dataset is described by N = 10K samples, D = 200 features, M = 4 components and cluster weights $[0.4, 0.3, 0.2, 0.1]$. In Figure 4.4, the x-axis hosts the VB-DPBM iterations whilst the y-axis represents the 15 clusters the algorithm has been initialized with (M = 15). When the variational algorithm begins (iteration 0 at the x-axis), all clusters are initialized by low and similar mixing weights (1/15 each one - light violet vertical colour at $x = 0$). As the algorithm progresses (iteration 1, 2 etc.), the 15 clusters update their mixing weight (interchange colours of white - purple) up until iteration 200, where only four of them increase (non white colour; white denotes zero percentage) and retain their colour trace (coloured rows of blue and purple hue), indicating convergence of VB-DPBM to four clusters. Convergence is determined when the increase in the ELBO value is negligible (see Chapter 2, Section 2.3 for details on the variational Mean Field algorithm). The resulted component indices are "1",

"4", "3" and "14" with estimated mixing weights $[0.4, 0.3, 0.2, 0.1]$ respectively (100% correctly clustered data).

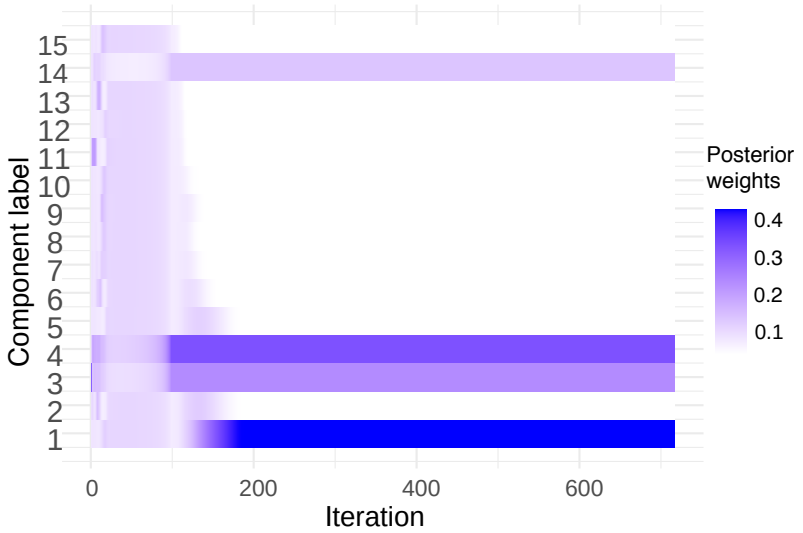


Figure 4.4 Mixing weights' evolution in VB-DPBM iterations. The simulated bounded continuous dataset includes $N = 10K$ samples, $D = 200$ features and number of components $M = 4$ with mixing weights $[0.4, 0.3, 0.2, 0.1]$. Colour intensity corresponds to the component's probability level at each iteration, *i.e.*, white colour implies 0 weight while dark blue weight of 0.4.

4.3.3 Comparison to Standard Methods

In the previous section, visual testing supplied a clear indication of the clustering performance of the VB-DPBM algorithm. However, to obtain concrete results of the performance, we have to quantify the achievement in multiple simulated data scenarios by exploiting a clustering measure. A rational approach would have been to treat the unsupervised algorithms as supervised problems when the ground truth is known, and apply the conventional classification indices (accuracy, F1 score (Goutte and Gaussier [51]) etc.). Nonetheless, despite of knowing the truth in our synthetic scenarios, we are still unable to exploit these measures due to the labelling problem in the variational Dirichlet Process mixtures, as well as in K-means, Hierarchical clustering and DBSCAN. More precisely, the VB-DPBM algorithm for example assigns randomly the labels to the components, *i.e.*, in Figure 4.4 the final components have labels "1", "3", "4"and "14", creating confusion when trying to match these tags with the ground truth where labels are normally given in a sequential manner $\{1, 2, 3, 4\}$. Moreover, all the unsupervised clustering methods that do not require pre-determination of the number of components cannot ensure they will find the exact same number of clusters. For instance, if VB-DPBM retrieves five clusters instead of the correct number four then it is infeasible to calculate the conventional classification measures that require same number of clusters as well as same labelling.

To deal with this situation, we need a measure that computes the accuracy of clustering regardless of the actual labelling, as long as the members of a cluster are allocated together. An astute way is to find which estimated labels correspond to which true ones so as to produce the highest clustering accuracy. To give an example, suppose we have the true label vector for 12 datapoints $[1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4]$ and the estimated vector viaVB-DPBM, $[14, 14, 4, 1, 11, 2, 4, 4, 14, 11, 11, 11]$. We therefore create the confusion matrix of those two vectors in Table 4.2.

		True labels			
		1	**2**	**3**	**4**
VB-DPBM labels	**14**	2	0	0	1
	11	0	0	0	3
	4	0	1	2	0
	1	0	2	0	0
	2	0	1	0	0

Table 4.2 Confusion matrix of true versus VB-DPBM component labels in a toy example with 12 datapoints. The grey boxes denote the counts of the correctly clustered datapoints.

In Table 4.2, we observe that the only two datapoints with true label "1" have been clustered as label "14" by the VB-DPBM algorithm. Thus, we figure out that the true label "1" corresponds to VB-DPBM label "14", having so far two datapoints that have been successfully allocated together. Regarding the four datapoints with true label "2", VB-DPBM manages to allocate together two of them under the VB-DPBM label "1", while the two left have been assigned different labels (VB-DPBM labels "4" and "2"). Hence, we count two datapoints (the highest count) as correctly clustered under the true label "2". Similarly, we count those datapoints that have been successfully assigned together into the rest true clusters "3" and "4". Those counts are highlighted in grey boxes in Table 4.2. Eventually, the accuracy measure is calculated as $(2+2+2+3)/12 = 0.75$, indicating 75% clustering accuracy of the VB-DPBM algorithm. This procedure can be straightforwardly executed by the Python function of Han [54] that uses the Hungarian algorithm to solve this bipartite graph (Asratian et al. [4]).

The next step, given this accuracy measure, is to compare the clustering performance of the VB-DPBM model with the non-probabilistic K-means, Hierarchical clustering and DBSCAN.

Regarding the principal advantage of the VB-DPBM model, this is its scalability to high dimensions (sample-wise and feature-wise). Hence, we are interested in checking the algorithm's clustering behaviour in two synthetic scenario sets: a) when the number of samples exponentially grow, while the number of features is fixed at $D =$ 100 (Table 4.3) and b) when the sample size is relatively low ($N =$200), while the

feature dimensions increase by one extra digit each time (Table 4.5). In the first set of simulations, the synthetic data have seven clusters $M = 7$ with mixing weights $[0.05, 0.1, 0.115, 0.135, 0.15, 0.175, 0.275]$, whereas in the second set $M = 3$ with weights $[0.6, 0.2, 0.2]$.

	Correctly clustered %			
Samples	**VB-DPBM**	**K-means**	**Hierarchical**	**DBSCAN**
$\boldsymbol{N}$**=1M**	100 (7)	81.66 (5)	-	99.80 (7)
$\boldsymbol{N}$**=100K**	100 (7)	81.64 (5)	-	99.23 (7)
$\boldsymbol{N}$**=10K**	100 (7)	81.79 (5)	99.67 (7)	99.17 (7)
$\boldsymbol{N}$**=1K**	100 (7)	81.50 (5)	100 (7)	97.40 (7)

Table 4.3 Clustering performance of VB-DPBM, K-means, Hierarchical clustering and DBSCAN on four bounded continuous synthetic of increasing sample size (N = 1K to N = 1M), fixed number of features $D = 100$ and number of components $M = 7$ with the mixing weights being $[0.05, 0.1, 0.115, 0.135, 0.15, 0.175, 0.275]$. The rates represent the percentage of correctly clustered observations and the values inside the parentheses the determined number of components (except K-means and Hierarchical where the number of clusters is fixed by the Elbow method). Dash line denotes algorithm's inability to scale in such large sample scenarios.

In Table 4.3, the VB-DPBM model manages to cluster without error the bounded continuous data of all the different sample sizes of 1K, 10K, 100K and 1M. Specifically, for the N = 1M case, VB-DPBM requires only 55 minutes to converge, while the Markov chain Monte Carlo algorithms would take several days/weeks. K-means, Hierarchical clustering and DBSCAN converge faster than VB-DPBM, as expected, due to the considerably lower amount of parameter estimations (non-probabilistic algorithms whereas VB-DPBM probabilistic). Given that K-means and Hierarchical clustering do not determine the number of clusters, we select a candidate number by the SSE analysis (Elbow method, Marutho et al. [87]). K-means seems to perform satisfactory on this dataset, however not as successfully as VB-DPBM, with average accuracy almost 82% and five clusters being determined by the SSE analysis instead of seven, in all the different sample size cases. Hierarchical analysis is precise in the lowest two samples (N = 1K and 10K) yet in the greater sample size scenarios is unable to scale and returns no results (dash lines). DBSCAN presents high accuracy in every case, with the peak value 99.8% achieved in N = 1M. Overall, VB-DPBM wins over all three clustering methods by constantly performing 100% accuracy.

To re-enforce these results, we populate the simulations in each scenario by changing the seed 20 times (in R: `set.seed(`x`)` with $x = 123, 124, ...$ etc.) in order to end up with 20 simulations *per* sample size case. The mean clustering performance is recorded along with the standard deviation in Table 4.4. This indicates the consistency and

superiority of the VB-DPBM algorithm - highest accuracy and lowest variance in all sample sizes.

	Correctly clustered %			
Samples (20 in each)	**VB-DPBM**	**K-means**	**Hierarchical**	**DBSCAN**
***N*=1M**	99 (3.10)	82.25 (3.68)	-	98.90 (3.70)
***N*=100K**	98.25 (3.20)	81 (4.45)	-	98 (4.90)
***N*=10K**	98 (3.37)	82 (5.80)	97 (3.89)	97.20 (4.98)
***N*=1K**	97 (3.41)	82.50 (3.55)	96.30 (6.33)	96.40 (4.55)

Table 4.4 Mean clustering performance of VB-DPBM, K-means, Hierarchical clustering and DBSCAN based on 20 bounded continuous simulations for each sample size category (N = 1K to N = 1M), fixed number of features D = 100 and number of components M = 7 with the mixing weights being $[0.05, 0.1, 0.115, 0.135, 0.15, 0.175, 0.275]$. The main values correspond to the mean clustering accuracy - based on the 20 simulations in each scenario - of each algorithm. The value inside the parenthesis is the standard deviation.

	Correctly clustered %			
Features	**VB-DPBM**	**K-means**	**Hierarchical**	**DBSCAN**
***D*=100K**	80 (2)	70.50 (5)	79 (5)	82 (4)
***D*=10K**	100 (3)	76.50 (4)	99.50 (4)	95.50 (3)
***D*=1K**	99 (3)	98 (3)	98 (3)	94 (3)
***D*=100**	100 (3)	80 (2)	80 (2)	92.50 (2)
***D*=10**	99 (3)	74.50 (4)	93 (4)	85 (4)

Table 4.5 Clustering performance of VB-DPBM, K-means, Hierarchical clustering and DBSCAN on bounded continuous synthetic data of varying dimensions (D = 10 to D = 100K), fixed sample N = 200 and components M = 3 with mixing weights $[0.6, 0.2, 0.2]$. The rates correspond to the accuracy of the algorithm in correctly clustering the simulated datapoints and the values inside the parentheses to the determined number of components (except K-means and Hierarchical where the number of clusters is fixed by the Elbow method).

In regard to the performance of the clustering techniques in simulated datasets of small number of samples and escalating number of features, we refer to Table 4.5. We notice that VB-DPBM is accurate by correctly clustering 99% and 100% of the datapoints in all the cases with varying feature number between D = 10 and D = 10K. For the same cases, Hierarchical clustering and DBSCAN perform sufficiently well by both obtaining accuracy more than 94% in the datasets of D = 1K and D = 10K, whilst for the lower feature scenarios D = 10 and D = 100 Hierarchical achieves 93% and 80% respectively, whereas DBSCAN 85% and 92.5%. We comment here that DBSCAN retrieves three components - as many as the true number - in the 10K case, albeit the clustering performance is not 100% (= 95.5%). This shows that some samples

have been allocated into the wrong cluster, however the true number of components is retained. For example, in the cluster with true mixing weight 60%, the algorithm assigned correctly 55.5% and the remaining 4.5% was incorrectly allocated into the second cluster of true weight 20%, falsely changing it into 24.5%. The third and last cluster remained intact at 20%. Thus, the correctly clustered samples reach the percentage of 55.5% + 20% + 20% = 95.5%. Similar is the explanation for all those cases of three components, yet with clustering performance < 100%.

In principle, high clustering accuracy is associated with higher chances to retrieve the true number of clusters, however it is not necessary the true number of clusters will be determined. To illustrate this, accuracy 100% means the true amount of components, *i.e.* 10, is determined. An accuracy of 95% would denote that 5% of the data are incorrectly clustered. If the estimated components were 10 (as the truth), then this 5% would have been falsely allocated into one of the 9 wrong clusters. On the other hand, if the estimated components were 11 then this 5% would create a cluster of its own.

	$\boldsymbol{\pi}$		
Truth	0.6	0.2	0.2
VB-DPBM	0.6	0.4	

Table 4.6 The variational component weights for the synthetic bounded continuous dataset in Table 4.5 of $D = 100K$, $N = 200$ and true $M = 3$ with mixing weights [0.6,0.2,0.2] that corresponds to VB-DPBM clustering accuracy 80%. The true weights are also given.

Regarding the discussion of the rest of the results in Table 4.5, K-means has the less efficient performance in $D = 10$ and $D = 10K$ with accuracy 74.5% and 76.5% respectively. However, for the middle datasets of $D = 100$ and $D =$ 1K it performs similarly to Hierarchical clustering. With respect to the highest feature size $D =$ 100K, the four algorithms seem to differentiate only little with accuracies in the range of 70 – 82%. Nevertheless, the best performance is achieved by DBSCAN with accuracy 82% and immediately follows VB-DPBM with 80%. The rest two methods attain values between 70 and 79%. Regarding VB-DPBM in the $D =$ 100K, we notice that it determines two instead of three clusters, thus we inspect in Table 4.6 the variational allocation that returned less number of components. In this table, we display the three true mixing weights of the $D =$ 100K and $N = 200$ synthetic dataset of Table 4.5, as well as the corresponding VB-DPBM estimated weights. VB-DPBM correctly estimates the component with 0.6 weight, while merges into one the two components with weight 0.2, returning a mixing coefficient of 0.4. Consequently, VB-DPBM mis-clusters 20% of the datapoints, confirming the accuracy level of 80% it reaches for this dataset.

4.4 Discrete Synthetic Data

For the discrete simulations, we generate count data from Poisson mixtures and binary data from Binomial mixtures when the number of trials is one. Both mixture simulations assume independence across the features and across the samples as in the synthetic Beta mixture data in Section 4.3. Specifically, we follow the same simulation technique as in Section 4.3, by randomly generating the model parameters. In particular, for the count mixture datasets, we simulate component specific Poisson data, with the Poisson parameter λ_{dm} being generated from a uniform with hyperparameters (10,20). λ_{dm} is the parameter of the d^{th} feature in the m^{th} component (see model structure in Appendix B, Section B.2.1). Regarding the binary mixture datasets, we simulate component specific Bernoulli data, with the probability parameter p_{dm} being generated from a uniform (0.01, 0.99) so as to ensure it lives between $[0, 1]$ (see model definition in Chapter 3, Section 3.3.1). The number of features and samples, and the mixing weights values are function's inputs.

4.4.1 Clustering Count Data

Count data is a type of data in which the observations are non-negative integers $\{0, 1, 2, 3, 4, ...\}$ and represent occurrences of a particular characteristic, *i.e.*, number of methylated CpG sites (counts) in a DNA region like DMR (Differentially Methylated Region). When the counts are accompanied by the number of independent trials and the aim is to find the hidden clusters structure, the Dirichlet Process Binomial mixture model can be applied. In cases where the numbers of trials are not independent and unknown, Dirichlet Process Poisson mixture could be a better clustering choice. In this count simulation study, we implement the variational Dirichlet Process Poisson mixture (VB-DPPM) on count data with non-fixed trials. This hierarchical model also takes into account the overdispersion of the data due to the Gamma prior on the Poisson parameter $\boldsymbol{\lambda}$ (model structure in Appendix B, Section B.2.1). In our applications, we fix the overdispersion by choosing the hyperparameters of this Gamma prior. To fully account for the randomness of the over-dispersion, the hierarchical Dirichlet Process Negative Binomial mixture model with random overdispersion should be employed instead (see details in Miao et al. [95]).

Hyperparameters and Initialization

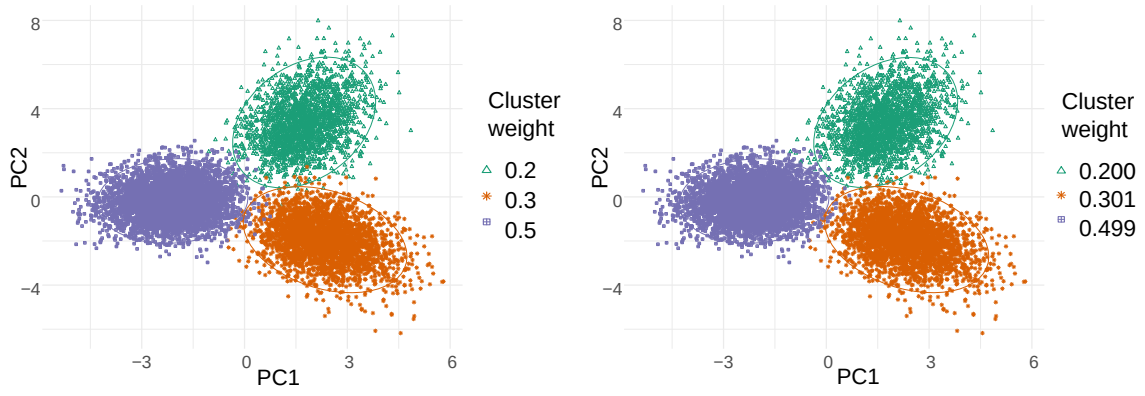
Prior to the VB-DPPM application on the simulated counts, we need to set the hyperparameters of the hierarchical Dirichlet Process Poisson mixture model. The hyperparameter vector $\boldsymbol{\phi}_0 = [\phi_{01}, ..., \phi_{0M}]$, which is related to the stick-breaking point $\boldsymbol{w}$ and therefore the mixing weights $\boldsymbol{\pi}$, is similarly fixed as in the Dirichlet Process

Beta mixture by a low value (each element of the $\boldsymbol{\phi}_0$ vector is set as 0.01) to promote the support from the data. The hyperparameters (a_{0dm}, b_{0dm}) of the Gamma prior on λ_{dm}, where λ_{dm} is the parameter of the d^{th} feature in the m^{th} Poisson component, are set to 1 to avoid extremely large values of λ_{dm}.

In regard to the initialization, because of the annealing addition the performance of the VB-DPPM algorithm is thinly affected by any selection of initial values. Hence, and to avoid extra complexity, we initialize the variational parameters with the corresponding hyperparameters. With respect to the stick-breaking point variational parameter of the m^{th} component ϕ_m, this is initialized by the corresponding hyperparameter ϕ_{0m} by adding also up a small, yet component specific value, facilitating the distinction of the clusters. For instance, ϕ_m is initialized by $\phi_{0m} + \epsilon_m$, with $\epsilon_m < 0.001$.

Simulation Scenario

In this part, we challenge the power of VB-DPPM by applying it on a synthetic large count dataset of N = 10K samples, D = 100 features and M = 3 clusters with mixing weights: $[0.2, 0.3, 0.5]$. In Figure 4.5, the true clustering (Subfigure 4.5a) and the fitted VB-DPPM clustering (Subfigure 4.5b) is displayed in two principal components. The visual comparison shows the accurate performance of VB-DPPM by correctly revealing three clusters with estimated mixing weights $[0.200, 0.301, 0.499]$ (see weights in Subfigure 4.5b).



(a) True clusters (b) VB-DPPM clusters

Figure 4.5 Comparison of true density plot with VB-DPPM density plot in two-dimensions (Principal Component analysis is used to present the clustering in two dimensions) of a count simulated dataset of $N = 10K$, $D = 100$ and true $M = 3$ with mixing weights $[0.2, 0.3, 0.5]$. Clusters are: (a) the true ones and (b) the VB-DPPM ones. The cluster weight of each group is given on the right, along with the corresponding colour and datapoint symbol.

4.4.2 Comparison to Standard Methods

	Correctly clustered %			
Samples	**VB-DPPM**	**K-means**	**Hierarchical**	**DBSCAN**
$\boldsymbol{N}$**=1M**	94.12 (7)	77.25 (5)	-	77.50 (5)
$\boldsymbol{N}$**=100K**	93.81 (6)	77.05 (5)	-	75.80 (5)
$\boldsymbol{N}$**=10K**	99.98 (7)	85 (5)	72.50 (3)	57.77 (4)
$\boldsymbol{N}$**=1K**	100 (7)	100 (7)	100 (7)	68.10 (5)

Table 4.7 Clustering performance of VB-DPPM, K-means, Hierarchical clustering and DBSCAN on four count synthetic data of escalating sample size (N = $1K$ to N = $1M$), fixed number of features D = 100 and components M = 7 with mixing weights $[0.05, 0.1, 0.115, 0.135, 0.15, 0.175, 0.275]$. The rates represent the percentage of correctly clustered observations and the values inside the parentheses the determined number of components (except K-means and Hierarchical where the number of clusters is fixed by the Elbow method). Dash line denotes algorithm's inability to scale in such large sample scenarios.

After the graphical comparison with the ground truth, we provide extra simulation tests with the rates of correctly clustered datapoints *via* the VB-DPPM algorithm, alongside the performance of K-means, Hierarchical clustering and DBSCAN. Similar simulation frameworks are created and evaluated as in the VB-DPBM applications in Section 4.3.3. Specifically, the first set of scenarios in Table 4.7 refers to synthetic count datasets with M = 7, D = 100 and increasing number of samples (N = 1K to N = 1M), whilst the second in Table 4.8 to small sample-wise datasets (N = 200) with M = 3 and increasing number of features (D = 10 to D = 100K).

The evaluation of the clustering methods in Table 4.7 conveys dominance of the VB-DPPM algorithm in all the different sample sizes, with the accuracy level being higher than 93%. K-means performs efficiently in the two lower datasets (N = 1K and N = 10K), whereas its accuracy lowers down to approximately 77% in the larger scenarios. Hierarchical clustering is faultless in N = 1K, however in N = 10K has poorer performance with rate 72.5%. In the large-scale scenarios, Hierarchical cannot scale and thus produces no results (dash lines). Concerning DBSCAN, it generally presents inferior performance on these synthetic datasets compared to the rest of the methods, by even reaching the low level of 57.77% on the N = 10K case.

Regarding the clustering evaluation in multiple feature scenarios in Table 4.8, VB-DPPM is once again the winning algorithm. It specifically reaches 100% in the high-feature datasets, while it drops to 80% in D = 10 with five estimated clusters instead of three. However, K-means, Hierarchical and DBSCAN perform less efficiently than VB-DPPM on the same small synthetic dataset, with the highest accuracy amongst three to be obtained by K-means (78%).

	Correctly clustered %			
Features	**VB-DPPM**	**K-means**	**Hierarchical**	**DBSCAN**
***D*=100K**	100 (3)	100 (3)	100 (3)	99.50 (3)
***D*=10K**	100 (3)	100 (3)	100 (3)	99 (3)
***D*=1K**	100 (3)	70 (5)	98.50 (5)	96 (3)
***D*=100**	97.50 (4)	77 (4)	98 (4)	78 (2)
***D*=10**	80 (5)	78 (4)	73.50 (4)	70 (2)

Table 4.8 Clustering performance of VB-DPPM, K-means, Hierarchical clustering and DBSCAN on count synthetic data of escalating feature dimensions (D = 10 to D = 100K), fixed sample N = 200 and components M = 3 with mixing weights $[0.6, 0.2, 0.2]$. The rates correspond to the accuracy of the algorithm in correctly clustering the simulated datapoints and the values inside the parentheses to the determined number of components (except K-means and Hierarchical where the number of clusters is fixed by the Elbow method).

Counts with Confounding Parameters

In the previous simulated scenarios, the datasets are exempted from confounding parameters, such as sex, age, ethnicity etc., that may affect the clustering credibility. For that reason, we arbitrarily choose to generate a set of synthetic scenarios of counts, with two confounding parameters L = 2, D = 1K features and varying sample sizes N = 100 to N = 100K (Table 4.9) in order to assess the clustering performance of the VB-DPPM with covariates (Chapter 3, Section 3.3.3) and the VB-DPPM without covariates. The aim is to show the superiority of VB-DPPM with covariates over the plain VB-DPPM when confounding factors exist.

	Correctly clustered %	
Features	**VB-DPPM with covariates**	**VB-DPPM**
***N*=100K**	85.80 (4)	42 (5)
***N*=10K**	85.20 (4)	40 (5)
***N*=1K**	84.70 (4)	60.20 (4)
***N*=100**	73 (2)	70 (2)

Table 4.9 Clustering performance of VB-DPPM with covariates and VB-DPPM without covariates, on count simulations where confounding parameters exist. The performance is tracked for increasing sample sizes, with fixed features D = 1K, number of components M = 3 with mixing weights $[0.2, 0.3, 0.5]$ and number of confounding parameters L = 2.

In Table 4.9, the VB-DPPM with covariates performs sufficiently well in clustering the large-scale count datasets (N = 1K to N = 100K) by reaching accuracy between 84% and 86%. Regarding the mis-clustered 14 – 16%, this is due to the extra fourth

component that VB-DPPM estimates, while the true number is three. In contrast, the VB-DPPM without covariates shows poor performance in the same scenarios, with accuracy levels between 40 – 60% and estimated number of components four or five, instead of three. However, both methods perform similarly on the simulated dataset of low number of samples $N = 100$, with VB-DPPM with covariates reaching 73% and VB-DPPM without 70%. In conclusion, this simulation study shows that VB-DPPM with covariates needs to be preferred over the plain VB-DPPM when confounding factors are present.

4.4.3 Clustering Binary Data

Binary are the data that can take on only two possible states. An appropriate distribution to model such type of data is Bernoulli, a sub-category of the Binomial distribution when the number of trials is one. In this analysis, our interest lies on uncovering hidden clusters based on binary variables and therefore, a suitable probabilistic method is the Dirichlet Process mixture of Bernoulli distributions. This hierarchical model is also known as Dirichlet Process Beta-Binomial mixture (with fixed number of trials at one) due to the Beta priors imposed upon the Bernoulli parameter matrix $\boldsymbol{p}$ of dimensions $M \times D$, with p_{dm} referring to the probability of success of the d^{th} feature in the m^{th} component (model hierarchy in Chapter 3, Subsection 3.3.1). The randomly drawn p_{dm} from a Beta distribution provides conjugate Bayesian inference and also contributes to capturing the overdispersion in the data.

Hyperparameters and Initialization

The inferential procedure for the Dirichlet Process Bernoulli mixture involves Variational Inference and therefore, the VB-DPBerM is applied (model in Chapter 3, Subsection 3.3.1). The hyperparameters of the prior distributions are specified as well as the variational parameters. In particular, the Beta prior on p_{dm} has hyperparameters a_{0dm} and b_{0dm}, which are both given the value 1 (uniform distribution) so as to express equally favourable values for p_{dm}. For the hyperparameter vector $\boldsymbol{\phi_0} = [\phi_{01}, ..., \phi_{0M}]$ of the stick-breaking point $\boldsymbol{w}$ prior, we set a low value 0.01, as in the VB-DPBM and VB-DPPM, in order to assign low weight on the initial clusters and let the data to gradually determine the mixing coefficients.

Regarding the variational parameters initialization, we straightforwardly give initial values equal to the hyperparameters. For example, a_{dm} is initialized by a_{0dm} and b_{dm} by b_{0dm}. With regards to $\boldsymbol{\phi} = [\phi_1, ..., \phi_M]$, each element is initialized by the corresponding ϕ_{0_m} with the addition of the low component specific value $\epsilon_m < 0.001$ to help the distinction of the clusters.

4.4.4 Comparison to Standard Methods

Having already seen the process of visually evaluating the clustering performance of the variational Dirichlet Process mixtures, we move directly to quantitative comparisons of VB-DPBerM with standard non-probabilistic methods. Similar simulation scenarios to VB-DPBM and VB-DPPM are constructed in order to assess VB-DPBerM in multiple cases. To recap the scenarios structure, we firstly simulate binary datasets of increasing sample size (N = 1K to N = 1M), with fixed number of features D = 100 and number of components equal to M = 7 with mixing weights $[0.05, 0.1, 0.115, 0.135, 0.15, 0.175, 0.275]$ (Table 4.10). The second set of scenarios includes synthetic binary datasets of growing feature size (D = 10 to D = 100K), fixed sample size N = 200 and three components M = 3 whose mixing weights are $[0.6, 0.2, 0.2]$ (Table 4.11).

	Correctly clustered %			
Samples	**VB-DPBerM**	**K-means**	**Hierarchical**	**DBSCAN**
***N*=1M**	96 (7)	88 (8)	-	90.86 (7)
***N*=100K**	96 (7)	86.67 (8)	-	85.50 (8)
***N*=10K**	95 (6)	98 (7)	98 (7)	93 (7)
***N*=1K**	95 (6)	83.70 (6)	97 (7)	90 (6)

Table 4.10 Clustering performance of VB-DPBerM, K-means, Hierarchical clustering and DBSCAN on binary synthetic data of varying sample size (N = $1K$ to N = $1M$), fixed number of features D = 100 and components M = 7 with mixing weights $[0.05, 0.1, 0.115, 0.135, 0.15, 0.175, 0.275]$. The values represent the percentage of correctly clustered observations. Dash line denotes algorithm's inability to scale in such large sample scenarios.

In Table 4.10, we observe that VB-DPBerM is overall the most consistent clustering algorithm for the synthetic binary scenarios with seven clusters, by successfully obtaining accuracy of 95% in the lower sample sizes: N = 1K, N = 10K and 96% in the high-scaled ones: N = 100K and N = 1M. Hierarchical clustering is accurate at N = 1K and N = 10K with clustering performance at 97 – 98%, however, as before, it cannot scale at the higher sample sizes and thus no results are returned. DBSCAN and K-means have on average close performance that varies between 84 – 91%. In regard to the small sample sizes and escalating feature dimensions in Table 4.11, VB-DPBerM is undoubtedly the most suitable algorithm, given those synthetic binary datasets, with accuracy level at 100% in all scenarios, except the small dataset of D = 10 where it lightly drops to 97%. Hierarchical clustering and DBSCAN present similar clustering performance with accuracy levels between 80 and 90% in the middle scenarios D = 100 and D = 10K, while at D = 10 and D = 100K both reach higher than 90% rates. K-means is the clustering algorithm with the most unstable performance, since it reaches higher than

98% in the external datasets of low and high number of features (D = 10 and D = 100K), whilst in the middle cases the accuracy level fluctuates between 66 and 72%.

	Correctly clustered %			
Features	**VB-DBerPM**	**K-means**	**Hierarchical**	**DBSCAN**
***D*=100K**	100 (3)	100 (3)	100 (3)	98.42 (3)
***D*=10K**	100 (3)	72 (5)	83 (2)	82 (4)
***D*=1K**	100 (3)	66 (5)	85 (4)	84.2 (4)
***D*=100**	100 (3)	69 (5)	82 (4)	80.3 (4)
***D*=10**	97 (3)	98 (2)	95 (4)	90 (4)

Table 4.11 Clustering performance of VB-DPBerM, K-means, Hierarchical clustering and DBSCAN on binary synthetic data of varying dimensions (D = 10 to D = 100K), fixed sample size N = 200 and components M = 3 with mixing weights $[0.6, 0.2, 0.2]$. The rates correspond to the accuracy of the algorithm in successfully clustering the simulated datapoints.

4.5 Further Simulation Analysis

In the previous sections, we created individual simulation scenarios of various data structures (small and high sample and feature sizes) to assess the clustering performance of the probabilistic variational Dirichlet Process mixtures. In this section, we are interested in evaluating the performance of VB-DPBM, VB-DPPM and VB-DPBerM in synthetic bounded continuous, count and binary data respectively, with specific dimensions that resemble the dataset sizes of our real applications in Chapter 5. The reason is to further evaluate those probabilistic methods in similar scenarios to the real datasets, aiming at increasing the credibility of modeling DNA methylation data by variational Dirichlet Process mixtures. More precisely, we simulate 20 synthetic datasets from each data type (bounded continuous, counts, binary) of N = 200 samples, D = 40 features and M = 4 clusters with mixing weights $[0.3, 0.3, 0.3, 0.1]$. We then apply the appropriate variational method, VB-DPBM, VB-DPPM or VB-DPBerM, as well K-means, Hierarchical clustering and DBSCAN and track the accuracy rate. Eventually, we report in Table 4.12 the average value based on the 20 replicates, along with the corresponding standard deviation.

In Table 4.12, we compare column wise the clustering rates. In the bounded continuous simulations, VB-DPBM performs accurately by clustering correctly 100% of the datapoints in all the 20 replicates. Second in row comes DBSCAN with average accuracy 95.8% and 5.56 standard deviation, then K-means with 93.7% and 6.02 standard deviation and Hierarchical with 83.1% and the considerably high standard deviation of 10 units. In the counts simulations, VB-DPPM reaches again the highest performance of 95.3% with the lowest deviation of 5.60, while second arrives K-means with 93.5%

and 5.82 standard deviation. Hierarchical and DBSCAN fall behind with accuracy around 72%, which is a result of high fluctuations of the accuracy rates (standard deviation of 10 to 12 units). Regarding the binary simulations, VB-DPBerM has on average the highest accuracy rate of 91.2% with K-means diverging by only 1.2%, while the variability is relatively similar (around 5.40 – 5.50). Hierarchical obtains 85.2% average clustering performance and DBSCAN 80.6%, with their standard deviation reaching high levels (between 10 and 11).

	Correctly clustered %		
Methods	**Bounded continuous**	**Counts**	**Binary**
VB-DPBM	100 (0)		
VB-DPPM		95.30 (5.60)	
VB-DPBerM			91.20 (5.50)
K-means	93.70 (6.02)	93.50 (5.82)	90 (5.44)
Hierarchical	83.10 (10)	71.35 (11.37)	85.20 (10.78)
DBSCAN	95.80 (5.56)	72 (10)	80.60 (11.02)

Table 4.12 Average clustering performance of VB-DPBM, VB-DPPM, VB-DPBerM, K-means, Hierarchical clustering and DBSCAN based on 20 simulations in each data type category: bounded continuous, counts and binary. All the synthetic scenarios concern $N = 200$ samples, $D = 40$ features and number of components $M = 4$ with mixing weights $[0.3, 0.3, 0.3, 0.1]$. The values correspond to the mean accuracy of each algorithm in clustering the corresponding data type. The value inside the parenthesis is the standard deviation.

In conclusion, based on these simulations, we have evidence that the variational Dirichlet Process mixtures are appropriate candidates for the DNA methylation applications in Chapter 5, given their successful performance in synthetic datasets with dimensions similar to those of the real datasets.

4.6 *A posteriori* Feature Selection

Having applied the variational Dirichlet Process mixture models on simulated scenarios and tested their clustering performance, we proceed with the final step that concerns the selection of the discriminative features *per* component. For this selection, we require the fitted component distributions so as to apply the discriminative measure from Chapter 2, equation (2.78).

For illustrative purposes and to ease the understanding around this *a posteriori* selection step, we simulate a new bounded continuous dataset in Table 4.13 of $N = 1$K samples, three clusters, $M = 3$, with mixing weights $[0.6, 0.2, 0.2]$ and only $D = 3$ features. The aim is to implement the VB-DPBM model and obtain the discriminative set of features

for each of the three components. The fitted component distributions are utilized to calculate the feature selection measure $A_m(h)$ in equation (2.78). We recall that m is the index of the m^{th}component and h the set of features used to compute the measure. For each component, we record the value of $A_m(h)$ for all the feature combinations, intending to choose the set h which corresponds to the highest value.

	Discriminative accuracy						
Features combinations →	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
Component 1	0.749	0.795	0.757	0.825	0.917	0.804	1
Component 2	0.999	0.998	0.998	0.999	0.994	0.999	1
Component 3	0.658	0.918	0.827	0.851	0.870	0.975	1

Table 4.13 Feature selection *per* component after the implementation of VB-DPBM on a synthetic dataset of $N = 1K, D = 3$ and $M = 3$ with mixing weights $[0.6, 0.2, 0.2]$. The discriminative measure $A_m(h)$ is calculated for each feature combination within the cluster.

In Table 4.13, the ultimate discriminative accuracy values (100%) are attained in the full dimensions $\{1, 2, 3\}$ for all three components. Nonetheless, we look whether the next in sequence accuracy values are high enough. For component 1, the next high accuracy is achieved by the set $\{1, 3\}$. In component 2, all the sets of features are individually important to discriminate this component from the rest, hence we select the smallest in size set which also bears the highest value. This is feature $\{1\}$ with discriminative accuracy 0.999. Regarding component 3, the next highest accuracy is reached by the set of $\{2, 3\}$.

To summarize, feature $\{1\}$ seems to discriminate component 2 *per se* and also component 1 in conjunction with feature $\{3\}$. As for component 3, this can be discriminated by the information provided by $\{2, 3\}$. In this toy example, we found that each component had a different set of discriminative features, however, the joint set includes all three dimensions resulting in no feature reduction.

The challenging part though arises when the number of features is considerably high and therefore, selection is required to reduce the feature space complexity. For this framework, we generate a synthetic binary dataset of sample size N = 1K, M = 3 with mixing weights $[0.6, 0.2, 0.2]$ and D = 1K. We then apply the VB-DPBerM algorithm which successfully determines three clusters with accuracy 97%. The reason we choose to work on binary scenarios for this illustrative application is because of the sparse structure the binary data usually suffer from, *i.e.*, numerous dimensions with mostly zero values and hence negligible contribution.

Regarding the previous three-dimensional toy example, it was feasible to enumerate all the possible combinations of feature dimensions ($2^3 - 1 = 7$). For the 10K-feature data in our new simulated study, this number is unattainable and thus, the forward

method is recruited as discussed in Chapter 2, Section 2.7. Table 4.14 demonstrates the iterative forward selection scheme of features in the synthetic binary dataset. Each component column contains the discriminative accuracy $A_m(h)$ values *per* iteration. The convergence value (current value differs from previous by less than 10^{-3} units) is framed in grey colour and corresponds to the selected set of features. Component 1 is defined by nine features and components 2 and 3 by 10 each one. Overall, the total number of unique features is 26, which corresponds to 97.4% feature reduction, leaving space to only the informative ones.

	Discriminative accuracy		
Iteration	**Component 1**	**Component 2**	**Component 3**
1	0.706	0.689	0.544
2	0.779	0.778	0.584
3	0.807	0.818	0.615
4	0.831	0.843	0.641
5	0.892	0.923	0.684
6	0.966	0.951	0.817
7	0.976	0.960	0.893
8	0.981	0.986	0.960
9	0.994	0.986	0.996
10	0.994	0.993	0.995
11	-	0.993	0.995
Number of important features per component	9	10	10
Number of important features in total	26 out of 1K		

Table 4.14 Forward selection of features *per* component based on the discriminative measure $A_m(h)$. The coloured boxes denote the convergence value of the measure. The selected number of features for each component, as well as the total important features, are given in the end. The data concern a binary simulated dataset of $N = 200, D = 1K$ and $M = 3$ with mixing weights $[0.6, 0.2, 0.2]$, modelled by VB-DPBerM.

Thereafter, we graphically assess the clustering performance of the significantly reduced binary dataset ($D_{\text{sel}} = 26$ features). Our hope is to retrieve the true number of clusters ($M = 3$) and accurate mixing estimates in relation to the ground truth (weights: $[0.2, 0.3, 0.5]$). Figure 4.6 shows the clustering performance of VB-DPBerM in the synthetic dataset of $N = 1$K, $M = 3$ and $D_{\text{new}} = 26$. For reasons of graphical representation, we illustrate the two logistic principal components. Logistic Principal Component Analysis is a dimensionality reduction tool for multivariate binary data (Lee et al. [72]). The figure communicates the achievement of VB-DPBerM in correctly clustering the datapoints into three clusters of mixing weights $0.19, 0.29$ and 0.52, by

exploiting only 26 out of the 1K features. The clustering accuracy reaches the 98% level displaying effective feature selection by the $A_m(h)$ measure.

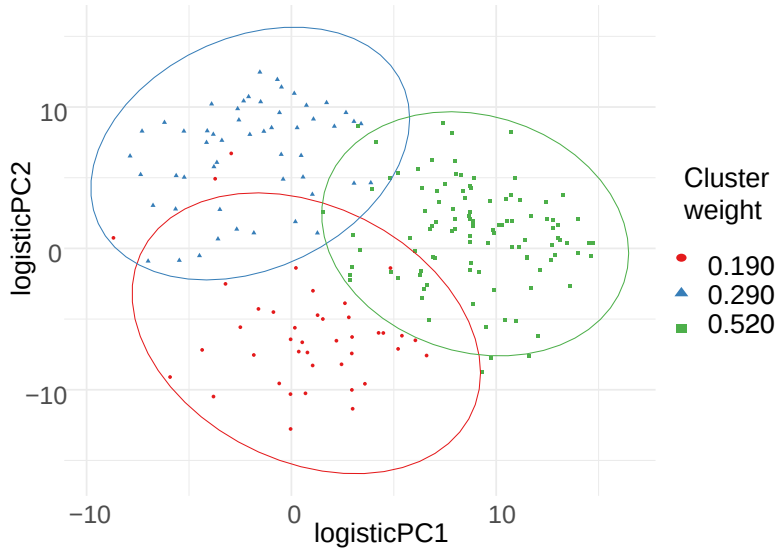


Figure 4.6 Clustering performance of VB-DPBerM on a synthetic binary dataset of $M = 3$ clusters with mixing weights $[0.2, 0.3, 0.5]$, $N = 200$ and $D_{\text{sel}} = 26$ features selected by the discriminative measure $A_m(h)$ (the original D was equal to 1K). For reasons of graphical representation in two dimensions, the logistic Principal Component Analysis is employed. Each cluster bears a distinct colour and shape point. The estimated mixing weights are given on the right hand side of the graph.

4.7 Summary

In this chapter, we assessed the clustering power of the continuous and discrete variational mixture models in synthetic datasets. In particular, we evaluated the clustering performance of the variational Dirichlet Process Beta mixture, the variational Dirichlet Process mixture of Poisson distributions and the variational Dirichlet Process Bernoulli mixture. The VB-DPBM, VB-DPPM and VB-DPBerM easily scaled on large datasets, both sample and feature wise, providing speedy and notably accurate results in scenarios of $N >> D$ and $D >> N$. Clustering comparisons to the non-probabilistic K-means, Hierarchical cluster analysis and DBSCAN boosted the confidence for preferring the variational Dirichlet Process mixture models for clustering bounded continuous data as well as counts and binary data. Regarding the occurrence of confounding parameters, a simulation study clearly recommended using methods which take those factors into account such as the variational Dirichlet Process Poisson mixture model with covariates. Finally, we showed that selection of the discriminative features *per* component *via* the $A_m(h)$ measure offered a practical way in *a posteriori* revealing the salient features for each component and overall reducing the dimensions without loss of information.

Chapter 5

Analysis of DNA Methylation Data

5.1 Overview

This chapter is focused on the analysis of real molecular datasets related to DNA methylation and specifically, Cytosine methylation. In particular, we seek to unearth the hidden clusters of individuals based on their methylation profile, as well as to specify those DNA regions (here DMRs) whose degree of methylation is responsible for discriminating each cluster from the rest of the groups. The general aim is to identify clinically relevant subgroups for the early prognosis or diagnosis of diseases.

Regarding feature discrimination, several methods have been developed to do selection of features that are important for all the subgroups simultaneously, such as in Tadesse et al. [130] and in Kim et al. [68]. In the former paper, clustering is achieved through a Bayesian Finite multivariate Gaussian mixture model which is inferred by the Reversible Jump Markov chain Monte Carlo, while the latter does the clustering *via* a Bayesian Dirichlet Process multivariate Gaussian mixture model inferred by the split-merge Markov Chain Monte Carlo algorithm of Jain and Neal [64]. In both papers, the variable selection is accomplished with the addition of a latent binary variable that indicates which features contribute or not in the overall group structure. On the other hand, Raftery and Dean [115] use Finite mixture models for the clustering part while they do variable selection by comparing in pairs models of nested subsets of features through approximate Bayes Factors. Despite the usefulness of these works in selecting important features, they lack feature discrimination *per* cluster. Consequently, we decide to use the discriminative measure of Lin et al. [76] (Chapter 2, Section 2.7) to determine whether there are features and specifically methylated DNA regions (iDMRs) that may be significant for specific clusters and not necessarily for all. Particularly, the algorithm provides three types of information: 1) features that are not important for

any cluster, 2) features that are important for all and 3) features that are important for some but not for the other clusters.

For the purpose of this chapter, we analyze blood samples of artificially and naturally conceived neonates recorded with the Beckwith-Wiedemann syndrome (congenital disorder related to overgrowth, Weksberg et al. [147]) and without the syndrome (control group). For each blood sample, the examined features are established imprinted Differentially Methylated Regions (iDMRs). Imprinted genes are those expressed only in one of the two parental chromosomes (Ferguson-Smith [44]), while DMRs are regions in the genome with different methylation patterns among multiple samples (patients, cells, tissues etc.) (Neidhart [103]). Therefore, in our analysis, looking at iDMR level rather than CpG works as a dimensionality reduction technique, which also promotes the relaxation of correlation between the methylation values at different iDMRs due to the aggregation of the correlated methylation values of sequential CpGs within an iDMR.

For the dataset of neonates (samples) and iDMRs (features), we derive three measures, each of which carries similar information but expressed differently. Concerning the first measure, the median of the methylation beta-intensities across the CpGs within an iDMR is calculated for each iDMR and each individual. In the second measure, for each CpG site, individuals with methylation level below or above the controls median methylation level ± 3 standard deviations (SDs) confidence interval are considered as 1 (significantly affected CpG) and those inside the 3SDs confidence interval are considered as 0 (non-significantly affected CpG). Then the number of significantly affected CpGs *per* iDMR is counted. Regarding the third measure, for each iDMR, individuals with a median methylation level below or above the controls median ± 3 standard deviations (SDs) confidence interval are considered as 1 (significantly affected iDMR) and those inside the 3SDs confidence interval are considered as 0 (non-significantly affected iDMR). We acknowledge the relatively low dimensionality of the dataset (228 neonates × 33 iDMRs) by reporting that this was the only available set of data at the moment of the study meeting the desirable requirements for the purpose of this analysis. Further implementations of the same models can be conducted in the future for larger available datasets of similar nature.

The goal in analyzing the same data of the same cohort but from three different perspectives is to discover which of these three measures is more informative in revealing the association of aberrantly methylated artificially conceived newborns with rare developmental disorders, as well as in indicating potential onset of another developmental disorder in the future, given the recorded methylation in certain iDMRs. For this reason, we start by analyzing each data type separately (median beta-intensities defined as “beta methylation data” from now on, number of affected CpGs *per* iDMR defined as

"count methylation data" and significantly/non-significantly modified iDMRs defined as "binary methylation data") to eventually end up to a consensus of the three analyses. In particular, we cluster each dataset based on the appropriate variational Dirichlet Process mixture model and then do iDMR selection by Lin et al. [76] to retain only the total discriminative iDMRs. Subsequently, we implement our variational models on the reduced iDMR datasets and derive the final clusters for each dataset. As a last step, we apply the discriminative measure again to perform iDMR discrimination *per* cluster.

5.2 Applications on DNA Methylation in Neonates

The current analysis is based on blood samples of 228 neonates, with 22 bearing a rare developmental disorder called Beckwith-Wiedemann Syndrome (BWS) - responsible for overgrowth such as macroglossia (Weksberg et al. [147]), while the rest neonates are recorded as BWS free. Moreover, an extra factor is considered regarding their way of conception. 78 neonates have been conceived naturally (20 of them have the disorder), whereas the remaining 150 (two of them with the disorder) are conceived artificially through Assisted Reproductive Technologies (ART) (Zegers-Hochschild et al. [153]). For each neonate, 33 of the most known imprinted differentially methylated regions (iDMRs) as defined in Monk et al. [98] and also reported in Ochoa et al. [106] are studied. The analysis of methylation at iDMRs is the standard methodology for clinical molecular analysis of congenital disorders worldwide (Ochoa et al. [106]). For each iDMR, the number of CpGs defining the region and their location in the genome can be found in Table 1 in Ochoa et al. [106]. The number of CpGs per iDMR varies between 5 and 76.

In order to perform a methylation profiling at imprinted regions in a cohort of newborns naturally conceived and newborns conceived by ART procedures, with and without Beckwith-Wiedemann Syndrome, we combined multiple public available datasets from methylation array platforms (450K and EPIC array). For these datasets, we analysed only those probes that overlap with imprinting regions and are common between the two platforms (984 probes). To avoid a batch effect caused by differences between platforms, each array-specific dataset was individually processed and then merged together. For validity reasons, we prove later that the platform/array is not a confounding parameter in the clustering procedure. After filtering by quality parameters, DNA methylation data is filtered by 33 iDMRs[1] (Ochoa et al. [106]).

The raw data are available in the GEO with accession number GSE166531 and GSE131433, corresponding to Ochoa et al. [106] and Novakovic et al. [105] respectively.

[1] iDMRs covered by at least five probes/CpGs.

Regarding the data pre-processing (flowchart in Figure 5.1), the IDAT data files from 450K and EPIC are analysed separately by the Bioconductor R package `ChAMP` (Chip analysis Methylation Pipeline by Tian et al. [136]), where the files are filtered by their intensity value and probes with detection p-value < 0.01 are removed. Further filtering removes probed CpGs that fall near an SNP, align to multiple locations with bwa or come from the X and Y chromosomes (Zhou et al. [157]). For the extracted beta-intensities, the NAs are discarded and the rest are imputed by the KNN (K-nearest neighbors algorithm). The imputed data are then normalised by the BMIQ filtering method (Teschendorff et al. [135]). In the normalised data the batch effect is predicted by the SVD method (Singular Value Decomposition) and the batch correction is performed by ComBat (Hansen et al. [56]). The obtained beta-intensities are then used to create the three previously discussed measures: a) median beta-intensity *per* iDMR[2] ("beta methylation data"), b) number of significantly affected CpGs *per* iDMR ("count methylation data") and c) significantly/non-significantly affected iDMR ("binary methylation data").

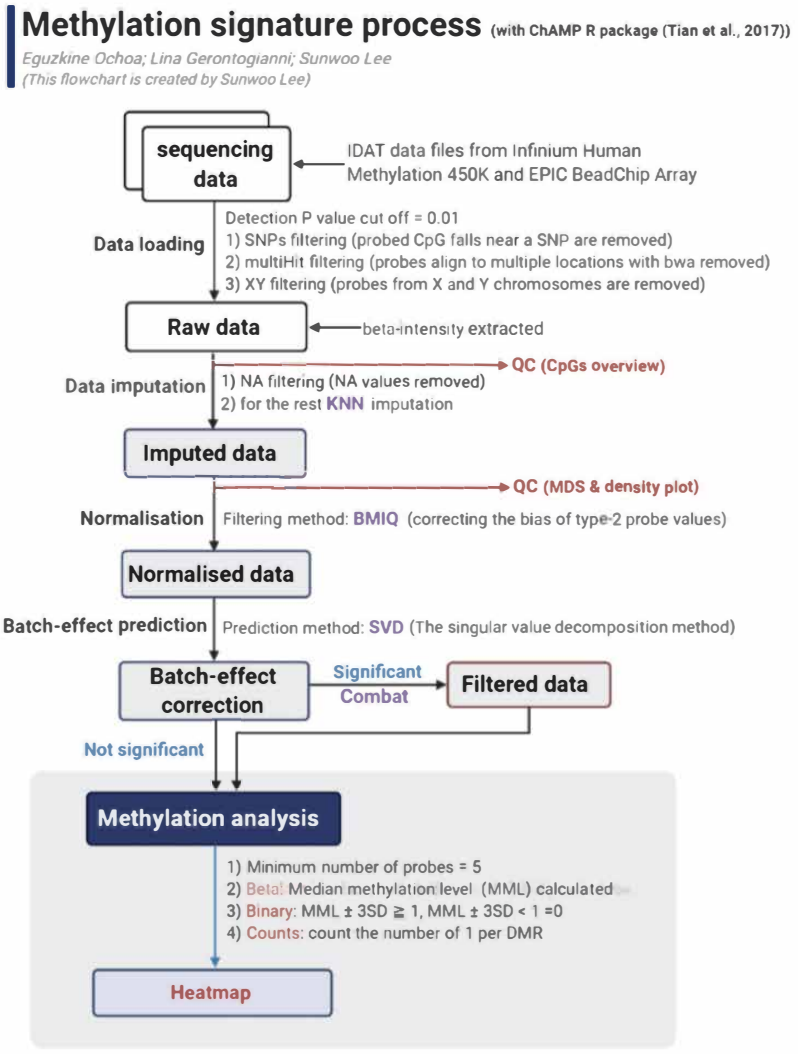


Figure 5.1 Data pre-processing flowchart of the real methylation data.

[2]Median of the CpG beta-intensities within an iDMR.

Regarding the choice of neonates for this analysis, testing methylation on newborns (fixed 0 age) and not children or adults of differing age results in fixing the confounding parameter *age*, as well as the *sex* factor given we do not have serious evidence (to our knowledge and according to the results in this due course analysis) that methylation at birth - at least in iDMRs - can be significantly affected by the newborn's gender. Generally, confounding factors have to be adjusted to prevent any chance of unavoidably clustering based on the grouping provided by them rather than the iDMR methylation level. For these real neonate data we are able to confidently apply our "variational Dirichlet Process mixture models without covariates" given we have already accounted for the sex and age covariates. Nonetheless, we could alternatively use the proper model with covariates specified in Chapter 3 in case of un-adjusted influencing parameters, *i.e.,* methylation data of individuals with different age.

For each one of the three measures ("beta methylation data", "count methylation data" and "binary methylation data") we choose not to apply the commonly used Gaussian mixtures due to their unrestricted support range. We conversely apply the appropriate variational Dirichlet Process mixture model based on the data type. Specifically, the "beta methylation data" are modelled by the variational Dirichlet Process Beta mixture (VB-DPBM), the "count methylation data" by the variational Dirichlet Process Poisson mixture (VB-DPPM) and the "binary methylation data" by the variational Dirichlet Process Bernoulli mixture (VB-DPBerM). At this point, we stress again that the three mixture models assume that methylation values between different iDMRs are independent. Generally, we are confident that claiming independence between methylation levels of different iDMRs is not a strict violation, considering that dependence is mostly present amongst the level of methylation in CpGs within an iDMR.

For clarification, in the subsequent analyses "Platform" is a categorical variable referring to the 450K and the EPIC arrays, "Sex" indicates the newborn's gender whilst "Status" is a categorical variable denoting the BWS cases and controls as follows

- case: neonate with BWS conceived naturally (20 in total)
- case-ART: neonate with BWS conceived through ART (2 in total)
- ctrl: neonate without BWS conceived naturally (58 in total)
- ctrl-ART: neonate without BWS conceived through ART (148 in total).

5.2.1 Beta Methylation Data

The first instance of data to be analyzed deals with the median beta-intensities *per* iDMR. The data are bounded in $[0, 1]$ and therefore, the VB-DPBM algorithm is employed for clustering. The structure of the analysis starts with the implementation

of the VB-DPBM on the dataset with all the iDMRs, then follows the selection of the total discriminative iDMRs and ends with the implementation of the VB-DPBM on the dataset with only the discriminative iDMRs. The reason we proceed with this two stages procedure is to reduce the noise in the data due to the participation of non-discriminative iDMRs, and form the final clusters based only on the important iDMRs. We clarify that each heatmap in this chapter, bears the same labelling for the clusters, *i.e.*, C1, C2 etc., but the clusters are dataset specific. For example, C1 in Figure 5.2 is different than C1 in Figure 5.4 and thus we let the caption of each figure or table to define the dataset that each specific group label corresponds to.
In Figure 5.2, the beta methylation data with all the iDMRs are clustered in four groups. Cluster C1 contains neonates with hypomethylation records (blue colour) in their KCNQ1OT1:TSSDMR. We observe that these neonates are the ones with the BWS disorder ("Status" column), showing that the algorithm successfully revealed the cases of BWS based solely on the methylation level of their iDMRs and especially KCNQ1OT1:TSSDMR. The rest three clusters, C2, C3 and C4, that contain newborns without the BWS, do not straightforwardly show differences in the methylation pattern.

Another observation is made with respect to the "Platform" variable. C1 contains only the 450K array data and C2, C3 the EPIC data, implying the possible existence of a platform effect that may lead to clustering based on the platform instead of the methylation level. To ensure that the "Platform" does not have an impact on the analysis, we remove it by clustering the residuals from the Beta regression model with fixed covariate the "Platform" variable (Ferrari and Cribari-Neto [45]) (we vectorize the beta methylation data and run the R `betareg` function by Zeileis et al. [154]). The extracted "sweighted" residuals, which are proved to be normally distributed (Espinheira et al. [39]), are free from the platform effect and thus are used to find the hidden clusters according to the methylation level. Specifically, on the Beta regression residuals with all the iDMRs we implement the variational Dirichlet Process Gaussian mixture algorithm (with independent features, Appendix B, Subsection B.3.1), defined as VB-DPGM for simplicity, and obtain the clusters in Figure 5.3 (note that the residuals are scaled for graphical reasons solely, such that 0 residual values correspond to 50% methylation and -4 and 4 residuals to 0% and 100% respectively). The clustered residuals heatmap shows great similarities with the one for the original beta methylation data with all the iDMRs (Figure 5.2). Both agree in clustering by 99.56% (rate of data points that have been identically assigned together in the two clusterings, described in Chapter 4, Section 4.3.3), determining four clusters with similar methylation patterns. This result hints against the platform effect. In order to validate this outcome, we check whether the coefficient of the "Platform" covariate in the Beta regression we applied before is statistically non-significant. Eventually, the platform coefficient is 0.008 and has p-adjusted value equal to $0.679 > 0.05$. This is strong evidence of non-

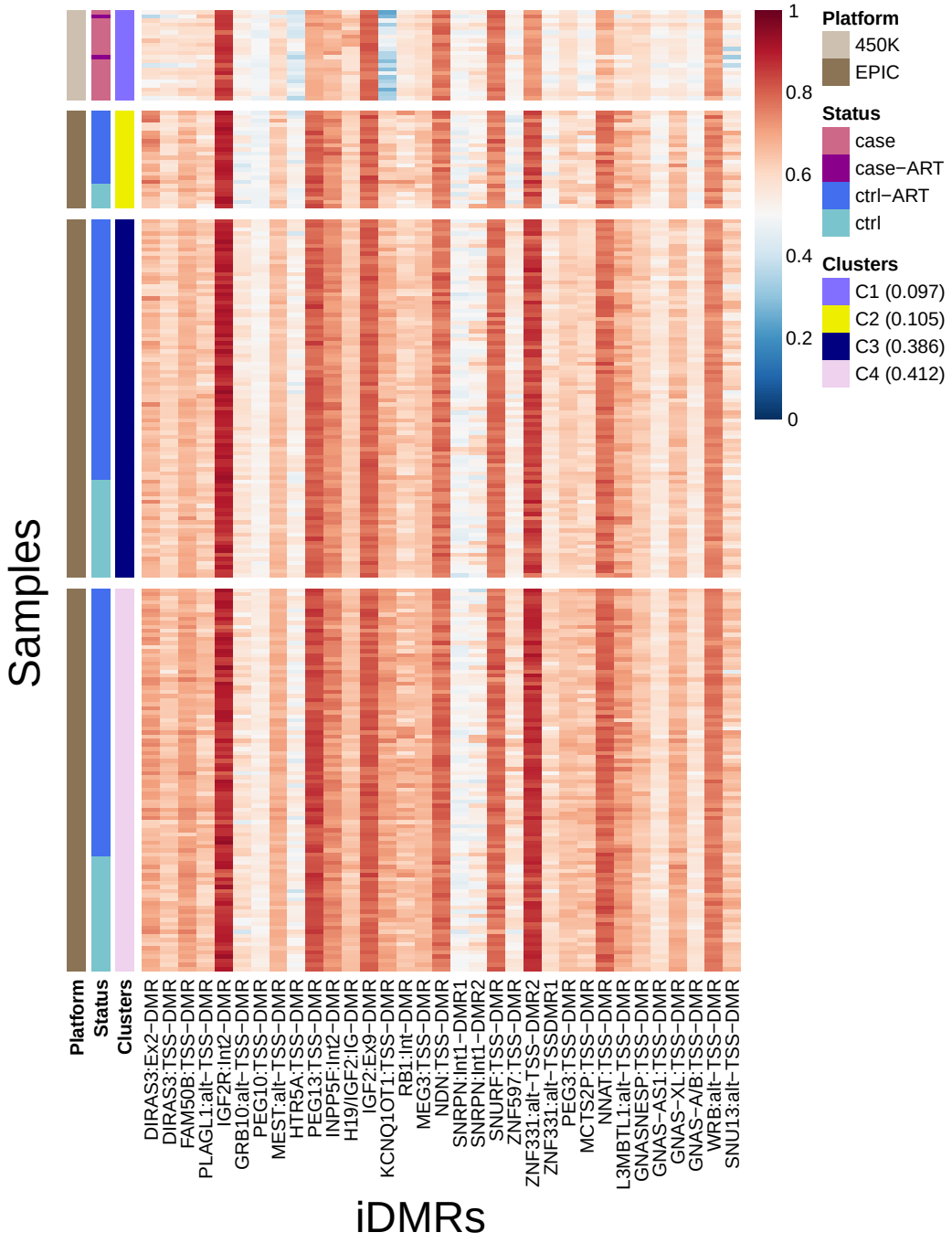


Figure 5.2 Clustered heatmap of the beta neonatal methylation intensities *via* the VB-DPBM. The x-axis represents the iDMRs (33 in total), while y-axis the samples (228 neonates). The colour scale of the beta-intensities starts from blue (0% methylation), continues to white (50% methylation) and ends up to red (100% methylation). On the left of the x-axis, the Clusters column shows the group in which the observations have been allocated to, in different colour (mixing weights are displayed on the right of the heatmap for each cluster). Status and Platform are also given for each cluster. The clusters are presented in an increasing mixing weight sequence (smallest cluster on top, largest at the bottom).

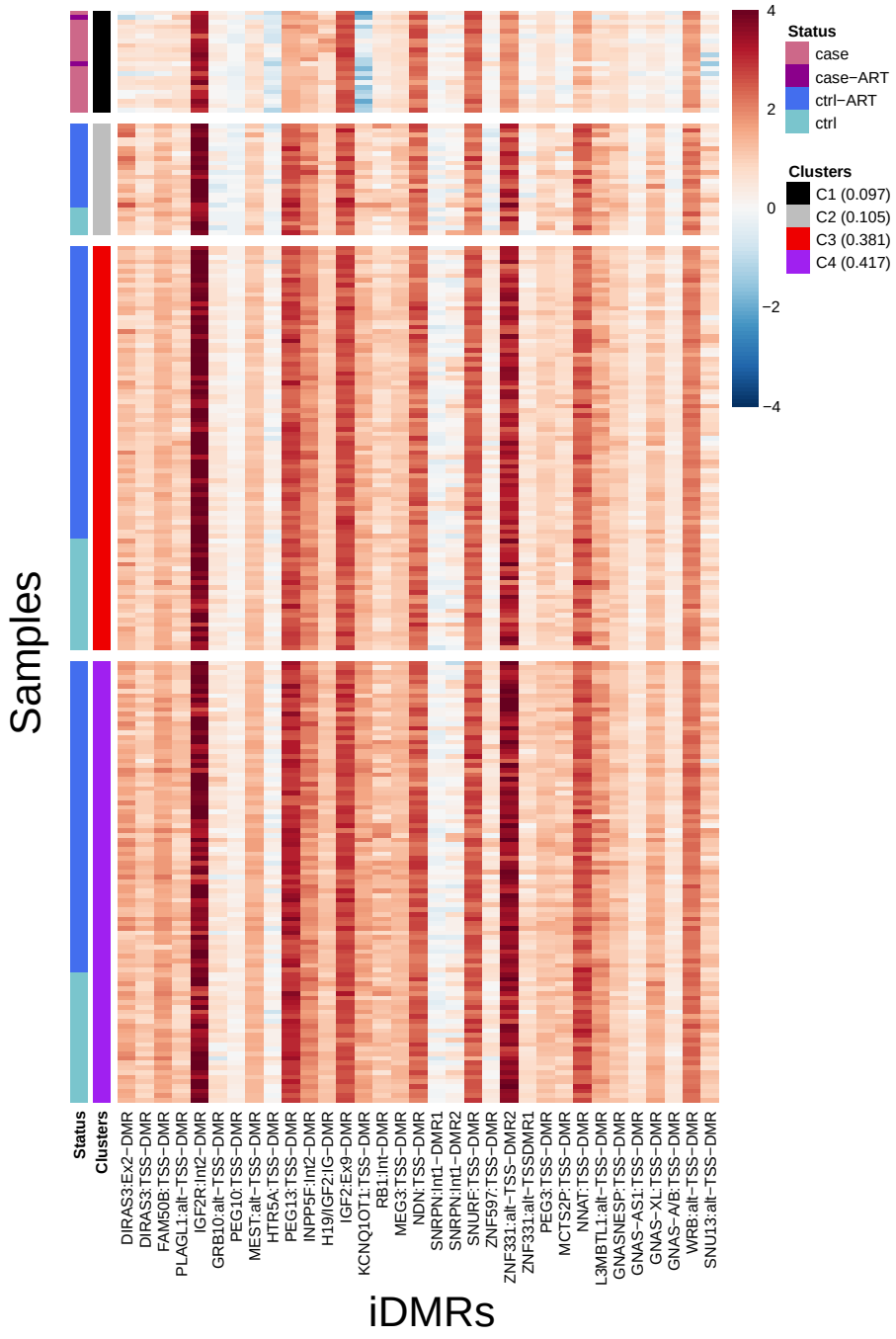


Figure 5.3 Clustered heatmap of the residuals from the Beta regression on the beta neonatal methylation intensities. The residuals are re-scaled for graphical reasons such that 0 values correspond to 50% methylation and $-4, 4$ to $0\%, 100\%$ methylation respectively. The clustering is achieved via the VB-DPGM. The x-axis represents the iDMRs (33 in total), while y-axis the samples (228 neonates). The colour scale of the residuals starts from blue (0% methylation), continues to white (50% methylation) and ends up to red (100% methylation). On the left of the x-axis, the Clusters column shows the group in which the observations have been allocated to, in different colour (mixing weights are displayed on the right of the heatmap for each cluster). Status is also given for each cluster. The clusters are presented in an increasing mixing weight sequence (smallest cluster on top, largest at the bottom).

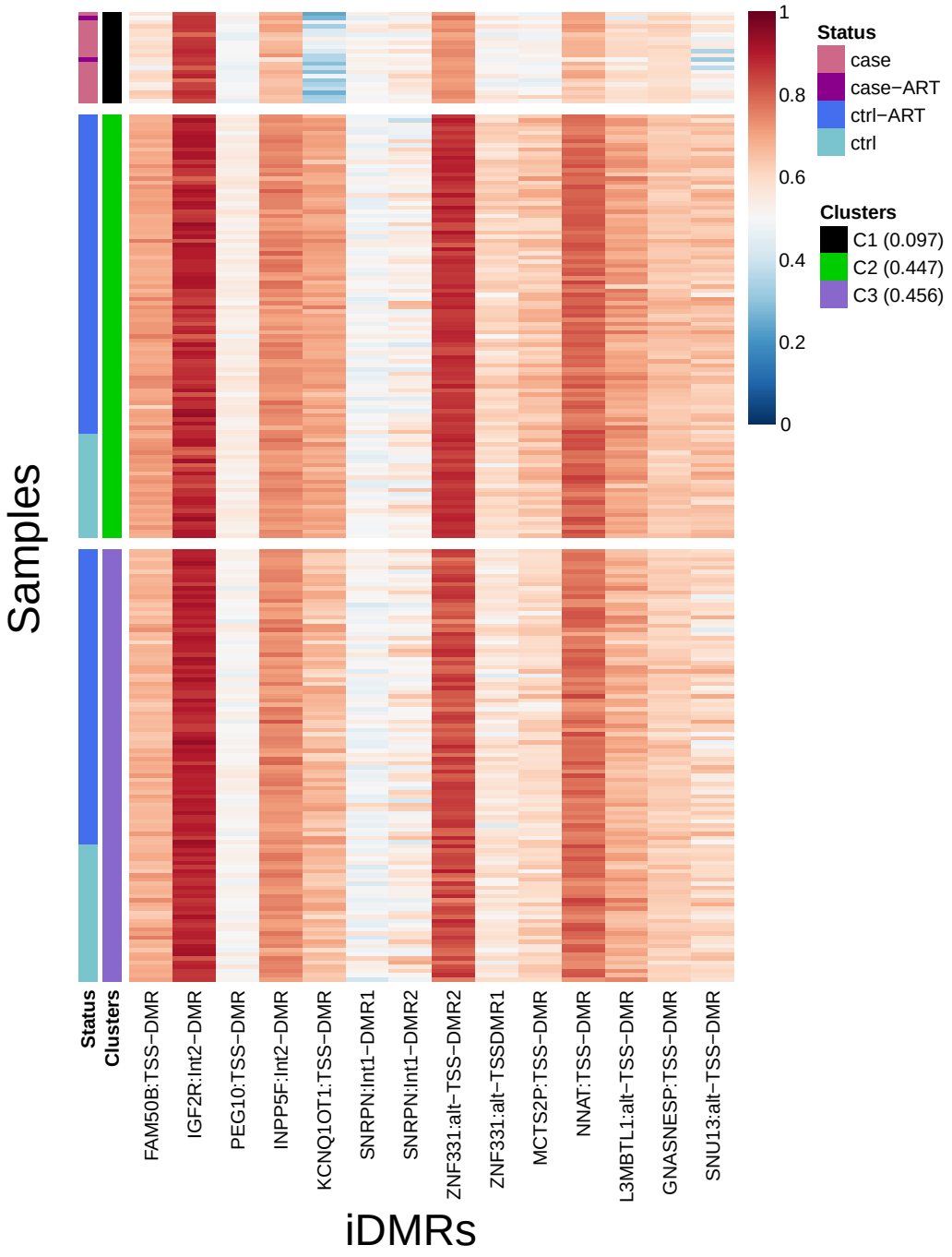


Figure 5.4 Clustered heatmap of the beta neonatal methylation data -*via* the VB-DPBM- only for the discriminative iDMRs. The x-axis represents the reduced in number iDMRs (14 in total), while y-axis the samples (228 neonates). The colour scale of the beta-intensities begins with blue (0% methylation), carries on with white (50% methylation) and ends up to red (100% methylation). The Clusters column displays the three neonates' groups in different colours (mixing weights are given on the legend). Status is also laid out for each cluster. The clusters are presented in an increasing mixing weight sequence (smallest cluster on top, largest at the bottom).

significant platform effect. Therefore, we carry on the analysis based on the original beta methylation data and not the residuals, since the data source (450K or EPIC) appears to have no confounding effect.

	Number of clusters	**Percentage of agreement (%)**
All iDMRs	4	79.82
Discriminatory iDMRs	3	

Table 5.1 Number of clusters in the beta methylation dataset with all the iDMRs and with only the discriminatory ones. The percentage of agreement, based on the measure described in Chapter 4, Section 4.3.3, is also calculated. Rate equal to 100% indicates common clusters (all data points are identically allocated in the full and reduced dataset), while the opposite (values close to 0%) implies completely different clusters.

Returning to Figure 5.2 and the clusters of the original beta methylation data with all the iDMRs, we observe that C2, C3 and C4 do not visually show clear distinction in terms of methylation patterns. Hence, by selecting the discriminative iDMRs *via* the Lin et al. [76] discriminative measure (Chapter 2, forward selection algorithm 7) and applying again the VB-DPBM algorithm on the reduced dataset, we aim at finding clusters with different methylation patterns.

In Figure 5.4, the final clusters of the beta methylation data with only the discriminative iDMRs (names on the x-axis) are three in total, instead of four as in the case with all the iDMRs (Figure 5.2). The agreement in clustering between the clusters derived from all the iDMRs and only from the discriminative is equal to 79.82% (Table 5.1). Since the clustering agreement is relatively high (almost 80%, implying relatively low cluster differentiation after the iDMR reduction) and the number of iDMRs is already considerably reduced from 33 to 14 (see discriminative iDMRs on the x-axis in Figure 5.4), we decide not to decrease further the feature number and stop at 14 discriminative iDMRs. Therefore, we proceed with the description of the three final clusters: C1 includes 9.7% of the total neonates, with all of them having the BWS disorder, while C2 and C3, contain 44.7% and 45.6% of the total neonates, that are BWS free.

Cluster	**Control**	**Case**	**Control-ART**	**Case-ART**
C1	-	90.90	-	9.10
C2	24.50	-	75.49	-
C3	31.73	-	68.27	-

Table 5.2 Frequencies (%) of Status categories (Control, Case, Control-ART, Case-ART) *per* VB-DPBM cluster of the beta methylation dataset that contains only the discriminative iDMRs (see Figure 5.4).

Table 5.2 exhibits the frequencies of the BWS cases and controls (both ART and non-ART) *per* cluster. C1 consists in the majority of naturally conceived neonates

with BWS and C2, as well as C3, are mostly comprised of neonates conceived by ART without BWS.

Having retrieved the three final clusters in the beta methylation data with only the discriminative iDMRs, we apply again the discriminative measure and find the important iDMRs *per* the final clusters. In Table 5.3, the discriminative iDMRs are given with order of addition, as well as accuracy level[3]. In particular, C1 is almost exclusively formed due to the hypomethylation of the KCNQ1OT1:TSS-DMR that provides discriminative accuracy 91.7%, validating our speculations about the importance of this imprinted region based on the heatmap in Figure 5.2. Specifically, significant hypomethylation of the KCNQ1OT1:TSS-DMR works as a diagnostic for the BWS (Weksberg et al. [148]), leading to the conclusion that we have correctly selected this iDMR as the most important for C1, since C1 is the cluster of neonates with BWS. However, if we want to reach a higher discriminative accuracy, we can include also one more iDMR, the SNRPN:Int1-DMR2, whose methylation (medium level) adds a 7.77% in the discrimination of C1, given the KCNQ1OT1:TSS-DMR involvement.

Discriminative iDMRs	**C1**	**C2**	**C3**
KCNQ1OT1:TSS-DMR	$\checkmark_1$ (0.917)		
SNRPN:Int1-DMR1			$\checkmark_1$ (0.727)
SNRPN:Int1-DMR2	$\checkmark_2$ (+0.077)	$\checkmark_1$ (0.605)	
ZNF331:alt-TSS-DMR2			$\checkmark_2$
ZNF331:alt-TSSDMR1		$\checkmark_2$ (+0.384)	$\checkmark_6$ (+0.066)
L3MBTL1:alt-TSS-DMR			$\checkmark_5$
GNASNESP:TSS-DMR			$\checkmark_3$ (+0.205)
SNU13:alt-TSS-DMR			$\checkmark_4$

Table 5.3 Cluster discrimination by specific iDMRs, for each cluster of the beta methylation dataset with only the discriminative iDMRs. The check mark denotes the discriminative iDMR, the subscript next to the checkmark defines the entrance sequence (the forward selection order) of the corresponding iDMR and the number in the parenthesis shows the discriminative accuracy level we reach after the selection of this iDMR (only for the first selected iDMR). Subsequent selections display their addition on the accuracy by the "+" sign. iDMRs with no addition serve as intermediate steps for reaching higher accuracy at the next forward iterations. The last added iDMR signifies convergence of the forward selection algorithm at 10^{-3}.

Regarding C2, the primary iDMR whose methylation level discriminates this cluster from the rest is the SNRPN:Int1-DMR2, with accuracy 60.5%, while the ZNF331:alt-TSSDMR1 increases the discriminative accuracy by 38.4%. As for C3, the first added

[3]We denote that the last added iDMR *per* cluster is the one corresponding to convergence of the forward selection algorithm at 10^{-3} units.

iDMR is the SNRPN:Int1-DMR1 with 72.7% contribution in the discrimination, whilst four more are added with the ZNF331:alt-TSS-DMR2, L3MBTL1:alt-TSS-DMR and SNU13:alt-TSS-DMR not increasing the discriminative accuracy *per se*, however their presence is important to reach the final total accuracy of 99.8%. As for the methylation pattern of these two clusters, we cannot find bold methylation differences between them by looking at the heatmap in Figure 5.4 (similar methylation colours).

Discriminative iDMRs of C2 and C3	**Wilcoxon test: p-value**
SNRPN:Int1-DMR1	0.235
SNRPN:Int1-DMR2	0.959
ZNF331:alt-TSS-DMR2	<0.001
ZNF331:alt-TSSDMR1	<0.001
L3MBTL1:alt-TSS-DMR	<0.001
GNASNESP:TSS-DMR	<0.001
SNU13:alt-TSS-DMR	<0.001

Table 5.4 Wilcoxon rank sum test for the difference in mean beta-intensities between C2 and C3, *per* discriminative iDMR of C2 and C3. P-values < 0.001 indicate significant methylation difference between the two clusters for this iDMR.

Consequently, to understand the partition mechanism for C2 and C3, we perform the Wilcoxon sum rank test on each of the discriminative iDMRs of these two clusters. From Table 5.4, we have strong evidence (p-value < 0.001) that all the C2 and C3 discriminative iDMRs have different mean methylation between these two clusters, except the SNRPN:Int1-DMR1 and SNRPN:Int1-DMR2. Therefore, even if the heatmap does not display differentiations on the methylation pattern, the VB-DPBM algorithm manages to discriminate C2 and C3 based on mean differences in specific discriminative iDMRs. However, it is not straightforward which of the control groups (C2 and C3) is more damaged in terms of iDMR alterations.

At this point, for the beta methylation data with the 14 discriminative iDMRs, we test the agreement in clustering between the VB-DPBM algorithm and the three non-probabilistic clustering methods, discussed in Chapter 4. Particularly, these are K-means, Hierarchical clustering and DBSCAN, and the agreement with the VB-DPBM is only 56%, 55% and 54% respectively. We then do comparisons only between the three non-probabilistic clustering methods and notice they do not agree in high degree either (approximately 50% concordance) in their clustering. Therefore, we conclude that the current dataset with the beta-intensities *per* discriminative iDMR is a challenging data type to conduct analysis on. We therefore proceed to the analysis of the count and binary methylation data, aiming at finding more informative and robust results regarding the clustering and the methylation pattern of each cluster.

5.2.2 Count Methylation Data

The second instance of data to be analyzed in this section concerns counts and specifically the number of significantly affected CpGs (alteration in normal methylation level) *per* iDMR. A suitable probabilistic clustering technique for this type of data is the VB-DPPM algorithm. More precisely, the VB-DPPM models the probability of events occurring in an interval such as the number of abnormally methylated CpGs sites within the iDMR. We prefer the variational Dirichlet Process Poisson mixture model over the variational Dirichlet Process Binomial mixture model (Appendix B, Subsection B.2.2) because the CpGs within an iDMR are not independent as the Binomial experiment would require.

In this section, we start the analysis by clustering the count methylation data with all the iDMRs present. Specifically, Figure 5.5 displays the six clusters in which the dataset is split, with C3 appearing to be the group of neonates that records the most significantly altered CpG sites compared to the rest of the groups, especially in the KCNQ10T1:TSS-DMR. The cluster C3 contains again all those newborns with the BWS phenotype, revealing the success of the VB-DPPM in clustering together the BWS cases. At this point we highlight the "Sex" column on the left of the figure which visually indicates the allocation of the two genders into each cluster. Table 5.5 quantifies this allocation.

Cluster	**Female**	**Male**
C1	3%	4%
C2	5%	3%
C3	5%	6%
C4	6%	8%
C5	9%	11%
C6	72%	68%

Table 5.5 Allocation of female and male neonates (in %) into the VB-DPPM clusters of the count methylation dataset.

In particular, females and males appear to be almost homogeneously assigned into the clusters. Given the Chi-squared test, we have strong evidence that gender is not a confounding parameter (Chisq.test = 1.33, p-value= 0.93), therefore the clustering is driven by the methylation profile and not the sex of the newborn. Regarding the rest of the clusters, C1, C2, C4, C5 and C6 refer to BWS free neonates and each one shows different levels of CpG alteration. The platform effect (EPIC, 450K) has been *a priori* removed from the count methylation dataset, during the construction of the counts, and there is no need for extra actions. Specifically, the median beta-intensities of the BWS controls from the EPIC array are used as reference level to declare a CpG site of

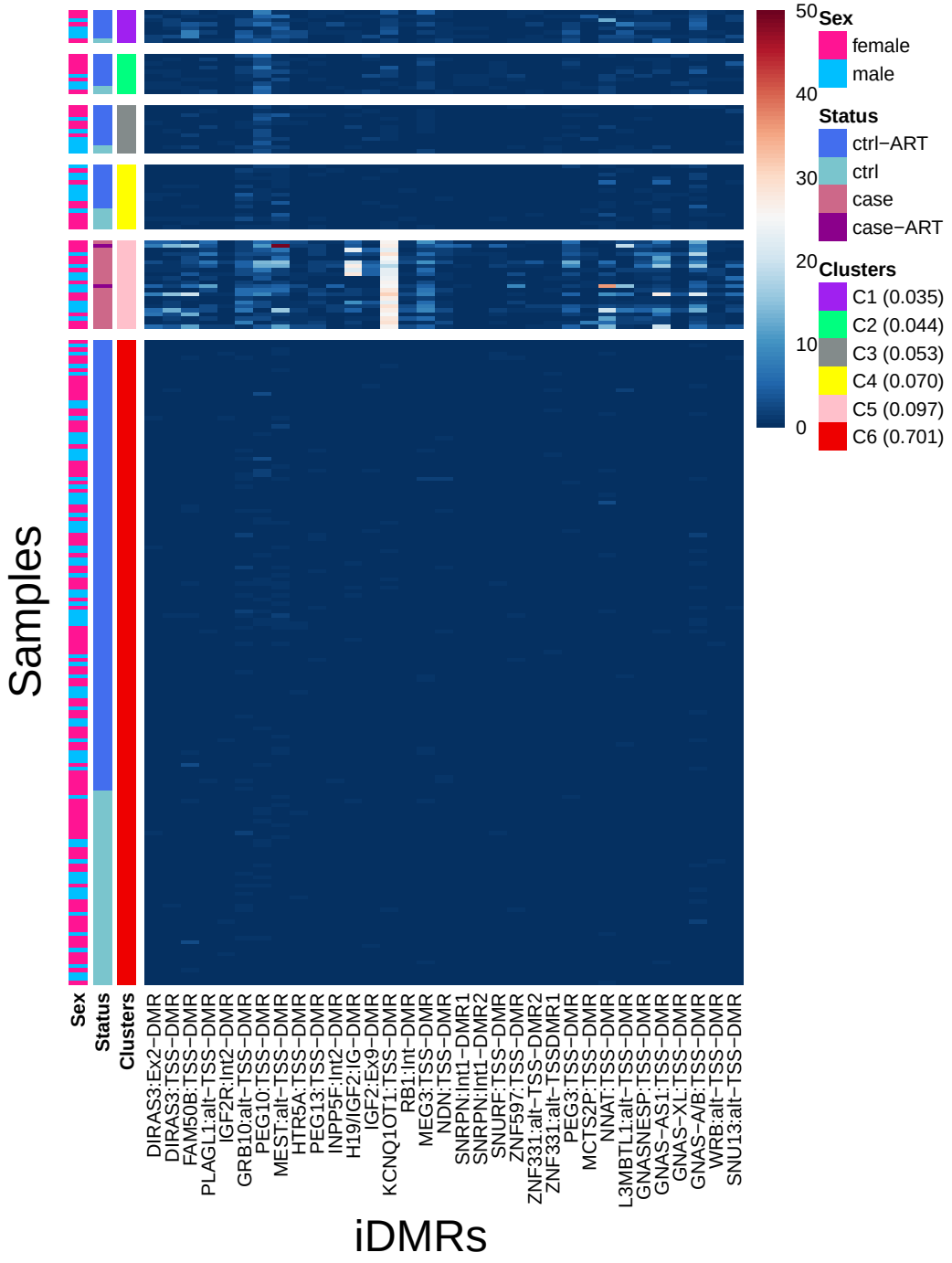


Figure 5.5 Clustered heatmap of the affected CpG counts in neonates, *via* the VB-DPPM. The x-axis bears the iDMRs (33), while y-axis the samples (228 neonates). The colour scale of the counts starts from blue (zero CpGs affected within iDMR), scales up to white (around 25 CpGs affected) and concludes to red (>40 altered CpGs). Clusters column on the left of x-axis displays the variational clusters in different colour (mixing weights are also given on the right). Status and Sex are shown for each cluster. The clusters are presented in an increasing mixing weight sequence (smallest cluster on top, largest at the bottom).

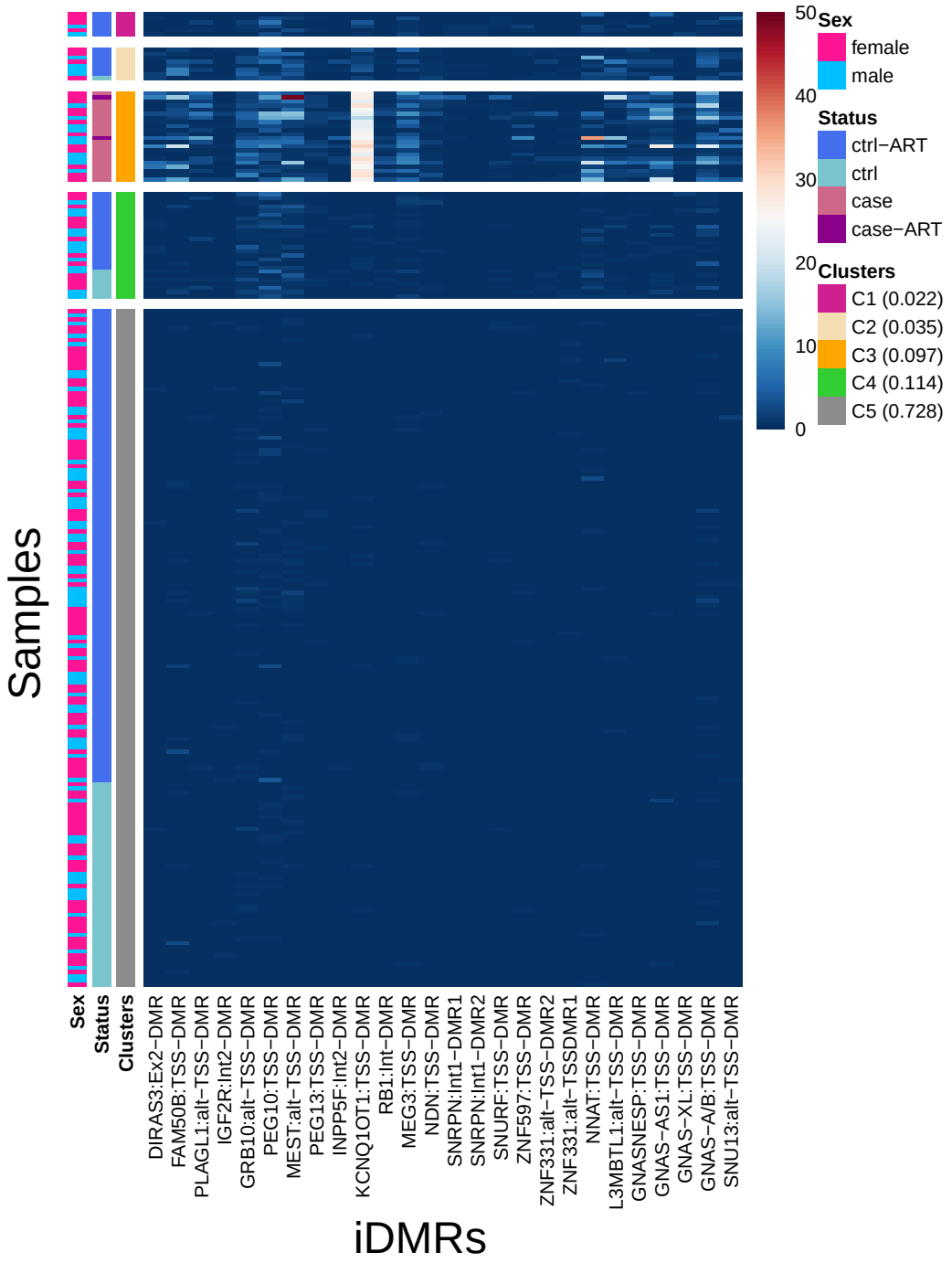


Figure 5.6 Clustered heatmap of the affected CpG counts in neonates -*via* VB-DPPM-only on the discriminative iDMRs. The x-axis corresponds to the iDMRs (26), while y-axis the samples (228 neonates). The colour scale of the counts begins with blue (zero CpGs affected within iDMR), rises up to white (around 25 CpGs affected) and concludes to red (> 40 altered CpGs). Clusters column on the left of x-axis shows the clusters in different colour (mixing weights and cluster indices are given on the legend). Status and Sex are displayed for each cluster too. The clusters are presented in an increasing mixing weight sequence (smallest cluster on top, largest at the bottom).

each neonate - either from 450K or EPIC platform - significantly or not significantly affected and thus, this contrast automatically adjusts the effect of the confounder.

Clustering on Discriminative iDMRs

As a subsequent analysis step, we clean the count methylation data by selecting the iDMRs that discriminate each cluster from the rest. We then join the discriminative iDMRs for each cluster into one set and apply the VB-DPPM algorithm again on the reduced dataset (only with the discriminative iDMRs). The clustering is displayed in Figure 5.6, where the final number of clusters is five instead of six (six was in Figure 5.5). The concordance between the clusters based on all the iDMRs (Figure 5.5) and those based only on the discriminative (Figure 5.6) is 90.35% (Table 5.6), denoting that the removal of the noisy iDMRs is not changing considerably the clustering results. Therefore, we conclude at not reducing further the number of iDMRs and retain the clustering results in Figure 5.6 for further analysis.

	Number of clusters	**Percentage of agreement (%)**
All iDMRs	6	90.35
Discriminatory iDMRs	5	

Table 5.6 Number of clusters in the count methylation dataset with all the iDMRs and with only the discriminatory ones. The percentage of agreement, based on the measure described in Chapter 4, Section 4.3.3, is also calculated. Rate equal to 100% indicates common clusters (all points are identically allocated in the full and reduced dataset), while the opposite (values close to 0%) implies considerably differing clusters.

Cluster	**Control**	**Case**	**Control-ART**	**Case-ART**
C1	-	-	100	-
C2	12.50	-	87.50	-
C3	-	90.90	-	9.10
C4	26.92	-	73.07	-
C5	30.12	-	69.87	-

Table 5.7 Frequencies (%) of Status categories (Control, Case, Control-ART, Case-ART) *per* VB-DPPM cluster of the count methylation dataset that contains only the discriminative iDMRs (see Figure 5.6) .

Concerning the allocation of the neonates into the final clusters, as in the beta methylation data analysis, one of the five groups incorporates 9.7% of the total neonates (cluster C3) that also bear the rare developmental BWS disorder and have been either naturally or artificially conceived. These newborns also seem to have the most aberrantly methylated CpG sites *per* iDMR (white coloured samples in C3). The

four remaining clusters refer to BWS controls neonates (ART and non-ART) whose iDMRs have gradually less affected sites, with decreasing order: C2, C1, C4 and finally C5. The distribution of controls and cases within each group is provided in Table 5.7. Cluster C3 consists mostly of BWS neonates conceived naturally, whereas C1, C2, C4 and C5 mainly of newborns without BWS that have been conceived through ART.

To assist in picturing the gradient of methylation alterations, we set the clusters in order of modification degree, starting from low to high - C5, C4, C1, C2 and C3. The cluster C3, with the highest alteration, consists of neonates with the BWS disorder, ascertaining again the success of the VB-DPPM algorithm in clustering together the BWS cases. The interest although is captured on the clusters of neonates without the BWS disorder who are also mainly conceived by ART (C5, C4, C1 and C2). In particular, we easily observe in the heatmap (Figure 5.6) this gradient of modifications for the controls groups that goes as follows:

1. C5: 72.8% neonates with negligible CpG alteration in all iDMRs
2. C4: 11.4% neonates with little CpG alteration in few iDMRs
3. C1: 2.2% neonates with some CpG alteration in some iDMRs
4. C2: 3.5% neonates with higher CpG alteration in some iDMRs.

Subsequently, this pattern manifests that neonates that have been artificially conceived, and do not have the BWS phenotype, may still have recorded some abnormal methylation on a few of their iDMRs, implying possible association of ART with potentially ongoing imprinting disorders.

Responsibilities

On a different note, we present here the superiority of the variational mixture models at providing a confidence level regarding the allocation of each neonate into a cluster. As we have discussed in Chapter 2, Section 2.6, mixture models allow a sample to be assigned into each cluster with a probability resulting in a vector of *responsibilities*. In Figure 5.7, we illustrate the responsibilities table in a heatmap form for the count methylation dataset. In this heatmap we can straightforwardly see that the BWS neonates of C2 and the controls of C1 and C3 are almost exclusively assigned into their cluster (red colour). Regarding C4 and C5, there are a few BWS-free neonates conceived by ART who slightly exchange participation between these two clusters and therefore are less confident to belong to only one (not exclusively allocated into their main cluster but have some probability to belong to the other too). This result supplies further information on a neonate level for future investigation. However, since the amount of those neonates is low, we proceed our analysis by assigning them into the cluster they are more probable to belong (final clusters in Figure 5.6).

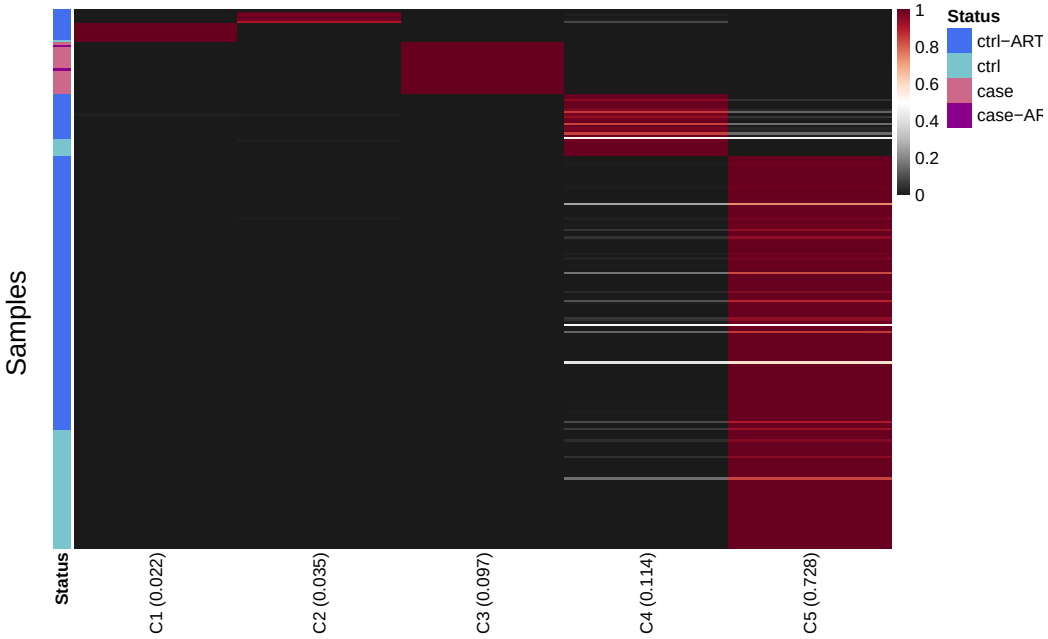


Figure 5.7 Heatmap of the responsibilities produced by the VB-DPPM for the count methylation dataset. Each row represents a neonate while the columns the final clusters (those in Figure 5.6). The color scale corresponds to the probability of a neonate to belong in each cluster (responsibilities). Red color denotes high probability (close to 1 or exact 1) and black low (close to 0 or exact 0). The Status column is also given.

Final Cluster Discrimination

Here we present and discuss the iDMRs that drive the discrimination of the five final clusters illustrated in Figure 5.6.[4] Table 5.8 presents the discriminative iDMRs along with their contribution into each cluster, and Figure 5.8 displays the proportion of methylation alteration for each of the discriminative iDMRs, grouped by the five clusters.

In particular, based on the information provided by both Table 5.8 and Figure 5.6, we enlist the most important iDMRs (first added) for each cluster separately, alongside their methylation alteration:

- **For C1:** SNRPN:Int1-DMR2 is an imprinting region associated to Prader–Willi syndrome (PWS) (Cassidy et al. [22]) and discriminates C1 by 96.3%. The neonates in this cluster have proportion of significantly affected CpGs in this iDMR between $0 - 16\%$, with median proportion at 8%.

[4]These iDMRs are not necessarily the most affected ones. It could be that an iDMR is significantly affected in one sub-population but unaffected in another leading the separation into two sub-populations for example.

- **For C2:** SNRPN:Int1-DMR2 is again the most salient and discriminates C2 by 94.9%. The neonates in this cluster have 0% (median) affected CpGs in this iDMR. Only few neonates have up to 4%.
- **For C3:** KCNQ1OT1:TSS-DMR is the diagnostic region for BWS and discriminates C3 by 89.5%. All neonates have 70 − 80% of their CpGs significantly affected for this iDMR.
- **For C4:** INPP5F:Int2-DMR discriminates C4 by 80.6%. These neonates have 0% CpGs affected for this iDMR, apart from an outlier newborn.
- **For C5:** MEST:alt-TSS-DMR is a diagnostic region for the Silver-Russell syndrome (SRS) (Wollmann et al. [151]) and discriminates C5 by 66%. Neonates have 0% (median) affected CpGs in this iDMR.

Discriminative iDMRs	**C1**	**C2**	**C3**	**C4**	**C5**
GRB10:alt-TSS-DMR			✓$_3$		✓$_5$ (+0.021)
PEG10:TSS-DMR		✓$_2$ (+0.015)	✓$_2$ (+0.010)		✓$_2$ (+0.022)
MEST:alt-TSS-DMR			✓$_{10}$ (+0.092)	✓$_3$ (+0.014)	✓$_1$ (0.660)
INPP5F:Int2-DMR	✓$_3$ (+0.006)			✓$_1$ (0.806)	
KCNQ1OT1:TSS-DMR			✓$_1$ (0.895)		
MEG3:TSS-DMR			✓$_7$ (+0.002)		✓$_6$ (+0.002)
NDN:TSS-DMR			✓$_9$		
SNRPN:Int1-DMR1	✓$_2$ (+0.002)				
SNRPN:Int1-DMR2	✓$_1$ (0.963)	✓$_1$ (0.949)	✓$_4$	✓$_4$ (+0.001)	
ZNF331:alt-TSS-DMR2				✓$_2$ (+0.016)	
ZNF331:alt-TSSDMR1			✓$_8$		✓$_4$ (+0.001)
GNAS-AS1:TSS-DMR			✓$_6$ (+0.001)	✓$_5$ (+0.032)	
GNAS-A/B:TSS-DMR			✓$_5$	✓$_6$ (+0.015)	✓$_3$ (+0.012)

Table 5.8 Cluster discrimination by specific iDMRs, for each cluster of the count methylation dataset with only the discriminative iDMRs. The check mark denotes the discriminative iDMR, the subscript next to the checkmark defines the entrance sequence (the forward selection order) of the corresponding iDMR and the number in the parenthesis shows the discriminative accuracy level we reach after the selection of this iDMR (only for the first selected iDMR). Subsequent selections display their addition on the accuracy by the "+" sign. iDMRs with no addition serve as intermediate steps for reaching higher accuracy at the next forward iterations. The last added iDMR signifies convergence of the forward selection algorithm at 10^{-3}.

General Discussion on Discrimination

SNRPN:Int1-DMR2 is not significantly altered methylation-wise in C2 and is lightly altered for C1. This could be a hint for possible risk of PWS onset for the artificially conceived neonates of C1 and maybe avoidance of this risk for C2. Nonetheless, C1 and C2 are comprised of only a little number of neonates, thus this assumption should be investigated deeper. On the other hand, KCNQ1OT1:TSS-DMR is still righteously

the main one responsible for discriminating C3 - the cluster of BWS cases - due to the high number of significantly affected CpGs compared to the rest of the clusters. One interesting observation concerns the additional selection of the GNAS-A/B:TSS-DMR and GNAS-AS1:TSSDMR for C3 (Table 5.8). GNAS-A/B and GNAS-AS1 are imprinting regions that work as diagnostics for PHP1b - a disorder that causes lack of response to parathyroid hormone (a hormone that manages vitamin's D, calcium's and phosphorous' levels in the blood, Mantovani et al. [85]). These diagnostic iDMRs are detected within C3 in Figure 5.6 as two regions wherein some neonates present relatively high number of affected CpGs (see also boxplots for these regions in Figure 5.8). This result could potentially indicate the existence or progression of the PHP1b disorder amongst some of the neonates with BWS.

Moreover, in Figure 5.8 we notice that the discriminative PEG10:TSS-DMR which is a diagnostic region for the Silver-Russell syndrome (SRS) (Wollmann et al. [151]) appears to be affected to some light extent (50% of the neonates have 2.5 – 7.5% affected CpGs with median proportion of affected CpGs 5%), potentially implying that artificially conceived neonates without the BWS may still have recorded some abnormal methylation in diagnostic regions for SRS that could hint potential risk of onset.

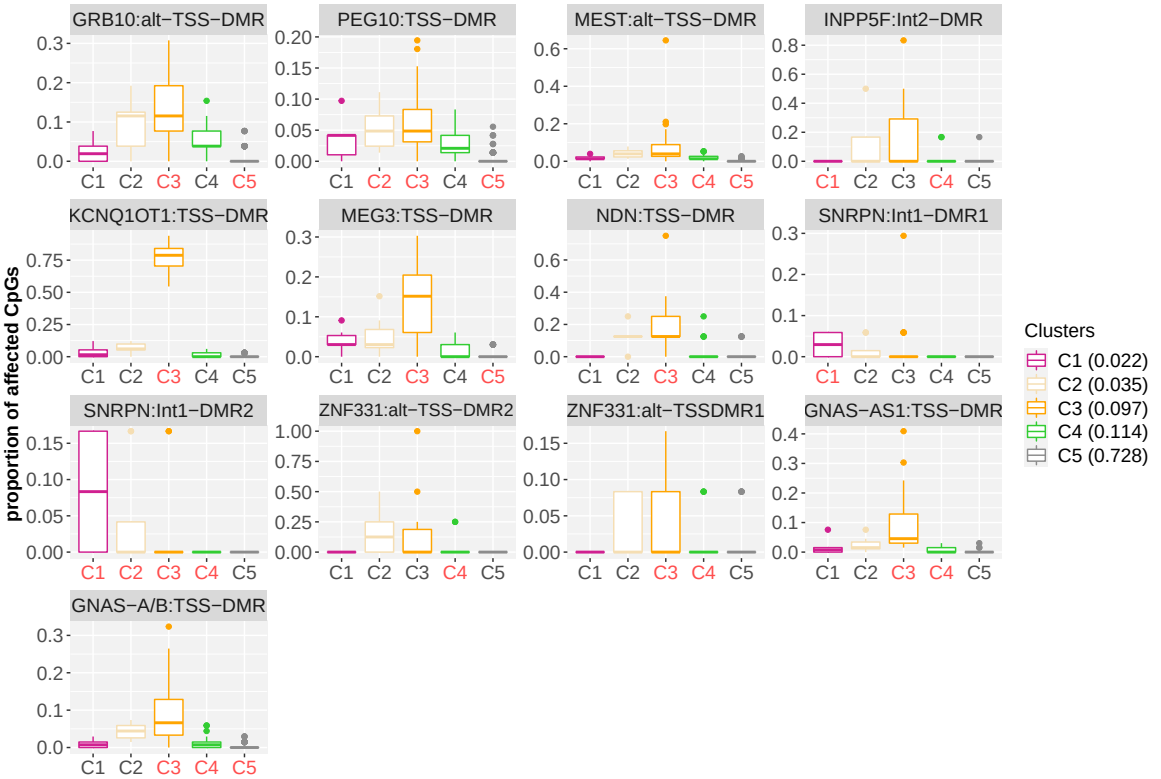


Figure 5.8 Boxplots of proportion of affected CpGs *per* discriminative iDMR, grouped by the clusters retrieved from the count methylation analysis. The iDMR is the title of each subplot and the clusters correspond to those in Figure 5.6, thus the boxplots have been coloured accordingly. In each iDMR plot, the clusters that are discriminated by the corresponding iDMR are highlighted in red on the x-axis. The scale of values on the y-axis is free for better resolution (fixed scale returns distorted resolution).

Comparison to Standard Clustering Tools

At this point and having completed the count methylation data analysis, we conduct an extra test with respect to the robustness of our results. Specifically, we compare the final five clusters from the VB-DPPM implementation with the clusters derived by K-means, Hierarchical clustering and DBSCAN. In contrast to the beta methylation analysis in Section 5.2.1, here the count analysis by VB-DPPM coincides in the clustering performance by 83% with DBSCAN, 85% with K-means and 82% with Hierarchical clustering. In conclusion, the analysis on count methylation data has been more informative and robust than on the beta methylation data, presenting explicitly the methylation modification level in each cluster and the potential association of ART with imprinting disorders (Novakovic et al. [105]) such as SRS and PHP1b.

5.2.3 Binary Methylation Data

The last analysis on the set of neonates is with regard to the binary methylation measure that refers to significantly or non-significantly affected iDMRs for each neonate. The binary nature of the data indicates the selection of a probabilistic clustering algorithm with binary support range such as the VB-DPBerM.

In this section, we start again the analysis by clustering *via* the VB-DPBerM the binary dataset with all the iDMRs. In Figure 5.9, we observe that the total number of clusters is nine, with the C1 to C7 showing aberrant methylation on specific iDMRs and C8 on most of the iDMRs, whereas C9 mainly appears with no alterations (non-affected iDMRs). However, to highlight important characteristics of the data, we find and remove the iDMRs with no discrimination ability, then we apply again the VB-DPBerM on the reduced dataset and obtain the final clusters in Figure 5.10.

Clustering on Discriminative iDMRs

Figure 5.10 presents the clusters based only on the discriminative iDMRs. The number of clusters drops from nine to three while the percentage of clustering agreement between the full dataset (all iDMRs) and the one with only the discriminative iDMRs is 85.53% (Table 5.9). Since the concordance rate is relative high (> 85%) and the number of iDMRs has been considerably reduced from 33 to 15 (see x-axis in Figure 5.10), we decide to stop any further feature reduction and continue the analysis based on C1, C2 and C3 in Figure 5.10.

In regard to the allocation of the newborns into the three clusters, we have strong evidence this is not driven by the sex differences (Chisq.test = 0.36, p-value= 0.83; see

also Table 5.10) but by their methylated iDMRs as depicted in the clustered heatmap in Figure 5.10.

	Number of clusters	**Percentage of agreement (%)**
All iDMRs	9	85.55
Discriminatory iDMRs	3	

Table 5.9 Number of clusters in the binary methylation dataset with all the iDMRs and with only the discriminatory ones. The percentage of agreement, based on the measure described in Chapter 4, Section 4.3.3, is also calculated. Rate equal to 100% indicates common clusters (all points are identically allocated in the full and reduced dataset), while the opposite (values close to 0%) implies considerably differing clusters.

Cluster	**Female**	**Male**
C1	1%	2%
C2	19%	18%
C3	80%	80%

Table 5.10 Allocation of female and male neonates (in %) into the VB-DPBerM clusters of the binary methylation dataset that contains only the discriminative iDMRs (see Figure 5.10).

In particular, C1 includes 1.3% of the total neonates, whose abnormal methylation is reported on their GNASNESP:TSSDMR. C2 is comprised of 18.9% of the total neonates, who present multiple significantly affected iDMRs, contrarily to C3 where the 79.8% of the samples refer to neonates with non important alteration in their iDMRs (grey colour).

Cluster	**Control**	**Case**	**Control-ART**	**Case-ART**
C1	33.33	-	66.66	-
C2	11.62	46.51	37.20	4.65
C3	28.57	-	71.42	-

Table 5.11 Frequencies (%) of Status categories (Control, Case, Control-ART, Case-ART) *per* VB-DPBerM cluster of the binary methylation dataset that contains only the discriminative iDMRs (see Figure 5.6) .

At this point, we stress out the low dimensionality of the real datasets under study. Updated conclusions may be revealed in larger sizes. However, we proceed with the discussion based on these results. Regarding the frequencies of the BWS cases and controls into the three clusters, C1 is a control group and especially a cluster of artificially conceived neonates, same as C3. Nonetheless, the interest is captured in C2. In the previous beta methylation and count methylation data analyses, neonates with BWS were all uniquely clustered in one cluster, while the newborns without

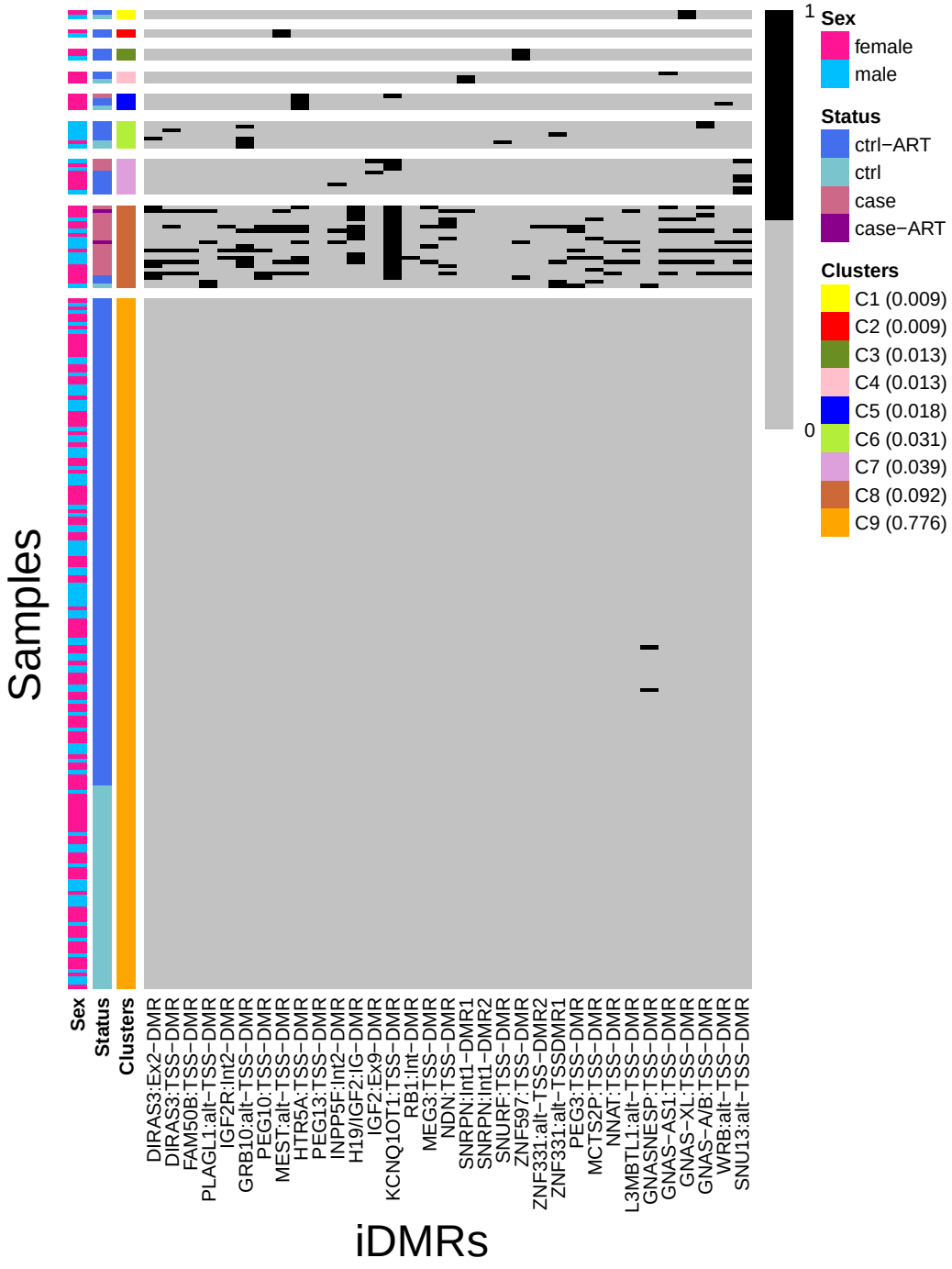


Figure 5.9 Clustered heatmap of the binary methylation data. Clustering achieved by the VB-DPBerM. The x-axis stores the iDMRs (33 in total), while y-axis the samples (228 neonates). The binary values are either grey (coded by 0, denoting non-significantly affected iDMR) or black (coded by 1, implying significantly affected iDMR). The Clusters column shows the data clusters in different colours (mixing weights are displayed on the right of the heatmap for each cluster). Status and Sex are also given for each cluster.

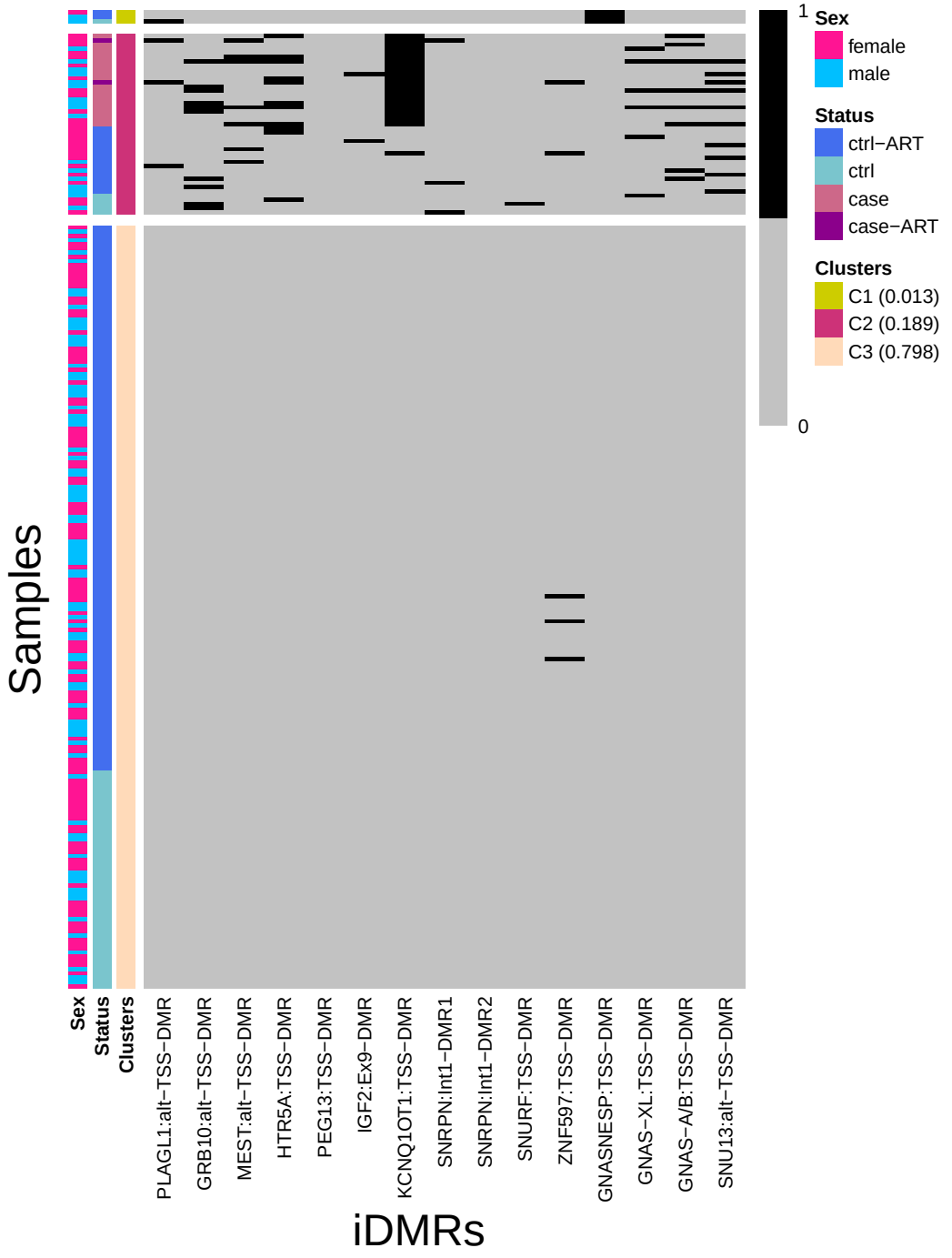


Figure 5.10 Clustered heatmap of the binary methylation data -*via* the VB-DPBerM- only for the discriminative iDMRs. The x-axis represents the reduced iDMRs (15 in total), while y-axis the samples (228 neonates). The binary values are either grey (coded by 0, denoting non-significantly affected iDMR) or black (coded by 1, implying significantly affected iDMR). The Clusters column displays the three neonates' clusters in different colours (mixing weights and cluster indices are given on the legend). Status and Sex are also provided for each cluster.

BWS in different ones. Here, C2, which is the cluster with the most affected iDMRs, encompasses both neonates with and without BWS (Table 5.11). Particularly,

- 51.16% are neonates with BWS
- 48.82% are neonates without BWS (with the majority conceived by ART)

Based on these results, we could suspect that neonates who have been mostly conceived by ART seem to have enough number of significantly altered iDMRs in order to be grouped along with the cases of BWS who have a lot of alterations. To enforce this conclusion, we apply the discriminative measure again, for each of the three clusters, so as to discover which iDMRs are important in discriminating C3, as well as C1 and C2.

Responsibilities

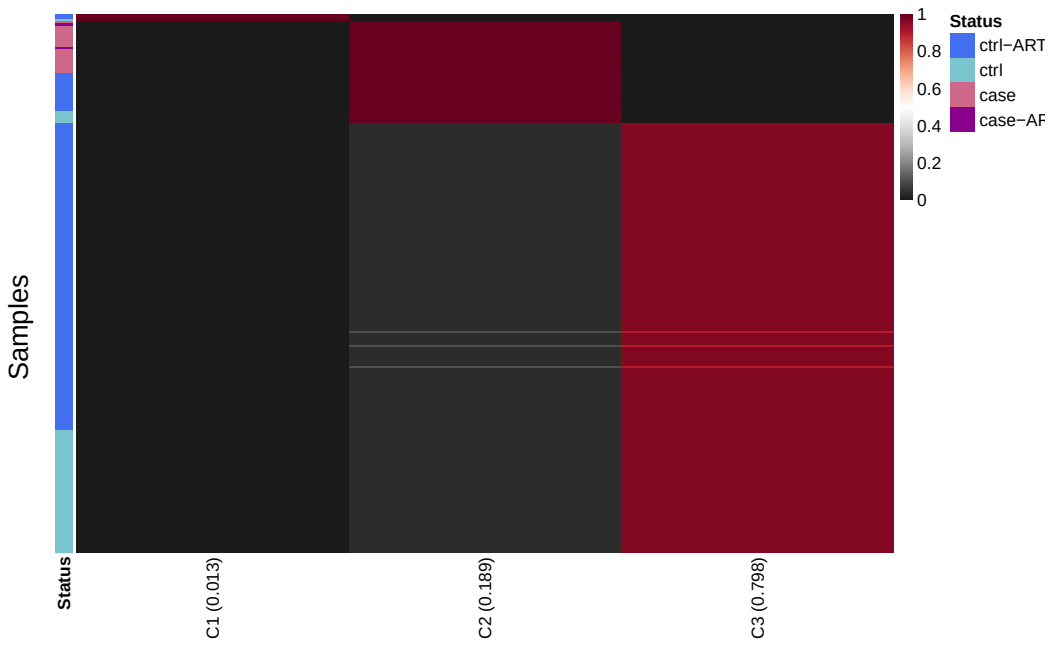


Figure 5.11 Heatmap of the responsibilities produced by the VB-DPBerM for the binary methylation dataset. Each row represents a neonate while the columns the final clusters (those in Figure 5.10). The color scale corresponds to the probability of a neonate to belong in each cluster (responsibilities). Red color denotes high probability (close to 1 or exact 1) and black low (close to 0 or exact 0). The Status column is also given.

Finally, we present the responsibilities heatmap for the binary methylation dataset in Figure 5.2.3. We observe that the neonates of C1 and C2 have been exclusively allocated into their cluster (red colour). Concerning the neonates of C3, these are controls conceived by ART and are primarily allocated into C3 (big red rectangular); however these neonates have also a non-zero but low probability to belong in C2 too

(grey rectangular), where C2 is the cluster of BWS cases. Nevertheless, this probability is considerably lower, therefore we are quite confident to keep those neonates into their main cluster (C3).

Discriminative iDMRs	**C1**	**C2**	**C3**
GRB10:alt-TSS-DMR		✓$_2$ (+0.018)	✓$_2$ (+0.030)
HTR5A:TSSDMR		✓$_8$ (+0.002)	✓$_7$ (+0.027)
IGF2:Ex9-DMR			✓$_5$
KCNQ1OT1:TSS-DMR		✓$_1$ (0.700)	✓$_1$ (0.685)
SNRPN:Int1DMR1		✓$_6$ (+0.002)	✓$_8$ (+0.002)
ZNF597:TSSDMR		✓$_5$	✓$_6$ (+0.003)
GNASNESP:TSS-DMR	✓$_1$ (0.978)	✓$_4$ (+0.017)	✓$_4$ (+0.015)
GNASXL:TSSDMR		✓$_7$ (+0.028)	
SNU13:alt-TSS-DMR		✓$_3$ (+0.012)	✓$_3$ (+0.011)

Table 5.12 Cluster discrimination by specific iDMRs, for each cluster of the binary methylation dataset having removed the non-discriminatory iDMRs. The check mark denotes the discriminative iDMR, the subscript next to the checkmark defines the entrance sequence (the forward selection order) of the corresponding iDMR and the number in the parenthesis shows the discriminative accuracy level we reach after the selection of this iDMR (only for the first selected iDMR). Subsequent selections display their addition on the accuracy by the "+" sign. iDMRs with no addition serve as intermediate steps for reaching higher accuracy at the next forward iterations. The last added iDMR signifies convergence of the forward selection algorithm at 10^{-3} units.

Final Cluster Discrimination

Here we present and discuss the iDMRs that drive the discrimination of the three final clusters illustrated in Figure 5.10. Table 5.12 presents the discriminative iDMRs along with their contribution into each cluster, and Figure 5.12 displays the proportion of neonates with significant alteration in each of the discriminative iDMRs, grouped by the three clusters.

- **For C1:** GNASNESP:TSS-DMR is a PHP1b associated iDMR and discriminates C1 by 97.8%. The neonates in this cluster are all significantly affected in this iDMR.
- **For C2:** KCNQ1OT1:TSS-DMR is the diagnostic region for BWS and discriminates C2 by 70%. 53% of the neonates in this cluster are significantly affected in this iDMR.
- **For C3:** KCNQ1OT1:TSS-DMR is the most important here too and discriminates C3 by 68.5%. All neonates in this cluster are not affected in this iDMR (0%).

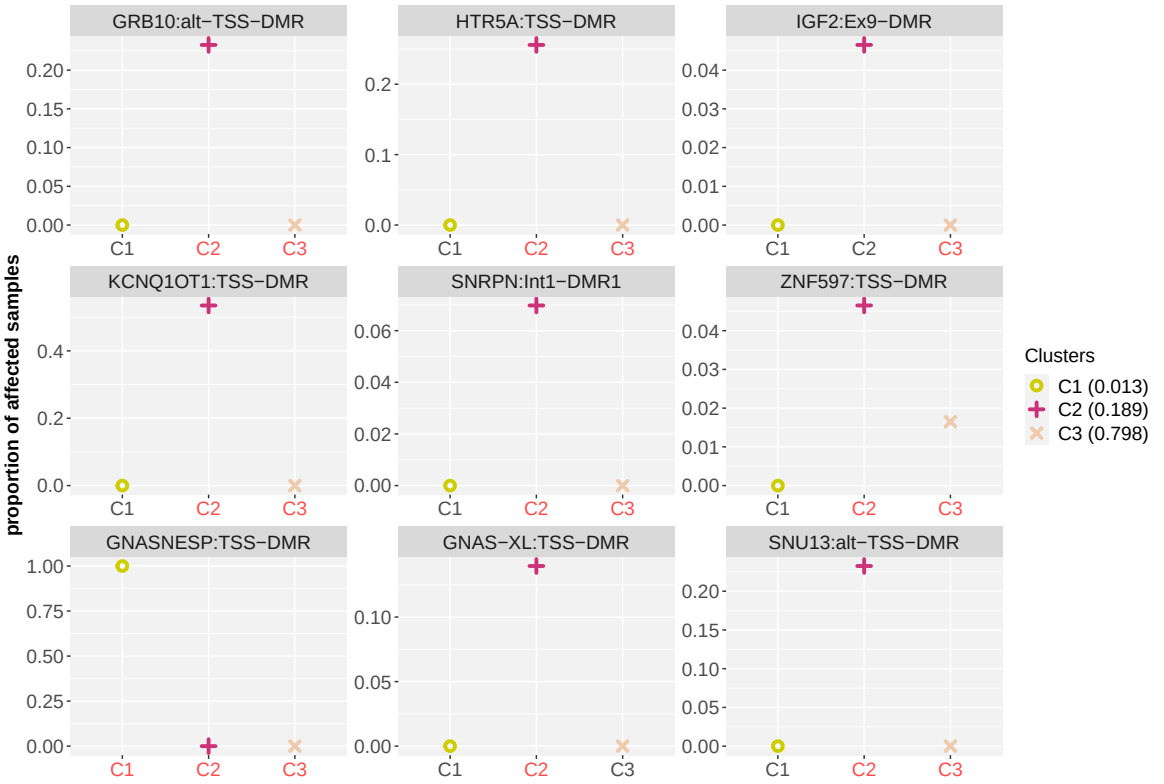


Figure 5.12 Plots of proportion of affected samples *per* cluster (binary methylation analysis), within the discriminative iDMR. The iDMR is the title of each subplot and the clusters correspond to those in Figure 5.10, thus the points have been coloured accordingly, as well as given shapes cluster-wise. In each iDMR plot, the clusters that are discriminated by the corresponding iDMR are highlighted in red on the x-axis. The scale of values on the y-axis is free for better resolution (fixed scale returns distorted resolution).

General Discussion on Discrimination

The irregular methylation of GNASNESP:TSS-DMR in C1 neonates could imply that newborns without BWS may still bear the possibility of developing the PHP1b disorder. However, this is a result based solely on a very small sample size (C1 contains only three neonates) and cannot be taken for granted. As for KCNQ1OT1:TSS-DMR, it discriminates by rights C3 and C2 since in the former newborns are BWS free with unaffected methylation in this iDMR, whereas in the latter the neonates are BWS cases and are considerably affected in this genomic region. For the C2 cluster, there is one more interesting discriminative iDMR, the GRB10:alt-TSS-DMR, which is associated to the SRS. The abnormal methylation record of this iDMR in neonates with BWS, and in neonates conceived artificially without BWS (both newborn types clustered in C2) could hint over existence or onset of the SRS disorder in BWS cases or controls conceived by ART.

Comparison to Standard Clustering Tools

To finish the binary analysis, we compute the agreement of our clustering results with K-means, Hierarchical clustering and DBSCAN. In particular, the VB-DPBerM clusters coincide with the ones suggested from K-means by 88.2%, from Hierarchical by 90% and from DBSCAN by 95%, signifying robust clusters. In summary of the binary methylation results, we have reasons to assume that artificially conceived neonates without the BWS phenotype who although have significantly affected iDMRs, like the GRB10:alt-TSS-DMR and the GNASNESP:TSS-DMR, may present high risk of developing imprinting disorders in the future such as SRS or PHP1b. Moreover, neonates with BWS may have or be prone to develop extra imprinting disorders (SRS and/or PHP1b), based on irregular methylation records on the same GRB10:alt-TSS-DMR and GNASNESP:TSS-DMR iDMRs.

5.3 Consensus Results

Having performed each analysis separately (beta methylation data, count methylation data and binary methylation data), we decide to summarise the results of the most informative ways of analysing DNA methylation - binary and counts - in order to create a consensus that could open discussion regarding the risk of onset of developmental disorders and the association of ART with abnormal DNA methylation in iDMRs[5]. Following this, we comment upon the common discriminative iDRMs in both analyses.

Risk groups

iDMR alteration	Binary analysis	Count analysis
Signalling	C1, C2	C1, C2, C3, C4
Non-signalling	C3	C5

Table 5.13 Level of iDMR alteration in the clusters of the binary and count analysis. The signalling and non-signalling clusters are supplied *per* analysis.

We begin by grouping in Table 5.13 the clusters of each analysis (counts and binary[6]) into two categories. The first one is called “Signalling” and accommodates clusters wherein the neonates have some or many iDMRs abnormally methylated and there is signal of high aberrant methylation. The second one is called “Non-signalling” and includes clusters wherein the neonates have little or none iDMRs aberrantly methylated

[5]This work is not a hypothesis testing. It is more like a descriptive analysis of the summarised clustering results with the intention to capture potential patterns of ART association with imprinting disorder and provoke discussion for further investigations.

[6]The clusters refer to the final clusters based on only the discriminative iDMRs for each analysis. Specifically, for the count methylation analysis, C1 to C5 correspond to the clusters in Figure 5.6 and for the binary methylation analysis C1, C2 and C3 to the clusters in Figure 5.10.

(low signal). This grouping is based on the earlier discussion made in each analysis separately.

Predicated on Table 5.13, we define three groups of neonates according to their allocation into "Signalling" or "Non-signalling" clusters in each of the two analyses. For instance, if a newborn is assigned into C2 in the binary analysis and into C3 in the count analysis (both "Signalling" clusters) then this neonate is placed into the High Risk group in Table 5.14. For this newborn, we would expect to have some aberrantly methylated iDMRs given the two methods agreed to allocate it into "Signalling" clusters. This could hint the predisposition or existence of an imprinting disorder for this neonate. Therefore, these three risk groups could also be associated with low, medium or high signalling concerns of potential risk for onset of a rare developmental disorder or predisposition to develop one.

Here we provide the definition for each group:

- The high risk category is defined as the group of neonates who have been allocated into "Signalling" clusters in both analyses. Specifically, these are the newborns that in the binary methylation analysis were assigned into C1 or C2 (high signal) *AND* in the count methylation analysis into C1, C2, C3 or C4 (high signal).
- The moderate risk category is defined as the group of neonates who have been allocated into a "Signalling" cluster in one analysis, while in the other into a "Non-signalling". These are the newborns that in the binary analysis were assigned into C1 or C2 (high signal) *AND* in the count analysis into C5 (low signal), *OR* in the count analysis into C1, C2, C3 or C4 (high signal) and in the binary into C3 (low signal).
- The low risk category is defined as the group of neonates who have been allocated into "Non-signalling" clusters in both analyses. These are newborns that in binary analysis were assigned into C3 (low signal) *AND* in the count analysis into C5 (low signal).

Table 5.14 exhibits the distribution of the three risk levels within the BWS cases and controls. We notice that all cases of BWS are allocated into the high risk category, as we normally expected, whilst neonates without BWS (controls/controls-ART) are mainly assigned within the low risk category (82.8%/73.6%). Nonetheless, controls, and especially conceived by ART, seem to occupy a place into the moderate and high risk categories too (18.9% and 7.4% respectively) potentially insinuating probable association of ART with the onset or predisposition of rare developmental imprinting disorders (based on recorded methylation patterns in their imprinted regions).

	Controls	Controls-ART	Cases (ART/non-ART)
High Risk	6.9% (4)	7.4% (11)	100% (22)
Moderate Risk	10.3% (6)	18.9% (28)	0% (0)
Low Risk	82.8% (48)	73.6% (109)	0% (0)

Table 5.14 Allocation rates of risk levels within the naturally and artificially conceived neonates (Controls/Controls-ART), as well as neonates with BWS (Cases). The risk levels are regarding the significance of aberrantly affected iDMRs, associated to the potential onset of imprinting disorders (High Risk, Moderate Risk and Low Risk). The allocation percentages are computed column wise (given Status category). The parentheses include the number of newborns in each category.

Common discriminative iDMRs

Finally, we append the common discriminative iDMRs (regardless of being first contributors or not in the discrimination of a cluster) in the count and binary methylation analysis, along with information on their methylation profile for the clusters they discriminate, provided by Figure 5.8 and Figure 5.12:

- **GRB10:alt-TSS-DMR**:
  - Binary: C1 $\rightarrow$ 25% affected neonates and C3 $\rightarrow$ 0% affected neonates
  - Counts: C3 $\rightarrow$ 9 – 20% affected CpGs for 50% of the neonates (those in the interquartile range) and C5 $\rightarrow$ 0% except two neonates with 4 – 9%
- **KCNQ1OT1:TSS-DMR**:
  - Binary: C2 $\rightarrow$ 53% affected neonates and C3 $\rightarrow$ 0% affected neonates
  - Counts: C3 $\rightarrow$ 70 – 80% affected CpGs for 50% of the neonates
- **SNRPN:Int1DMR1**:
  - Binary: C2 $\rightarrow$ 7% affected neonates and C3 $\rightarrow$ 0% affected neonates
  - Counts: C1 $\rightarrow$ 0 – 6% affected CpGs for 50% of the neonates (those in the interquartile range).

In a nutshell, three discriminative iDMRs are common in the count and binary analyses. GRB10:alt-TSS-DMR (associated with SRS) appears to be rather affected methylation-wise in some clusters of neonates, and not significantly affected in others (both analyses). On the other hand, KCNQ1OT1:TSS-DMR (diagnostic region for BWS) has significantly aberrant methylation (both analyses), whereas SNRPN:Int1DMR1 appears in both with low alteration. These results, along with the risk groups, can open the path for inputs and further discussion from experts in the field of imprinting disorders.

5.4 Summary

The analyses of the same data and same birth cohort by three different measures (beta methylation data, count methylation data and binary methylation data) revealed the

importance of approaching the same problem from multiple perspectives achieving more robust conclusions, while giving preference to the most informative approaches. In the beta methylation data, the clustering algorithm managed to cluster together neonates with the BWS disorder predicated on their hypomethylated iDMRs (mainly KCNQ1OT1:TSS-DMR). However, the rest two clusters of newborns without the BWS were not easily distinguishable in terms of methylation alteration, rendering this version the least informative to deduce further results. In the binary methylation data, the algorithm identified a subgroup of controls (mostly conceived by ART) with a profile of imprinting alterations close to cases with congenital imprinting alterations, indicating possible association of ART with imprinting disorders (Hattori et al. [58]). In the count methylation data, our method determined five clusters with four of them in controls. The counts of affected CpGs allowed to detect a gradient of alterations in imprinting regions of neonates without the BWS who were mostly conceived artificially, implying potential risk of an imprinting disorder development.

In conclusion, our variational Dirichlet Process mixture models demonstrated successful performance as clustering tools for methylation applications in birth cohorts, as illustrated in this chapter. The consensus results of the informative binary and count analysis showed inclination towards the possible impact of ART as aggravating factor for imprinting disorders onset (Hattori et al. [58]). The discriminative iDMRs that appeared in both analyses were the diagnostic region for BWS and two more iDRMs (one associated with SRS outcome) that could be meaningful for additional studies within the framework of the imprinting disorders.

Chapter 6

Conclusions and Discussion

In this final chapter, we summarize the conclusions of this thesis - especially the real data analysis presented in the previous chapter - and discuss the utility of our proposed approaches, along with directions for future work. Overall, this doctoral research presents novel toolkits for analyzing DNA methylation data measured in different ways due to the specific platform used or tailored transformed for the aim of the analysis. Moreover, it includes advanced methods for clustering discrete methylation data affected by confounding parameters such as sex, age, ethnicity etc. The adjustment for confounding effects has only been made for continuous methylation data (beta-intensities) by modelling the beta regression residuals instead of the original data. However, cluster analysis for discrete DNA methylation measurements that accounts for confounding effects has never been accomplished before due to the difficulty in specifying the residuals' distribution. This is the motive that inspired us to build model-based clustering algorithms that allow for confounding parameters into an internal regression process, avoiding the specification of the residuals' distribution. In this respect, the regression-clustering algorithms presented in this thesis are an important addition to the literature. Furthermore, one more valuable contribution of this work concerns the adoption of the second step analysis that follows the model-based clustering inference. This is the posterior selection of those features - here genetic loci - that are most important for segregating individuals into groups.

6.1 Summary

To test the performance and applicability of the proposed methods, we performed simulation tests for the models employed in our real analysis chapter. The conclusions from the simulations tests and the real data applications are briefly summarised below.

Chapter 4 showed that the models we derived for the DNA methylation cluster analysis - the variational non-Gaussian Dirichlet Process mixture models - were capable of clustering successfully both discrete and bounded continuous synthetic datasets. Specifically, the variational Dirichlet Process Beta mixture model and the variational Dirichlet Process Poisson and Bernoulli mixture models were rarely dropping below 90% in clustering accuracy when applied to data scenarios with low feature size and high sample size. The same high level of accuracy was also observed in applications with datasets of high feature and low sample size, showing that our tools are robust regarding the different specification of the number of features and samples. On the other hand, the commonly used non-probabilistic K-means, Hierarchical clustering and DBSCAN were generally less consistent and more prone to less accurate results on the same simulated scenarios. Moreover, the implementation of the *a posteriori* feature selection measure of Lin et al. [76] on simulated scenarios of high feature size demonstrated that the reduction of the features' dimension does not affect the clustering process, since the simulated number of clusters is retrieved. This result indicates that the selected features carry enough amount of information for the clustering structure.

Regarding Chapter 5, the analysis of real data from artificially and naturally conceived neonates with and without the Beckwith-Wiedemann Syndrome (BWS) showed that the number of affected CpGs *per* iDMR (count methylation data) and the significantly or non-significantly affected iDMRs (binary methylation data) were more informative ways to measure DNA methylation compared to the aggregated beta-intensities *per* iDMR (beta methylation data). The two former measures clearly revealed the methylation pattern in each subgroup of neonates, as well as allowed to detect an upward trend of methylation alterations in neonates without the BWS who were mostly conceived artificially, implying potential risk of an imprinting disorder development. The consensus results of the count and binary methylation data analysis opened the path for discussions regarding the higher risk of an imprinting disorder onset when artificial reproductive technologies are present; however the results are not unequivocal and further studies need to be performed on higher datasets.

6.2 Discussion

Our proposed clustering tools, apart from the DNA methylation applications, can be applied to any scenario that aims to cluster data of identically distributed features (*i.e.,* all features are assumed to be Beta distributed, or Poisson, or Binomially) and for which the requirement of independence between features is met. In principle, our models bear certain fine characteristics which render them attractive as clustering tools.

In brief, they are:

- flexible
  - for each data type we propose the appropriate Bayesian hierarchical model (Poisson, Binomial, Bernoulli, Beta, Gaussian)
  - in each sub-population (cluster) the cluster-specific parameters are also feature-specific (we allow features to have different mean and variance within the same cluster)
- self-determining
  - each model determines automatically the number of components without the need of pre-fixing the right number
- informative
  - each observation (*i.e.*, neonate) has a probability to belong to each cluster (soft clustering) yielding a level of confidence for the allocation, as opposed to standard tools like K-means and Hierarchical clustering which assign completely into one cluster (hard-clustering)
  - each sub-population has its own estimated distribution (variational distribution) and hence, further analyses can be made in each cluster regarding mean, variance, shape of distribution etc.
- scalable
  - each model is learned via variational algorithms that easily apply to high dimensional datasets and provide fast results, in contrast to MCMC methods
- regulatory
  - they regulate/account for the impact of confounding parameters (we supply these models separately defined as "variational mixture models with covariates")
- instrumental
  - they benefit the application of the discriminative accuracy measure owing to the fact that each sub-population is modelled by a specific distribution. This discriminative measure helps at selecting those features that lead the segregation into the sub-populations.

Regarding the "regulatory characteristic", the utility and advantage of the "variational mixture models with covariates" is the permission for cluster-specific effects of the confounding parameters. More precisely, the effect of the covariates (confounding parameter) is not global; each sub-population is allowed to receive at different level this effect. For example, the influence of gender may be higher in some groups than others, hence our proposed methods will account for that.

Potential applications

The Bayesian clustering tools we propose could possibly be handy tools in the single-cell RNA sequencing workflow. Current pipelines (`Seurat` R package, latest version Hao

et al. [57]) apply graph-based methods like the unsupervised Louvain algorithm (Blondel et al. [15]) in order to cluster cells based on their gene expression[1]. We suspect that our model-based methodologies could efficiently apply to these type of data (number of gene reads) with the extra advantage of allowing the allocation of a cell into all clusters with some probability rather than hard clustering into one group with no level of confidence. Moreover, each group of cells would be modelled by an estimated distribution and thus, this information would launch the selection process of those "meta-genes" (principal components of genes) that discriminate the groups.

Computational performance

For our clustering tools, we enhance the algorithm's convergence speed by adopting mainly a vectorised code scheme (bypass redundant iterative executions - "for loops"). To give a basic example, for a dataset of 10K samples and 1K features the variational method converges in less than 2 minutes compared to a less vectorised code that will add a few more minutes.

Code availability

On the subject of code availability, the reader can find the code in https://github.com/Lina-Ger/VBmixtures for most of the presented models, along with dataset simulators for testing and the corresponding discriminative accuracy forward selection algorithm. All the algorithms are implemented in R, however someone can straightforwardly translate the available code to any language of preference due to the clarity of the implementation.

6.3 Future Research Directions

With respect to future directions, we plan to develop a user-friendly R package for variational model-based clustering via mixture models for discrete and continuous distributions. This package will include all the proposed models together with the discriminative feature selection. The user will be merely responsible for feeding the function with the dataset and selecting the type of model (Beta/ Gaussian/Poisson Dirichlet Process mixture etc.). For better performance, the user will be allowed to tune the initial variational values and set the model hypeparameters (or proceed with the default).

On another direction, we would like to investigate further the results on the real neonate dataset in Chapter 5. Particularly, this would concern implementation of our clustering algorithms on larger datasets (in terms of more iDMRs as well as larger

[1]Instead of the actual genes, clustering is applied on the gene principal components (PCA), referred as "meta-genes".

cohorts of artificially and naturally conceived neonates) with the aim to reveal a stronger association of ART with imprinting disorders. Ideally, we would also wish to discover the existence of cluster-specific phenotypes that co-drive the segregation of the sub-populations apart from occurrence or not of the Beckwith-Wiedemann syndrome.

Furthermore, we would like to check the performance of our variational non-Gaussian Dirichlet Process mixture models with covariates, presented in Chapter 3, on DNA count or binary methylation data that are affected by confounding parameters (*i.e.*, datasets with both neonates and adults of different sex). Specifically, we would suggest to cluster the count methylation data by our "variational Dirichlet Process Poisson mixture with covariates" and the binary by our "variational Dirichlet Process Bernoulli mixture with covariates". Moreover, we aspire to cluster real data produced not only by Illumina platforms but by whole-genome bisulfite sequencing techniques. This implementation would concern our variational Dirichlet Process Binomial mixture model (presented in Appendix B).

A further interesting venue would be to study the problem of clustering data where $N << D$ while the features are highly correlated. In particular, we would try to parametrise the feature covariance matrices by latent factors - factor analysis is a method that accounts for the correlation in multi-dimensional data. This way the features would collapse into independent factors and we would be able to cluster the reduced feature-wise dataset based on the variational factor analysers mixtures proposed by Ghahramani and Beal [50].

Bibliography

[1] Affinito, O., Palumbo, D., Fierro, A., Cuomo, M., De Riso, G., Monticelli, A., Miele, G., Chiariotti, L., and Cocozza, S. (2020). Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics*, 112(1):144–150.

[2] Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

[3] Annunziato, A. (2008). DNA packaging: nucleosomes and chromatin. *Nature Education*, 1(1):26.

[4] Asratian, A. S., Denley, T. M., and Häggkvist, R. (1998). *Bipartite Graphs and their Applications*, volume 131. Cambridge University Press, New York.

[5] Baragatti, M. (2011). Bayesian variable selection for probit mixed models applied to gene selection. *Bayesian Analysis*, 6(2):209–229.

[6] Bartlett, M. S. and Kendall, D. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society*, 8(1):128–138.

[7] Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

[8] Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., et al. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295.

[9] Bierkens, J. and Roberts, G. (2017). A piecewise deterministic scaling limit of lifted Metropolis–Hastings in the Curie–Weiss model. *The Annals of Applied Probability*, 27(2):846–882.

[10] Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1):6–21.

[11] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

[12] Blei, D. M. (2011). Variational Inference. *https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf*.

[13] Blei, D. M. and Jordan, M. I. (2006). Variational Inference for Dirichlet Process mixtures. *Bayesian Analysis*, 1(1):121–143.

[14] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

[15] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[16] Bottolo, L., Banterle, M., Richardson, S., Ala-Korpela, M., Järvelin, M.-R., and Lewin, A. (2021). A computationally efficient Bayesian seemingly unrelated regressions model for high-dimensional quantitative trait loci discovery. *Journal of the Royal Statistical Society: Series C*.

[17] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York.

[18] Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B*, 60(3):627–641.

[19] Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 57(3):473–484.

[20] Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456.

[21] Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174.

[22] Cassidy, S. B., Schwartz, S., Miller, J. L., and Driscoll, D. J. (2012). Prader-willi syndrome. *Genetics in Medicine*, 14(1):10–26.

[23] Changyong, F., Hongyue, W., Naiji, L., Tian, C., Hua, H., Ying, L., et al. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, 26(2):105.

[24] Chargaff, E. (2012). *The Nucleic Acids*. Elsevier, London.

[25] Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.

[26] Clancy, S. (2008). DNA transcription. *Nature Education*, 1(1):41.

[27] Clancy, S. and Brown, W. (2008). Translation: DNA to mRNA to protein. *Nature Education*, 1(1):101.

[28] Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184):215–219.

[29] Corduneanu, A. and Bishop, C. M. (2001). Variational Bayesian model selection for mixture distributions. In *Artificial Intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA.

[30] Das, P. M. and Singal, R. (2004). DNA methylation and cancer. *Journal of Clinical Oncology*, 22(22):4632–4642.

[31] Dechter, R. and Pearl, J. (1988). Network-based heuristics for constraint-satisfaction problems. In *Search in Artificial Intelligence*, pages 370–425. Springer, New York.

[32] Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. (2011). Evaluation of the Infinium Methylation 450k technology. *Epigenomics*, 3(6):771–784.

[33] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38.

[34] Denison, D. G., Holmes, C. C., Mallick, B. K., and Smith, A. F. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley & Sons, New York.

[35] Dong, C., Yoon, W., and Goldschmidt-Clermont, P. J. (2002). DNA methylation and atherosclerosis. *The Journal of Nutrition*, 132(8):2406S–2409S.

[36] Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):1–9.

[37] Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38(12):1378–1385.

[38] El-Arini, K. (2008). Dirichlet Processes. *https://www.cs.cmu.edu/~kbe/dp_tutorial.pdf*.

[39] Espinheira, P. L., Ferrari, S. L., and Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, 35(4):407–419.

[40] Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes: A booming present, a brighter future. *Oncogene*, 21(35):5427–5440.

[41] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231.

[42] Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Reports*, 10(8):1386–1397.

[43] Feng, H., Conneely, K. N., and Wu, H. (2014). A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*, 42(8):e69–e69.

[44] Ferguson-Smith, A. C. (2011). Genomic imprinting: The emergence of an epigenetic paradigm. *Nature Reviews Genetics*, 12(8):565–575.

[45] Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

[46] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

[47] Fränti, P. and Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, 93:95–112.

[48] Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831.

[49] Gershman, S. J. and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

[50] Ghahramani, Z. and Beal, M. J. (1999). Variational Inference for Bayesian mixtures of factor analysers. In *NIPS*, volume 12, pages 449–455.

[51] Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359. Springer, New York.

[52] Gupta, A. K. and Nagar, D. K. (2018). *Matrix Variate Distributions*. Chapman and Hall/CRC, London.

[53] Hammond, C. M., Strømme, C. B., Huang, H., Patel, D. J., and Groth, A. (2017). Histone chaperone networks shaping chromatin function. *Nature Reviews Molecular Cell Biology*, 18(3):141–158.

[54] Han, K. (2019). DTC. *https://github.com/k-han/DTC/blob/master/utils/util.py*.

[55] Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516.

[56] Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216.

[57] Hao, Y., Hao, S., Andersen-Nissen, E., III, W. M. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. B., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*.

[58] Hattori, H., Hiura, H., Kitamura, A., Miyauchi, N., Kobayashi, N., Takahashi, S., Okae, H., Kyono, K., Kagami, M., Ogata, T., et al. (2019). Association of four imprinting disorders and ART. *Clinical Epigenetics*, 11(1):1–12.

[59] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

[60] Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168.

[61] Horváth, A. and Vértessy, B. G. (2010). A one-step method for quantitative determination of uracil in DNA by real-time PCR. *Nucleic Acids Research*, 38(21):e196–e196.

[62] Illingworth, R. S. and Bird, A. P. (2009). CpG islands–'a rough guide'. *FEBS Letters*, 583(11):1713–1720.

[63] Jabbari, K. and Bernardi, G. (2004). Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*, 333:143–149.

[64] Jain, S. and Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet Process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

[65] Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

[66] Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127.

[67] Kim, J. and Lee, J.-H. (2017). The validation of a Beta-binomial model for overdispersed binomial data. *Communications in Statistics-Simulation and Computation*, 46(2):807–814.

[68] Kim, S., Tadesse, M. G., and Vannucci, M. (2006). Variable selection in clustering via Dirichlet Process mixture models. *Biometrika*, 93(4):877–893.

[69] Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.

[70] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

[71] Lai, Y., Ping, Y., Xiao, K., Hao, B., and Zhang, X. (2018). Variational Bayesian Inference for a Dirichlet Process mixture of Beta distributions and application. *Neurocomputing*, 278:23–33.

[72] Lee, S., Huang, J. Z., and Hu, J. (2010). Sparse logistic principal components analysis for binary data. *The Annals of Applied Statistics*, 4(3):1579.

[73] Li, Y. and Turner, R. E. (2016). Rényi divergence Variational Inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081.

[74] Lim, U. and Song, M.-A. (2012). Dietary and lifestyle factors of DNA methylation. *Cancer Epigenetics*, pages 359–376.

[75] Lin, L. (2012). Bayesian variable selection in clustering and hierarchical mixture modeling. *Unpublished doctoral dissertation, https:// dukespace.lib.duke.edu/ dspace/ handle/ 10161/ 5846*.

[76] Lin, L., Chan, C., and West, M. (2016). Discriminative variable subsets in Bayesian classification with mixture models, with application in flow cytometry studies. *Biostatistics*, 17(1):40–53.

[77] Lind, M. I. and Spagopoulou, F. (2018). Evolutionary consequences of epigenetic inheritance.

[78] Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics 5*.

[79] Luo, Z.-Q. and Tseng, P. (1992). On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35.

[80] Ma, Z. and Leijon, A. (2011). Bayesian estimation of Beta mixture models with Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2160–2173.

[81] MacKay, D. J. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

[82] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland.

[83] Maksimovic, J., Oshlack, A., and Phipson, B. (2021). Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biology*, 22(1):1–26.

[84] Mandt, S., McInerney, J., Abrol, F., Ranganath, R., and Blei, D. (2016). Variational tempering. In *Artificial Intelligence and Statistics*, pages 704–712. PMLR.

[85] Mantovani, G., Elli, F., and Spada, A. (2012). GNAS epigenetic defects and pseudohypoparathyroidism: Time for a new classification? *Hormone and Metabolic Research*, 44(10):716–723.

[86] Marabita, F., Almgren, M., Lindholm, M. E., Ruhrmann, S., Fagerström-Billai, F., Jagodic, M., Sundberg, C. J., Ekström, T. J., Teschendorff, A. E., Tegnér, J., et al. (2013). An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics*, 8(3):333–346.

[87] Marutho, D., Hendra Handaka, S., Wijaya, E., and Muljono (2018). The determination of cluster number at K-mean using Elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 533–538.

[88] Mather, K., Mather, K., and Jinks, J. (1949). *Biometrical Genetics*, volume 162. Methuen London.

[89] McCullagh, P. and Nelder, J. A. (2019). *Generalized Linear Models*. Routledge, Milton Park.

[90] McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, volume 38. Marcel Dekker, New York.

[91] McLachlan, G. J. and Krishnan, T. (2007). *The EM Algorithm and Extensions*, volume 382. John Wiley & Sons, New York.

[92] McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and its Application*, 6:355–378.

[93] Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877.

[94] Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770.

[95] Miao, Y., Kook, J. H., Lu, Y., Guindani, M., and Vannucci, M. (2020). Scalable Bayesian variable selection regression models for count data. In *Flexible Bayesian Regression Modelling*, pages 187–219. Elsevier, London.

[96] Minka, T. (2005). Divergence measures and message passing. Technical report, Microsoft Research.

[97] Minka, T. P. (2013). Expectation propagation for approximate Bayesian inference. *arXiv preprint arXiv:1301.2294*.

[98] Monk, D., Morales, J., den Dunnen, J. T., Russo, S., Court, F., Prawitt, D., Eggermann, T., Beygo, J., Buiting, K., Tümer, Z., et al. (2018). Recommendations for a nomenclature system for reporting methylation aberrations in imprinted domains. *Epigenetics*, 13(2):117–121.

[99] Moore, L. D., Le, T., and Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38.

[100] Morris, K. V. (2012). *Non-coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection*. Horizon Scientific Press.

[101] Morris, T. J. and Beck, S. (2015). Analysis pipelines and packages for Infinium Humanmethylation450 Beadchip (450K) data. *Methods*, 72:3–8.

[102] Neal, R. M. (1993). *Probabilistic Inference using Markov Chain Monte Carlo methods*. Department of Computer Science, University of Toronto.

[103] Neidhart, M. (2015). *DNA methylation and complex human disease*. Academic Press.

[104] Nelder, J. A. and Baker, R. J. (2004). Generalized Linear Models. volume 4. Wiley Online Library, New York.

[105] Novakovic, B., Lewis, S., Halliday, J., Kennedy, J., Burgner, D. P., Czajko, A., Kim, B., Sexton-Oates, A., Juonala, M., Hammarberg, K., et al. (2019). Assisted reproductive technologies are associated with limited epigenetic variation at birth that largely resolves by adulthood. *Nature Communications*, 10(1):1–12.

[106] Ochoa, E., Lee, S., Lan-Leung, B., Dias, R. P., Ong, K. K., Radley, J. A., de Nanclares, G. P., Martinez, R., Clark, G., Martin, E., et al. (2021). Imprintseq, a novel tool to interrogate DNA methylation at human imprinted regions and diagnose multilocus imprinting disturbance. *Genetics in Medicine. https://doi.org/10.1016/j.gim.2021.10.011*.

[107] Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153.

[108] Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. (2014). MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, 30(17):2414–2422.

[109] Peterson, C. L. and Laniel, M.-A. (2004). Histones and histone modifications. *Current Biology*, 14(14):R546–R551.

[110] Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., and Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEpic BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1):1–17.

[111] Pitman, J. (2002). Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley.

[112] Plongthongkum, N., Diep, D. H., and Zhang, K. (2014). Advances in the profiling of DNA modifications: Cytosine methylation and beyond. *Nature Reviews Genetics*, 15(10):647–661.

[113] Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

[114] Prentice, R. L. (1974). A log Gamma model and its maximum likelihood estimation. *Biometrika*, 61(3):539–544.

[115] Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

[116] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B*, 59(4):731–792.

[117] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

[118] Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

[119] Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. *Frontiers in Genetics*, 5:324.

[120] Rockafellar, R. T. (1993). Lagrange multipliers and optimality. *SIAM Review*, 35(2):183–238.

[121] Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, 42(3):1–21.

[122] Scott, D. W. and Szewczyk, W. F. (2001). From kernels to mixtures. *Technometrics*, 43(3):323–335.

[123] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650.

[124] Smith, M. S., Loaiza-Maya, R., and Nott, D. J. (2020). High-dimensional copula variational approximation through transformation. *Journal of Computational and Graphical Statistics*, 29(4):729–743.

[125] Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

[126] Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41–45.

[127] Stunnenberg, H. G., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., Amin, V., Amit, I., Antonarakis, S. E., Aparicio, S., Arima, T., et al. (2016). The International Human Epigenome Consortium: A BLUEPRINT for scientific collaboration and discovery. *Cell*, 167(5):1145–1149.

[128] Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A., and Li, W. (2014). MOABS: Model based analysis of bisulfite sequencing data. *Genome Biology*, 15(2):1–12.

[129] Sun, Z., Chai, H. S., Wu, Y., White, W. M., Donkena, K. V., Klein, C. J., Garovic, V. D., Therneau, T. M., and Kocher, J.-P. A. (2011). Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC medical genomics*, 4(1):1–12.

[130] Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

[131] Tan, L. S. and Nott, D. J. (2018). Gaussian variational approximation with sparse precision matrices. *Statistics and Computing*, 28(2):259–275.

[132] Teh, Y. W. (2010). Dirichlet Process. *https://www.stats.ox.ac.uk/~teh/research/npbayes/Teh2010a.pdf*.

[133] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

[134] Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450K DNA methylation data. *Bioinformatics*, 29(2):189–196.

[135] Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Gayther, S. A., Apostolidou, S., Jones, A., Lechner, M., Beck, S., Jacobs, I. J., et al. (2009). An epigenetic signature in peripheral blood predicts active ovarian cancer. *PloS One*, 4(12):e8274.

[136] Tian, Y., Morris, T. J., Webster, A. P., Yang, Z., Beck, S., Feber, A., and Teschendorff, A. E. (2017). ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics*, 33(24):3982–3984.

[137] Titterington, D. et al. (2004). Bayesian methods for neural networks and related models. *Statistical Science*, 19(1):128–139.

[138] Titterington, D. M., Smith, A. F., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley & Sons, New York.

[139] Tjio, J. H. and Levan, A. (1956). The chromosome number of man. In *Problems of Birth Defects*, pages 112–118. Springer, New York.

[140] Vellai, T. and Vida, G. (1999). The origin of eukaryotes: The difference between prokaryotic and eukaryotic cells. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1428):1571–1577.

[141] Wang, D., Yan, L., Hu, Q., Sucheston, L. E., Higgins, M. J., Ambrosone, C. B., Johnson, C. S., Smiraglia, D. J., and Liu, S. (2012). IMA: an r package for high-throughput analysis of illumina's 450k infinium methylation data. *Bioinformatics*, 28(5):729–730.

[142] Wang, Z., Wu, X., and Wang, Y. (2018). A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. *BMC bioinformatics*, 19(5):15–22.

[143] Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

[144] Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., Yu, H., Jove, R., and Yuan, Y.-C. (2013). COHCAP: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis. *Nucleic acids research*, 41(11):e117–e117.

[145] Waterland, R. A. and Jirtle, R. L. (2003). Transposable elements: Targets for early nutritional effects on epigenetic gene regulation. *Molecular and Cellular Biology*, 23(15):5293–5300.

[146] Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738.

[147] Weksberg, R., Shuman, C., and Beckwith, J. B. (2010). Beckwith–Wiedemann syndrome. *European Journal of Human Genetics*, 18(1):8–14.

[148] Weksberg, R., Shuman, C., Caluseriu, O., Smith, A. C., Fei, Y.-L., Nishikawa, J., Stockley, T. L., Best, L., Chitayat, D., Olney, A., et al. (2002). Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith–Wiedemann syndrome. *Human Molecular Genetics*, 11(11):1317–1325.

[149] Wilhelm-Benartzi, C. S., Koestler, D. C., Karagas, M. R., Flanagan, J. M., Christensen, B. C., Kelsey, K. T., Marsit, C. J., Houseman, E. A., and Brown, R. (2013). Review of processing and analysis methods for DNA methylation array data. *British Journal of Cancer*, 109(6):1394–1402.

[150] Wilson, A. S., Power, B. E., and Molloy, P. L. (2007). DNA hypomethylation and human diseases. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1775(1):138–162.

[151] Wollmann, H., Kirchner, T., Enders, H., Preece, M., and Ranke, M. (1995). Growth and symptoms in Silver-Russell syndrome: Review on the basis of 386 patients. *European Journal of Pediatrics*, 154(12):958–968.

[152] Yang, X., Shao, X., Gao, L., and Zhang, S. (2016). Comparative DNA methylation analysis to decipher common and cell type-specific patterns among multiple cell types. *Briefings in Functional Genomics*, 15(6):399–407.

[153] Zegers-Hochschild, F., Adamson, G. D., de Mouzon, J., Ishihara, O., Mansour, R., Nygren, K., Sullivan, E., and Van der Poel, S. (2009). The International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) revised glossary on ART terminology, 2009. *Human Reproduction*, 24(11):2683–2687.

[154] Zeileis, A., Cribari-Neto, F., Gruen, B., Kosmidis, I., Simas, A. B., Rocha, A. V., and Zeileis, M. A. (2016). Package 'betareg'.

[155] Zhang, L., Meng, J., Liu, H., and Huang, Y. (2012). A nonparametric Bayesian approach for clustering bisulfate-based DNA methylation profiles. *BMC Genomics*, 13(6):1–9.

[156] Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biology*, 16(1):1–20.

[157] Zhou, W., Laird, P. W., and Shen, H. (2017). Comprehensive characterization, annotation and innovative use of Infinium DNA Methylation BeadChip probes. *Nucleic Acids Research*, 45(4):e22–e22.

[158] Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821.

Appendix A

A.1 Variational Lower Bound in Regression Models

We derive and present the Evidence Lower Bounds (ELBO), denoted as $\mathcal{L}(\boldsymbol{y}; q)$, for the variational regression models in Chapter 2, Section 2.5. Note that all the lower bounds are with respect to the corresponding variational parameters.

A.1.1 Single-response Linear Regression Model

The model is presented in subsection 2.5.1. The variational ELBO is

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}; q) = &-\frac{N}{2}\log(2\pi) - \frac{N}{2}\mathbb{E}_{\sigma^2}[\log\sigma^2] \\ &- \frac{1}{2}\frac{A_{q(\sigma^2)}}{B_{q(\sigma^2)}}\left\{\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}\right) + \operatorname{tr}\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\Sigma}_{q(\beta)}\right)\right\} \\ &- \frac{1}{2}\log|\boldsymbol{\Sigma}_\beta| - \frac{1}{2}\left(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta\right)^T\boldsymbol{\Sigma}_\beta^{-1}\left(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta\right) \\ &+ A\log B - \log\Gamma(A) - (A+1)\mathbb{E}_{\sigma^2}[\log\sigma^2] - B\frac{A_{q(\sigma^2)}}{B_{q(\sigma^2)}} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta)}| \\ &- A_{q(\sigma^2)}\log B_{q(\sigma^2)} + \log\Gamma(A_{q(\sigma^2)}) + (A_{q(\sigma^2)}+1)\mathbb{E}_{\sigma^2}[\log\sigma^2] + A_{q(\sigma^2)},\end{aligned} \tag{A.1}$$

where $\mathbb{E}_{\sigma^2}[\log\sigma^2] = \log B_{q(\sigma^2)} - \Psi(A_{q(\sigma^2)})$, since $\sigma^2 \sim \mathcal{IG}(A_{q(\sigma^2)}, B_{q(\sigma^2)})$.

A.1.2 Multi-response Linear Regression Model

The model is presented in subsection 2.5.1. The variational ELBO is

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}; q) = &-\frac{Nq}{2}\log(2\pi) - \frac{N}{2}\mathbb{E}_{\boldsymbol{\Sigma}}[\log|\boldsymbol{\Sigma}|] \\ &- \frac{1}{2}\operatorname{tr}\left\{(\nu_{\boldsymbol{\Sigma}} + N)\boldsymbol{Q}_{q(\boldsymbol{\Sigma})}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}\right)\right\}\end{aligned}$$

$$
\begin{aligned}
&- \frac{\nu_{\boldsymbol{\Sigma}} q}{2}\log(2\pi) - \log\Gamma_q(\frac{\nu_{\boldsymbol{\Sigma}}}{2}) - \frac{\nu_{\boldsymbol{\Sigma}}}{2}\log|\boldsymbol{Q}_{\boldsymbol{\Sigma}}| \\
&- \frac{\nu_{\boldsymbol{\Sigma}} + q + 1}{2}\mathbb{E}_{\boldsymbol{\Sigma}}[\log|\boldsymbol{\Sigma}|] \\
&- \frac{q}{2}\log|\boldsymbol{V}_\beta| - \frac{p}{2}|\boldsymbol{\Sigma}_\beta| \\
&- \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta)^T \boldsymbol{V}_\beta(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta)\right\} + \frac{q}{2}\log|\boldsymbol{V}_{q(\beta)}| + \frac{p}{2}|\boldsymbol{\Sigma}_{q(\beta)}| \\
&+ \frac{(\nu_{\boldsymbol{\Sigma}} + N)q}{2}\log(2\pi) + \log\Gamma_q(\frac{(\nu_{\boldsymbol{\Sigma}} + N)}{2}) + \frac{(\nu_{\boldsymbol{\Sigma}} + N)}{2}\log|\boldsymbol{Q}_{q(\boldsymbol{\Sigma})}| \\
&+ \frac{(\nu_{\boldsymbol{\Sigma}} + N) + q + 1}{2}\mathbb{E}_{\boldsymbol{\Sigma}}[\log|\boldsymbol{\Sigma}|],
\end{aligned}
\tag{A.2}
$$

where $\mathbb{E}_{\boldsymbol{\Sigma}}[\log|\boldsymbol{\Sigma}|] = \sum_{d=1}^{D}\Psi\left(\frac{\nu_{\boldsymbol{\Sigma}}+N+1-d}{2}\right) + D\ln 2 + \ln|\boldsymbol{Q}_{q(\boldsymbol{\Sigma})}|$, since $\boldsymbol{\Sigma} \sim \mathcal{IW}_q\left(\nu_{q(\boldsymbol{\Sigma})}, \boldsymbol{Q}_{q(\boldsymbol{\Sigma})}\right)$.

A.1.3 Linear Mixed Regression Model

The model is presented in subsection 2.5.2. The variational ELBO is

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{y}; q) = &-\frac{N}{2}\log(2\pi) - \frac{N}{2}\mathbb{E}_{\sigma_\epsilon^2}[\log\sigma_\epsilon^2] \\
&- \frac{A_\epsilon + \frac{N}{2}}{B_{q(\sigma_\epsilon^2)}}\frac{1}{2}\left\{\left(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta,\boldsymbol{u})}\right)^T\left(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta,\boldsymbol{u})}\right) + \mathrm{tr}\left(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\Sigma}_{q(\beta,\boldsymbol{u})}\right)\right\} \\
&- \frac{p}{2}\log(\sigma_\beta^2) - \frac{\sigma_\beta^{-2}}{2}\left\{\boldsymbol{\mu}_{q(\beta)}^T\boldsymbol{\mu}_{q(\beta)} + \mathrm{tr}\left(\boldsymbol{\Sigma}_{q(\beta)}\right)\right\} \\
&+ A_\epsilon\log B_\epsilon - \log\Gamma(A_\epsilon) - (A_\epsilon + 1)\mathbb{E}_{\sigma_\epsilon^2}[\log\sigma_\epsilon^2] - \frac{A_\epsilon + \frac{N}{2}}{B_{q(\sigma_\epsilon^2)}}B_\epsilon \\
&- \frac{1}{2}\sum_{l=1}^{r}K_l\mathbb{E}_{\sigma_{\boldsymbol{u}_l}^2}[\log(\sigma_{\boldsymbol{u}_l}^2)] - \sum_{l=1}^{r}\frac{A_{\boldsymbol{u}_l} + \frac{K_l}{2}}{B_{q(\sigma_{\boldsymbol{u}_l}^2)}}(B_{q(\sigma_{\boldsymbol{u}_l}^2)} - B_{\boldsymbol{u}_l}) \\
&+ \sum_{l=1}^{r}\left\{A_{\boldsymbol{u}_l}\log B_{\boldsymbol{u}_l} - \log\Gamma(A_{\boldsymbol{u}_l}) - (A_{\boldsymbol{u}_l} + 1)\mathbb{E}_{\sigma_{\boldsymbol{u}_l}^2}[\log\sigma_{\boldsymbol{u}_l}^2] - \frac{A_{\boldsymbol{u}_l} + \frac{K_l}{2}}{B_{q(\sigma_{\boldsymbol{u}_l}^2)}}B_{\boldsymbol{u}_l}\right\} \\
&+ \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta,\boldsymbol{u})}| \\
&- (A_\epsilon + \frac{N}{2})\log B_{q(\epsilon)} + \log\Gamma(A_\epsilon + \frac{N}{2}) + (A_\epsilon + \frac{N}{2} + 1)\mathbb{E}_{\sigma_\epsilon^2}[\log\sigma_\epsilon^2] + (A_\epsilon + \frac{N}{2}) \\
&+ \sum_{l=1}^{r}\left\{-A_{q(\sigma_{\boldsymbol{u}_l}^2)}\log B_{q(\sigma_{\boldsymbol{u}_l}^2)} + \log\Gamma(A_{q(\sigma_{\boldsymbol{u}_l}^2)}) + (A_{q(\sigma_{\boldsymbol{u}_l}^2)} + 1)\mathbb{E}_{\sigma^2_{q(\sigma_{\boldsymbol{u}_l}^2)}}[\log\sigma^2_{q(\sigma_{\boldsymbol{u}_l}^2)}]\right. \\
&\left. + A_{q(\sigma_{\boldsymbol{u}_l}^2)} + \frac{K_l}{2}\right\},
\end{aligned}
\tag{A.3}
$$

where $\mathbb{E}_{\sigma_\epsilon^2}[\log\sigma_\epsilon^2] = \log B_{q(\sigma_\epsilon^2)} - \Psi(A_{q(\sigma_\epsilon^2)})$, since $\sigma_\epsilon^2 \sim \mathcal{IG}(A_{q(\sigma_\epsilon^2)}, B_{q(\sigma_\epsilon^2)})$
and $\mathbb{E}_{\sigma_{\boldsymbol{u}_l}^2}[\log\sigma_{\boldsymbol{u}_l}^2] = \log B_{q(\sigma_{\boldsymbol{u}_l}^2)} - \Psi(A_{q(\sigma_{\boldsymbol{u}_l}^2)})$, since $\sigma_{\boldsymbol{u}_l}^2 \sim \mathcal{IG}(A_{q(\sigma_{\boldsymbol{u}_l}^2)}, B_{q(\sigma_{\boldsymbol{u}_l}^2)})$.

A.1.4 Probit Regression Model

The model is presented in subsection 2.5.3. The variational ELBO is

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}; q) = \boldsymbol{y}^T \log \Phi(\boldsymbol{X}\boldsymbol{\mu}_{q(\beta)}) + (\mathbf{1}_N - \boldsymbol{y})^T \log(\mathbf{1}_N - \Phi(\boldsymbol{X}\boldsymbol{\mu}_{q(\beta)})) \\ - \frac{1}{2}(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta)^T \boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta) + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta)}| - \frac{1}{2}\log|\boldsymbol{\Sigma}_\beta|.\end{aligned} \tag{A.4}$$

A.1.5 Probit Mixed Regression Model

The model is presented in subsection 2.5.4. The variational ELBO is

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}; q) = \boldsymbol{y}^T \log \Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta,\boldsymbol{u})}) + (\mathbf{1}_N - \boldsymbol{y})^T \log(\mathbf{1}_N - \Phi(\boldsymbol{C}\boldsymbol{\mu}_{q(\beta,\boldsymbol{u})})) \\ - \frac{1}{2}(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta)^T \boldsymbol{\Sigma}_\beta^{-1}(\boldsymbol{\mu}_{q(\beta)} - \boldsymbol{\mu}_\beta) + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta,\boldsymbol{u})}| - \frac{1}{2}\log|\boldsymbol{\Sigma}_\beta| \\ + \sum_{l=1}^{r}\left\{A_{\boldsymbol{u}_l}\log B_{\boldsymbol{u}_l} - \log\Gamma(A_{\boldsymbol{u}_l}) - \left(A_{\boldsymbol{u}_l} + \frac{K_l}{2}\right)\log B_{q(\sigma^2_{\boldsymbol{u}_l})} + \log\Gamma\left(A_{\boldsymbol{u}_l} + \frac{K_l}{2}\right)\right\} \\ - \frac{N}{2}\log 2\pi - \frac{1}{2}\left(\boldsymbol{\mu}_{q(\boldsymbol{z})} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta,\boldsymbol{u})}\right)^T\left(\boldsymbol{\mu}_{q(\boldsymbol{z})} - \boldsymbol{C}\boldsymbol{\mu}_{q(\beta,\boldsymbol{u})}\right).\end{aligned} \tag{A.5}$$

Appendix B

In this appendix, we present the variational derivation of the Finite and Dirichlet Process mixtures, as indicated in Chapter 3, Table 3.1. In particular, we exhibit the priors and the log-prior distributions and then we derive the variational distributions, along with their logarithmic version (necessary for the ELBO calculation). In all the subsequent models, $\boldsymbol{y}$ is the $N \times D$ matrix of observations, where N is the number of samples and D the number of features. The model parameters for each model are either $D \times M$ matrices, denoted by the dm subscript, or M vectors by the m subscript. M implies the number of components (this M is fixed and specified in the Finite mixture models or is a truncated number in the Dirichlet Process mixture models due to the stick-breaking point assumption). We also supply code snippets at the end for reader's reference.

B.1 Mean Field Finite Mixture Models

Here, we provide only the equations for the variational Finite Poisson mixture to avoid repetition, since the rest Finite mixtures in Table 3.1 are denoted as "easy to derive".

B.1.1 Variational Finite Poisson Mixture

$$\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{z} \sim \prod_{n=1}^{N} \prod_{d=1}^{D} \prod_{m=1}^{M} \left[\text{Poisson}(y_{nd}|\lambda_{dm})\right]^{z_{nm}} . \tag{B.1}$$

The likelihood is

$$P(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{z}) = \prod_{n=1}^{N} \prod_{d=1}^{D} \prod_{m=1}^{M} \left[\frac{{\lambda_{dm}}^{y_{nd}}}{y_{nd}!} \exp(-\lambda_{dm}) \right]^{z_{nm}} . \tag{B.2}$$

The log-likelihood is

$$\log P(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{z}) = \sum_{n=1}^{N} \sum_{d=1}^{D} \sum_{m=1}^{M} z_{nm} \left[y_{nd} \log \lambda_{dm} - \lambda_{dm} - \log y_{nd}! \right] . \tag{B.3}$$

The Categorical prior on the latent allocation is

$$P(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{m=1}^{M}\pi_m^{z_{nm}}. \tag{B.4}$$

The log-Categorical prior is

$$\log P(\boldsymbol{z}|\boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log\pi_m. \tag{B.5}$$

The Dirichlet prior on the mixing weights is

$$P(\boldsymbol{\pi}) = C^{-1}(\boldsymbol{\phi_0})\prod_{m=1}^{M}\pi_m^{(\phi_{0_m}-1)}, \quad \text{with} \quad C^{-1}(\boldsymbol{\phi_0}) = \frac{\Gamma\left(\sum_{m=1}^{M}\phi_{0m}\right)}{\prod_{m=1}^{M}\Gamma(\phi_{0m})}. \tag{B.6}$$

The log-Dirichlet prior is

$$\log P(\boldsymbol{\pi}) = -\log C(\boldsymbol{\phi_0}) + \sum_{m=1}^{M}(\phi_{0_m}-1)\log\pi_m. \tag{B.7}$$

The Gamma prior on the model parameter is

$$P(\boldsymbol{\lambda}) = \prod_{d=1}^{D}\prod_{m=1}^{M}\frac{{b_{0dm}}^{a_{0dm}}}{\Gamma(a_{0dm})}{\lambda_{dm}}^{(a_{0dm}-1)}\exp(-b_{0dm}\lambda_{dm}). \tag{B.8}$$

The log-Gamma prior is

$$\log P(\boldsymbol{\lambda}) = \sum_{d=1}^{D}\sum_{m=1}^{M}\left[a_{0dm}\log b_{0dm} - \log\Gamma(a_{0dm}) + (a_{0dm}-1)\log\lambda_{dm} - b_{0dm}\lambda_{dm}\right]. \tag{B.9}$$

The variational derivation of the latent allocation is

$$\begin{aligned}
\log q(\boldsymbol{z}) &\propto \mathbb{E}_{/\boldsymbol{z}}\left[\log P(\boldsymbol{y}|\boldsymbol{\lambda},\boldsymbol{z}) + \log P(\boldsymbol{z}|\boldsymbol{\pi})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{z}}\left[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{y_{nd}\log\lambda_{dm} - \lambda_{dm} - \log y_{nd}!\right\} + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log\pi_m\right] \\
&= \sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{y_{nd}\mathbb{E}_{\lambda_{dm}}\left[\log\lambda_{dm}\right] - \mathbb{E}_{\lambda_{dm}}\left[\lambda_{dm}\right] - \log y_{nd}!\right\} \\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\mathbb{E}_{\pi_m}\left[\log\pi_m\right] \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\sum_{d=1}^{D} y_{nd}\mathbb{E}_{\lambda_{dm}}\left[\log\lambda_{dm}\right] - \sum_{d=1}^{D}\mathbb{E}_{\lambda_{dm}}\left[\lambda_{dm}\right] - \sum_{d=1}^{D}\log y_{nd}!\right. \\
&\quad \left. + \mathbb{E}_{\pi_m}\left[\log\pi_m\right]\right\}.
\end{aligned} \tag{B.10}$$

Equation (B.10) reminds the logarithmic kernel of a Categorical density after we set the expression inside the brackets as $\log\rho_{nm}$

$$\log\rho_{nm} = \sum_{d=1}^{D} y_{nd}\mathbb{E}_{\lambda_{dm}}\left[\log\lambda_{dm}\right] - \sum_{d=1}^{D}\mathbb{E}_{\lambda_{dm}}\left[\lambda_{dm}\right] + \mathbb{E}_{\pi_m}\left[\log\pi_m\right], \tag{B.11}$$

concluding with

$$\log q(\boldsymbol{z}) \propto \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log\rho_{nm}. \tag{B.12}$$

In order for $q(\boldsymbol{z})$ to be equal and not just proportional to a Categorical density the variational parameter of $\boldsymbol{z}$ (here will be denoted as $\boldsymbol{r}$) should be constrained to belong in the $[0,1]$ interval and have $\sum_{m=1}^{M} r_{nm} = 1$. Hence, $r_{nm} = \rho_{nm} / \sum_{j=1}^{M} \rho_{nj}$. The variational log-Categorical density is

$$\log q(\boldsymbol{z}) = \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log r_{nm}. \tag{B.13}$$

The variational Categorical density is

$$q(\boldsymbol{z}) = \prod_{n=1}^{N} \prod_{m=1}^{M} {r_{nm}}^{z_{nm}}. \tag{B.14}$$

The variational derivation of the mixing weights is

$$\begin{aligned} \log q(\boldsymbol{\pi}) &\propto \mathbb{E}_{/\boldsymbol{\pi}} \left[\log P(\boldsymbol{z}|\boldsymbol{\pi}) + \log P(\boldsymbol{\pi}) \right] \\ &\propto \mathbb{E}_{/\boldsymbol{\pi}} \left[\sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log \pi_m + \sum_{m=1}^{M} (\phi_{0_m} - 1) \log \pi_m \right] \\ &= \sum_{m=1}^{M} \left(\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] + \phi_{0_m} - 1 \right) \log \pi_m. \end{aligned} \tag{B.15}$$

Equation (B.15) is the kernel of a Log-Dirichlet density if we set the parenthesis expression, except from the -1 term, as ϕ_m

$$\phi_m = \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] + \phi_{0_m}. \tag{B.16}$$

The variational log-Dirichlet density is

$$\log q(\boldsymbol{\pi}) = -\log C(\boldsymbol{\phi}) + \sum_{m=1}^{M} (\phi_m - 1) \log \pi_m. \tag{B.17}$$

The variational Dirichlet density is

$$q(\boldsymbol{\pi}) = C^{-1}(\boldsymbol{\phi}) \prod_{m=1}^{M} \pi_m^{(\phi_m - 1)}. \tag{B.18}$$

The variational derivation of the model parameter is

$$\begin{aligned} \log q(\boldsymbol{\lambda}) &\propto \mathbb{E}_{/\boldsymbol{\lambda}} \left[\log P(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{z}) + \log P(\boldsymbol{\lambda}) \right] \\ &\propto \mathbb{E}_{/\boldsymbol{\lambda}} \Bigg[\sum_{n=1}^{N} \sum_{d=1}^{D} \sum_{m=1}^{M} z_{nm} \left\{ y_{nd} \log \lambda_{dm} - \lambda_{dm} \right\} + \sum_{d=1}^{D} \sum_{m=1}^{M} \Big\{ a_{0dm} \log b_{0dm} \\ &\quad - \log \Gamma(a_{0dm}) + (a_{0dm} - 1) \log \lambda_{dm} - b_{0dm} \lambda_{dm} \Big\} \Bigg] \\ &= \sum_{m=1}^{M} \sum_{d=1}^{D} \Bigg\{ \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] y_{nd} \log \lambda_{dm} - \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] \lambda_{dm} \\ &\quad + (a_{0dm} - 1) \log \lambda_{dm} - b_{0dm} \lambda_{dm} \Bigg\} \\ &= \sum_{m=1}^{M} \sum_{d=1}^{D} \left\{ \left(\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] y_{nd} + a_{0dm} - 1 \right) \log \lambda_{dm} \right. \end{aligned}$$

$$- \left(\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] + b_{0dm} \right) \lambda_{dm} \Bigg\}. \tag{B.19}$$

Equation (B.19) is the kernel of a Log-Gamma density with

$$\begin{aligned} a_{dm} &= a_{0dm} + \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] y_{nd}, \\ b_{dm} &= b_{0dm} + \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}]. \end{aligned} \tag{B.20}$$

The variational log-Gamma density is

$$\log q(\boldsymbol{\lambda}) = \sum_{d=1}^{D} \sum_{m=1}^{M} \left[a_{dm} \log b_{dm} - \log \Gamma(a_{dm}) + (a_{dm} - 1) \log \lambda_{dm} - b_{dm} \lambda_{dm} \right]. \tag{B.21}$$

The variational Gamma density is

$$P(\boldsymbol{\lambda}) = \prod_{d=1}^{D} \prod_{m=1}^{M} \frac{{b_{dm}}^{a_{dm}}}{\Gamma(a_{dm})} {\lambda_{dm}}^{(a_{dm}-1)} \exp(-b_{dm} \lambda_{dm}). \tag{B.22}$$

The Evidence Lower Bound (ELBO) is

$$\begin{aligned} L(\boldsymbol{y}; q) &= \mathbb{E}_{\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\lambda}}[\log P(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{\pi})] - \mathbb{E}_{\boldsymbol{z}}[\log q(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{\pi})] - \mathbb{E}_{\boldsymbol{\lambda}}[\log q(\boldsymbol{\lambda})] \\ &= \mathbb{E}_{\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\lambda}}[\log P(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\lambda}) + \log P(\boldsymbol{z}|\boldsymbol{\pi}) + \log P(\boldsymbol{\pi}) + \log P(\boldsymbol{\lambda})] \\ &\quad - \mathbb{E}_{\boldsymbol{z}}[\log q(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{\pi})] - \mathbb{E}_{\boldsymbol{\lambda}}[\log q(\boldsymbol{\lambda})]. \end{aligned} \tag{B.23}$$

The explicit ELBO form is

$$\begin{aligned} L(\boldsymbol{y}; q) = &\sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{E}_{z_{nm}}[z_{nm}] \log \rho_{nm} - \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{E}_{z_{nm}}[z_{nm}] \log r_{nm} - \sum_{n=1}^{N} \sum_{d=1}^{D} \log y_{nd}! \\ &+ \log C(\boldsymbol{\phi}) - \log C(\boldsymbol{\phi_0}) + \sum_{m=1}^{M} (\phi_{0m} - \phi_m) \mathbb{E}_{\pi_m}[\log \pi_m] \\ &+ \sum_{d=1}^{D} \sum_{m=1}^{M} \Bigg[a_{0dm} \log b_{0dm} - \log \Gamma(a_{0dm}) + (a_{0dm} - a_{dm}) \mathbb{E}_{\lambda_{dm}}[\log \lambda_{dm}] \\ &- a_{dm} \log b_{dm} + \log \Gamma(a_{dm}) + (b_{dm} - b_{0dm}) \mathbb{E}_{\lambda_{dm}}[\lambda_{dm}] \Bigg]. \end{aligned} \tag{B.24}$$

The variational expectations and posterior estimates of the mixing weights are

$$\begin{aligned} \mathbb{E}_{z_{nm}}[z_{nm}] &= r_{nm}, \\ \mathbb{E}_{\pi_m}[\log \pi_m] &= \Psi(\phi_m) - \Psi\left(\sum_{m=1}^{M} \phi_m \right), \\ \mathbb{E}_{\lambda_{dm}}[\log \lambda_{dm}] &= \Psi(a_{dm}) - \log b_{dm}, \\ \mathbb{E}_{\lambda_{dm}}[\lambda_{dm}] &= \frac{a_{dm}}{b_{dm}}, \\ \pi_m &= \frac{\phi_{0m} + \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}]}{M \phi_{0m} + N}. \end{aligned} \tag{B.25}$$

B.2 Mean Field Dirichlet Process Mixture Models

Here, we present the variational Dirichlet Process mixture models of Table 3.1. In particular, the variational Dirichlet Process Poisson mixture model and the variational Dirichlet Process Binomial mixture model. The variational Dirichlet Process Gaussian mixture with independent features is given in R code version.

B.2.1 Variational Dirichlet Process Poisson Mixture

$$\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{z} \sim \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{m=1}^{M}\left[\text{Poisson}(y_{nd}|\lambda_{dm})\right]^{z_{nm}}. \tag{B.26}$$

The likelihood is

$$P(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{z}) = \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{m=1}^{M}\left[\frac{{\lambda_{dm}}^{y_{nd}}}{y_{nd}!}\exp(-\lambda_{dm})\right]^{z_{nm}}. \tag{B.27}$$

The log-likelihood is

$$\log P(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{z}) = \sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left[y_{nd}\log\lambda_{dm} - \lambda_{dm} - \log y_{nd}!\right]. \tag{B.28}$$

The Categorical prior on the latent allocation (stick-breaking point representation) is

$$P(\boldsymbol{z}|\boldsymbol{w}) = \prod_{n=1}^{N}\prod_{m=1}^{M}\left[w_m\prod_{j=1}^{m-1}(1-w_j)\right]^{z_{nm}}. \tag{B.29}$$

The log-Categorical prior is

$$\log P(\boldsymbol{z}|\boldsymbol{w}) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left[\log w_m + \sum_{j=1}^{m-1}\log(1-w_j)\right]. \tag{B.30}$$

The Beta prior on the mixing weights is

$$P(\boldsymbol{w}) = \prod_{m=1}^{M}\phi_{0m}(1-w_m)^{(\phi_{0m}-1)}. \tag{B.31}$$

The log-Beta prior is

$$\log P(\boldsymbol{w}) = \sum_{m=1}^{M}\left[\log\phi_{0m} + (\phi_{0m}-1)\log(1-w_m)\right]. \tag{B.32}$$

The Gamma prior on the model parameter is

$$P(\boldsymbol{\lambda}) = \prod_{d=1}^{D}\prod_{m=1}^{M}\frac{{b_{0dm}}^{a_{0dm}}}{\Gamma(a_{0dm})}{\lambda_{dm}}^{(a_{0dm}-1)}\exp(-b_{0dm}\lambda_{dm}). \tag{B.33}$$

The log-Gamma prior is

$$\log P(\boldsymbol{\lambda}) = \sum_{d=1}^{D}\sum_{m=1}^{M}\left[a_{0dm}\log b_{0dm} - \log\Gamma(a_{0dm}) + (a_{0dm}-1)\log\lambda_{dm} - b_{0dm}\lambda_{dm}\right]. \tag{B.34}$$

The variational derivation of the latent allocation is

$$
\begin{aligned}
\log q(\boldsymbol{z}) &\propto \mathbb{E}_{/\boldsymbol{z}}\left[\log P(\boldsymbol{y}|\boldsymbol{\lambda},\boldsymbol{z}) + \log P(\boldsymbol{z}|\boldsymbol{w})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{z}}\Bigg[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{y_{nd}\log\lambda_{dm} - \lambda_{dm} - \log y_{nd}!\right\} \\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\log w_m + \sum_{j=1}^{m-1}\log(1-w_j)\right\}\Bigg] \\
&= \sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{y_{nd}\mathbb{E}_{\lambda_{dm}}\left[\log\lambda_{dm}\right] - \mathbb{E}_{\lambda_{dm}}\left[\lambda_{dm}\right] - \log y_{nd}!\right\} \\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\mathbb{E}_{w_m}\left[\log w_m\right] + \sum_{j=1}^{m-1}\mathbb{E}_{w_m}\left[\log(1-w_j)\right]\right\} \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\Bigg\{\sum_{d=1}^{D} y_{nd}\mathbb{E}_{\lambda_{dm}}\left[\log\lambda_{dm}\right] - \sum_{d=1}^{D}\mathbb{E}_{\lambda_{dm}}\left[\lambda_{dm}\right] - \sum_{d=1}^{D}\log y_{nd}! \\
&\quad + \mathbb{E}_{w_m}\left[\log w_m\right] + \sum_{j=1}^{m-1}\mathbb{E}_{w_m}\left[\log(1-w_j)\right]\Bigg\}.
\end{aligned}
\tag{B.35}
$$

Equation (B.35) reminds the logarithmic kernel of a Categorical density after we set the expression inside the brackets as $\log\rho_{nm}$ (excluding $\log y$)

$$
\log\rho_{nm} = \sum_{d=1}^{D} y_{nd}\mathbb{E}_{\lambda_{dm}}\left[\log\lambda_{dm}\right] - \sum_{d=1}^{D}\mathbb{E}_{\lambda_{dm}}\left[\lambda_{dm}\right] + \mathbb{E}_{w_m}\left[\log w_m\right] + \sum_{j=1}^{m-1}\mathbb{E}_{w_m}\left[\log(1-w_j)\right], \tag{B.36}
$$

concluding with

$$
\log q(\boldsymbol{z}) \propto \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log\rho_{nm}. \tag{B.37}
$$

In order for $q(\boldsymbol{z})$ to be equal and not just proportional to a Categorical density, the variational parameter of $\boldsymbol{z}$ will be constrained to belong in the $[0,1]$ interval and have $\sum_{m=1}^{M} r_{nm} = 1$. Hence, $r_{nm} = \rho_{nm}/\sum_{j=1}^{M}\rho_{nj}$.

The variational log-Categorical density is

$$
\log q(\boldsymbol{z}) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\log r_{nm}. \tag{B.38}
$$

The variational Categorical density is

$$
q(\boldsymbol{z}) = \prod_{n=1}^{N}\prod_{m=1}^{M} {r_{nm}}^{z_{nm}}. \tag{B.39}
$$

The variational derivation of the mixing weights is

$$
\begin{aligned}
\log q(\boldsymbol{w}) &\propto \mathbb{E}_{/\boldsymbol{w}}\left[\log P(\boldsymbol{z}|\boldsymbol{w}) + \log P(\boldsymbol{w})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{w}}\left[\sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\log w_m + \sum_{j=1}^{m-1}\log(1-w_j)\right\} + \sum_{m=1}^{M}(\phi_{0m}-1)\log(1-w_m)\right] \\
&= \sum_{m=1}^{M}\left\{\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]\log w_m + \left(\sum_{n=1}^{N}\sum_{j=m+1}^{M}\mathbb{E}_{\boldsymbol{z}_{nj}}[z_{nj}] + \phi_{0m} - 1\right)\log(1-w_m)\right\}.
\end{aligned}
\tag{B.40}
$$

Equation (B.40) reminds the kernel of a Log-Beta density if we set

$$\begin{aligned}\delta_m &= 1 + \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}], \\ \phi_m &= \phi_{0m} + \sum_{n=1}^{N} \sum_{j=m+1}^{M} \mathbb{E}_{\boldsymbol{z}_{nj}}[z_{nj}].\end{aligned} \tag{B.41}$$

The variational log-Beta density is

$$\begin{aligned}\log q(\boldsymbol{w}) = \sum_{m=1}^{M} \Big\{ &\log\Gamma(\delta_m + \phi_m) - \log\Gamma(\delta_m) - \log\Gamma(\phi_m) + (\delta_m - 1)\log w_m \\ &+ (\phi_m - 1)\log(1 - w_m) \Big\}.\end{aligned} \tag{B.42}$$

The variational Beta density is

$$q(\boldsymbol{w}) = \prod_{m=1}^{M} \frac{\Gamma(\delta_m + \phi_m)}{\Gamma(\delta_m)\Gamma(\phi_m)} {w_m}^{(\delta_m - 1)} (1 - w_m)^{(\phi_1 - 1)}. \tag{B.43}$$

The variational derivation of the model parameter is

$$\begin{aligned}\log q(\boldsymbol{\lambda}) &\propto \mathbb{E}_{/\boldsymbol{\lambda}}\left[\log P(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{z}) + \log P(\boldsymbol{\lambda})\right] \\ &\propto \mathbb{E}_{/\boldsymbol{\lambda}}\Bigg[\sum_{n=1}^{N} \sum_{d=1}^{D} \sum_{m=1}^{M} z_{nm} \left\{ y_{nd} \log \lambda_{dm} - \lambda_{dm} \right\} + \sum_{d=1}^{D} \sum_{m=1}^{M} \Big\{ a_{0dm} \log b_{0dm} \\ &\quad - \log\Gamma(a_{0dm}) + (a_{0dm} - 1)\log\lambda_{dm} - b_{0dm}\lambda_{dm} \Big\} \Bigg] \\ &= \sum_{m=1}^{M} \sum_{d=1}^{D} \Bigg\{ \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] x_{nd} \log\lambda_{dm} - \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] \lambda_{dm} \\ &\quad + (a_{0dm} - 1)\log\lambda_{dm} - b_{0dm}\lambda_{dm} \Bigg\} \\ &= \sum_{m=1}^{M} \sum_{d=1}^{D} \Bigg\{ \left(\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] y_{nd} + a_{0dm} - 1 \right) \log\lambda_{dm} \\ &\quad - \left(\sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] + b_{0dm} \right) \lambda_{dm} \Bigg\}.\end{aligned} \tag{B.44}$$

Equation (B.44) is the kernel of a Log-Gamma density if we set the two parentheses as

$$\begin{aligned}a_{dm} &= a_{0dm} + \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}] x_{nd}, \\ b_{dm} &= b_{0dm} + \sum_{n=1}^{N} \mathbb{E}_{z_{nm}}[z_{nm}].\end{aligned} \tag{B.45}$$

The variational log-Gamma density is

$$\log q(\boldsymbol{\lambda}) = \sum_{d=1}^{D} \sum_{m=1}^{M} \left[a_{dm} \log b_{dm} - \log\Gamma(a_{dm}) + (a_{dm} - 1)\log\lambda_{dm} - b_{dm}\lambda_{dm} \right]. \tag{B.46}$$

The variational Gamma density is

$$P(\boldsymbol{\lambda}) = \prod_{d=1}^{D} \prod_{m=1}^{M} \frac{{b_{dm}}^{a_{dm}}}{\Gamma(a_{dm})} {\lambda_{dm}}^{(a_{dm} - 1)} \exp(-b_{dm}\lambda_{dm}). \tag{B.47}$$

The Evidence Lower Bound (ELBO) is

$$\begin{aligned} L(\boldsymbol{y}; q) =& \mathbb{E}_{\boldsymbol{z},\boldsymbol{w},\boldsymbol{\lambda}}[\log P(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{w})] - \mathbb{E}_{\boldsymbol{z}}[\log q(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{w})] - \mathbb{E}_{\boldsymbol{\lambda}}[\log q(\boldsymbol{\lambda})] \\ =& \mathbb{E}_{\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{\lambda}}[\log P(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\lambda}) + \log P(\boldsymbol{z}|\boldsymbol{w}) + \log P(\boldsymbol{w}) + \log P(\boldsymbol{\lambda})] \\ & - \mathbb{E}_{\boldsymbol{z}}[\log q(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{w})] - \mathbb{E}_{\boldsymbol{\lambda}}[\log q(\boldsymbol{\lambda})]. \end{aligned} \tag{B.48}$$

The explicit ELBO form is

$$\begin{aligned} L(\boldsymbol{y}; q) =& \sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{E}_{z_{nm}}[z_{nm}] \log \rho_{nm} - \sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{E}_{z_{nm}}[z_{nm}] \log r_{nm} - \sum_{n=1}^{N}\sum_{d=1}^{D} \log y_{nd}! \\ & + \sum_{m=1}^{M} [\log \phi_{0m} + (\phi_{0m} - 1) \log(1 - w_m)] \\ & - \sum_{m=1}^{M} \Big\{ \log \Gamma(\delta_m + \phi_m) - \log \Gamma(\delta_m) - \log \Gamma(\phi_m) + (\delta_m - 1) \log w_m \\ & + (\phi_m - 1) \log(1 - w_m) \Big\} \\ & + \sum_{d=1}^{D}\sum_{m=1}^{M} \Big[a_{0dm} \log b_{0dm} - \log \Gamma(a_{0dm}) + (a_{0dm} - a_{dm}) \mathbb{E}_{\lambda_{dm}}[\log \lambda_{dm}] \\ & - a_{dm} \log b_{dm} + \log \Gamma(a_{dm}) + (b_{dm} - b_{0dm}) \mathbb{E}_{\lambda_{dm}}[\lambda_{dm}] \Big]. \end{aligned} \tag{B.49}$$

The variational expectations and posterior estimates of the mixing weights are

$$\begin{aligned} \mathbb{E}_{z_{nm}}[z_{nm}] &= r_{nm}, \\ \mathbb{E}_{w_m}[\log w_m] &= \Psi(\phi_m) - \Psi(\delta_m + \phi_m), \\ \mathbb{E}_{w_m}[\log(1 - w_m)] &= \Psi(\delta_m) - \Psi(\delta_m + \phi_m), \\ \mathbb{E}_{\lambda_{dm}}[\log \lambda_{dm}] &= \Psi(a_{dm}) - \log b_{dm}, \\ \mathbb{E}_{\lambda_{dm}}[\lambda_{dm}] &= \frac{a_{dm}}{b_{dm}}, \\ \pi_m &= \mathbb{E}[w_m] \prod_{j=1}^{m-1} \big(1 - \mathbb{E}[w_j]\big). \end{aligned} \tag{B.50}$$

B.2.2 Variational Dirichlet Process Binomial Mixture

$$\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{z} \sim \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{m=1}^{M} [\text{Binomial}(y_{nd}|s_{nd}, p_{dm})]^{z_{nm}}. \tag{B.51}$$

The likelihood is

$$P(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{z}) = \prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{m=1}^{M} \left[\binom{s_{nd}}{y_{nd}} (p_{dm})^{y_{nd}} (1 - p_{dm})^{(s_{nd} - y_{nd})} \right]^{z_{nm}}. \tag{B.52}$$

The log-likelihood is

$$\log P(\boldsymbol{y}|\boldsymbol{p},\boldsymbol{z}) = \sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left[\log\binom{s_{nd}}{y_{nd}} + y_{nd}\log p_{dm} + (s_{nd}-y_{nd})\log(1-p_{dm})\right]. \tag{B.53}$$

The Categorical prior on the latent allocation (stick-breaking point representation) is

$$P(\boldsymbol{z}|\boldsymbol{w}) = \prod_{n=1}^{N}\prod_{m=1}^{M}\left[w_m\prod_{j=1}^{m-1}(1-w_j)\right]^{z_{nm}}. \tag{B.54}$$

The log-Categorical prior is

$$\log P(\boldsymbol{z}|\boldsymbol{w}) = \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left[\log w_m + \sum_{j=1}^{m-1}\log(1-w_j)\right]. \tag{B.55}$$

The Beta prior on the mixing weights is

$$P(\boldsymbol{w}) = \prod_{m=1}^{M}\phi_{0m}(1-w_m)^{(\phi_{0m}-1)}. \tag{B.56}$$

The log-Beta prior is

$$\log P(\boldsymbol{w}) = \sum_{m=1}^{M}\left[\log\phi_{0m} + (\phi_{0m}-1)\log(1-w_m)\right]. \tag{B.57}$$

The Beta prior on the model parameter is

$$P(\boldsymbol{p}) = \prod_{d=1}^{D}\prod_{m=1}^{M}\frac{\Gamma(a_{0dm}+b_{0dm})}{\Gamma(a_{0dm})\Gamma(b_{0dm})}{p_{dm}}^{(a_{0dm}-1)}(1-p_m)^{(b_{0dm}-1)}. \tag{B.58}$$

The log-Beta prior is

$$\begin{aligned}\log P(\boldsymbol{p}) = \sum_{d=1}^{D}\sum_{m=1}^{M}\Big[&\log\Gamma(a_{0dm}+b_{0dm}) - \log\Gamma(a_{0dm}) - \log\Gamma(b_{0dm})\\ &+ (a_{0dm}-1)\log p_{dm} + (b_{0dm}-1)\log(1-p_{dm})\Big].\end{aligned} \tag{B.59}$$

The variational derivation of the latent allocation is

$$\begin{aligned}\log q(\boldsymbol{z}) &\propto \mathbb{E}_{/\boldsymbol{z}}\left[\log P(\boldsymbol{y}|\boldsymbol{\lambda},\boldsymbol{z}) + \log P(\boldsymbol{z}|\boldsymbol{w})\right]\\
&\propto \mathbb{E}_{/\boldsymbol{z}}\Bigg[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{y_{nd}\log p_{dm} + (s_{nd}-y_{nd})\log(1-p_{dm})\right\}\\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\log w_m + \sum_{j=1}^{m-1}\log(1-w_j)\right\}\Bigg]\\
&= \sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{y_{nd}\mathbb{E}_{p_{dm}}\left[\log p_{dm}\right] + (s_{nd}-y_{nd})\mathbb{E}_{p_{dm}}\left[\log(1-p_{dm})\right]\right\}\\
&\quad + \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\left\{\mathbb{E}_{w_m}\left[\log w_m\right] + \sum_{j=1}^{m-1}\mathbb{E}_{w_m}\left[\log(1-w_j)\right]\right\}\\
&= \sum_{n=1}^{N}\sum_{m=1}^{M} z_{nm}\Bigg\{\sum_{d=1}^{D} y_{nd}\mathbb{E}_{p_{dm}}\left[\log p_{dm}\right] + \sum_{d=1}^{D}(s_{nd}-y_{nd})\mathbb{E}_{p_{dm}}\left[\log(1-p_{dm})\right]\\
&\quad + \mathbb{E}_{w_m}\left[\log w_m\right] + \sum_{j=1}^{m-1}\mathbb{E}_{w_m}\left[\log(1-w_j)\right]\Bigg\}.\end{aligned} \tag{B.60}$$

Equation (B.60) is the logarithmic kernel of a Categorical density after we set the expression inside the brackets as $\log \rho_{nm}$

$$\begin{aligned}\log \rho_{nm} = &\sum_{d=1}^{D} y_{nd} \mathbb{E}_{p_{dm}} [\log p_{dm}] + \sum_{d=1}^{D} (s_{nd} - y_{nd}) \mathbb{E}_{p_{dm}} [\log(1 - p_{dm})] \\ &+ \mathbb{E}_{w_m} [\log w_m] + \sum_{j=1}^{m-1} \mathbb{E}_{w_m} [\log(1 - w_j)],\end{aligned} \tag{B.61}$$

concluding with

$$\log q(\boldsymbol{z}) \propto \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log \rho_{nm}. \tag{B.62}$$

In order for $q(\boldsymbol{z})$ to be equal and not just proportional to a Categorical density, the variational parameter of $\boldsymbol{z}$ (here will be denoted as $\boldsymbol{r}$) will be constrained to belong in the $[0, 1]$ interval and have $\sum_{m=1}^{M} r_{nm} = 1$. Hence, $r_{nm} = \rho_{nm} / \sum_{j=1}^{M} \rho_{nj}$. The variational log-Categorical density is

$$\log q(\boldsymbol{z}) = \sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \log r_{nm}. \tag{B.63}$$

The variational Categorical density is

$$q(\boldsymbol{z}) = \prod_{n=1}^{N} \prod_{m=1}^{M} {r_{nm}}^{z_{nm}}. \tag{B.64}$$

The variational derivation of the mixing weights is

$$\begin{aligned}\log q(\boldsymbol{w}) &\propto \mathbb{E}_{/\boldsymbol{w}} [\log P(\boldsymbol{z}|\boldsymbol{w}) + \log P(\boldsymbol{w})] \\ &\propto \mathbb{E}_{/\boldsymbol{w}} \left[\sum_{n=1}^{N} \sum_{m=1}^{M} z_{nm} \left\{ \log w_m + \sum_{j=1}^{m-1} \log(1 - w_j) \right\} + \sum_{m=1}^{M} (\phi_{0m} - 1) \log(1 - w_m) \right] \\ &= \sum_{m=1}^{M} \left\{ \sum_{n=1}^{N} \mathbb{E}_{z_{nm}} [z_{nm}] \log w_m + \left(\sum_{n=1}^{N} \sum_{j=m+1}^{M} \mathbb{E}_{\boldsymbol{z}_{nj}} [z_{nj}] + \phi_{0m} - 1 \right) \log(1 - w_m) \right\}.\end{aligned} \tag{B.65}$$

Equation (B.65) reminds the kernel of a Log-Dirichlet density if we set

$$\begin{aligned}\delta_m &= 1 + \sum_{n=1}^{N} \mathbb{E}_{z_{nm}} [z_{nm}], \\ \phi_m &= \phi_{0m} + \sum_{n=1}^{N} \sum_{j=m+1}^{M} \mathbb{E}_{\boldsymbol{z}_{nj}} [z_{nj}].\end{aligned} \tag{B.66}$$

The variational log-Dirichlet density is

$$\begin{aligned}\log q(\boldsymbol{w}) = \sum_{m=1}^{M} \Big\{ &\log \Gamma(\delta_m + \phi_m) - \log \Gamma(\delta_m) - \log \Gamma(\phi_m) + (\delta_m - 1) \log w_m \\ &+ (\phi_m - 1) \log(1 - w_m) \Big\}.\end{aligned} \tag{B.67}$$

The variational Dirichlet density is

$$q(\boldsymbol{w}) = \prod_{m=1}^{M} \frac{\Gamma(\delta_m + \phi_m)}{\Gamma(\delta_m) \Gamma(\phi_m)} {w_m}^{(\delta_m - 1)} (1 - w_m)^{(\phi_1 - 1)}. \tag{B.68}$$

The variational derivation of the model parameter is

$$\begin{aligned}
\log q(\boldsymbol{p}) &\propto \mathbb{E}_{/\boldsymbol{p}}\left[\log P(\boldsymbol{y}|\boldsymbol{p},\boldsymbol{z}) + \log P(\boldsymbol{p})\right] \\
&\propto \mathbb{E}_{/\boldsymbol{p}}\Bigg[\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m=1}^{M} z_{nm}\left\{y_{nd}\log p_{dm} + (s_{nd}-y_{nd})\log(1-p_{dm})\right\} \\
&\quad + \sum_{d=1}^{D}\sum_{m=1}^{M}\Big\{(a_{0dm}-1)\log p_{dm} + (b_{0dm}-1)\log(1-p_{dm})\Big\} \\
&= \sum_{m=1}^{M}\sum_{d=1}^{D}\Bigg\{\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]y_{nd}\log p_{dm} + \sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}](s_{nd}-y_{nd})\log(1-p_{dm}) \\
&\quad + (a_{0dm}-1)\log p_{dm} + (b_{0dm}-1)\log(1-p_{dm})\Bigg\} \\
&= \sum_{m=1}^{M}\sum_{d=1}^{D}\Bigg\{\Bigg(\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]y_{nd} + a_{0dm} - 1\Bigg)\log p_{dm} \\
&\quad + \Bigg(\sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}](s_{nd}-y_{nd}) + b_{0dm} - 1\Bigg)\log(1-p_{dm})\Bigg\}.
\end{aligned} \tag{B.69}$$

Equation (B.69) reminds the kernel of a Log-Beta density if we set the two parentheses as

$$\begin{aligned}
a_{dm} &= a_{0dm} + \sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}]y_{nd}, \\
b_{dm} &= b_{0dm} + \sum_{n=1}^{N}\mathbb{E}_{z_{nm}}[z_{nm}](s_{nd}-y_{nd}).
\end{aligned} \tag{B.70}$$

The variational log-Beta density is

$$\begin{aligned}
\log P(\boldsymbol{p}) = \sum_{d=1}^{D}\sum_{m=1}^{M}\Big[&\log\Gamma(a_{dm}+b_{dm}) - \log\Gamma(a_{dm}) - \log\Gamma(b_{dm}) \\
&+ (a_{dm}-1)\log p_{dm} + (b_{dm}-1)\log(1-p_{dm})\Big].
\end{aligned} \tag{B.71}$$

The variational Beta density is

$$P(\boldsymbol{p}) = \prod_{d=1}^{D}\prod_{m=1}^{M}\frac{\Gamma(a_{dm}+b_{dm})}{\Gamma(a_{dm})\Gamma(b_{dm})}{p_{dm}}^{(a_{dm}-1)}(1-p_m)^{(b_{dm}-1)}. \tag{B.72}$$

The Evidence Lower Bound (ELBO) is

$$\begin{aligned}
L(\boldsymbol{y};q) =& \mathbb{E}_{\boldsymbol{z},\boldsymbol{w},\boldsymbol{p}}[\log P(\boldsymbol{y},\boldsymbol{z},\boldsymbol{p},\boldsymbol{w})] - \mathbb{E}_{\boldsymbol{z}}[\log q(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{w})] - \mathbb{E}_{\boldsymbol{p}}[\log q(\boldsymbol{p})] \\
=& \mathbb{E}_{\boldsymbol{z},\boldsymbol{\pi},\boldsymbol{p}}[\log P(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{p}) + \log P(\boldsymbol{z}|\boldsymbol{w}) + \log P(\boldsymbol{w}) + \log P(\boldsymbol{p})] \\
&- \mathbb{E}_{\boldsymbol{z}}[\log q(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{\pi}}[\log q(\boldsymbol{w})] - \mathbb{E}_{\boldsymbol{\lambda}}[\log q(\boldsymbol{p})].
\end{aligned} \tag{B.73}$$

The explicit ELBO form is

$$\begin{aligned}
L(\boldsymbol{y};q) = &\sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}_{z_{nm}}[z_{nm}]\log\rho_{nm} - \sum_{n=1}^{N}\sum_{m=1}^{M}\mathbb{E}_{z_{nm}}[z_{nm}]\log r_{nm} \\
&+ \sum_{m=1}^{M}\left[\log\phi_{0m} + (\phi_{0m}-1)\log(1-w_m)\right]
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{m=1}^{M} \Big\{ \log\Gamma(\delta_m + \phi_m) - \log\Gamma(\delta_m) - \log\Gamma(\phi_m) + (\delta_m - 1)\log w_m \\
&+ (\phi_m - 1)\log(1 - w_m) \Big\} \\
&+ \sum_{d=1}^{D} \sum_{m=1}^{M} \Big\{ \log\Gamma(a_{0dm} + b_{0dm}) - \log\Gamma(a_{dm} + b_{dm}) - \log\Gamma(a_{0dm}) - \log\Gamma(b_{0dm}) \\
&+ \log\Gamma(a_{0dm}) + \log\Gamma(b_{0dm}) + (a_{0dm} - a_{dm})\mathbb{E}_{\boldsymbol{p}}[\log p_{dm}] \\
&+ (b_{0dm} - b_{dm})\mathbb{E}_{\boldsymbol{p}}[\log(1 - p_{dm})] \Big\}.
\end{aligned}
\tag{B.74}
$$

The variational expectations and posterior estimates of the mixing weights are

$$
\begin{aligned}
\mathbb{E}_{z_{nm}}[z_{nm}] &= r_{nm}, \\
\mathbb{E}_{w_m}[\log w_m] &= \Psi(\phi_m) - \Psi(\delta_m + \phi_m), \\
\mathbb{E}_{w_m}[\log(1 - w_m)] &= \Psi(\delta_m) - \Psi(\delta_m + \phi_m), \\
\mathbb{E}_{p_{dm}}[\log p_{dm}] &= \Psi(a_{dm}) - \Psi(a_{dm} + b_{dm}), \\
\mathbb{E}_{p_{dm}}[(1 - p_{dm})] &= \Psi(b_{dm}) - \Psi(a_{dm} + b_{dm}), \\
\pi_m &= \mathbb{E}[w_m] \prod_{j=1}^{m-1} \big(1 - \mathbb{E}[w_j]\big).
\end{aligned}
\tag{B.75}
$$

B.3 Code Snippet

Here, we provide a code snippet with our main R function for the variational Dirichlet Process Gaussian mixture model with independent features. This is a frame of reference for the reader to help her understand the general code structure of the variational algorithm in mixture models.

B.3.1 Variational Dirichlet Process Gaussian Mixture (independent features)

The main variational function is given with its inputs and outputs defined.

```
#### safe computation of logsumexp (included in the main function)
# inputs:
  ## y: scalar
# output:
  ## lse: safe log sum exp value
logsumexp <- function(y)
  {
  # Computes log(sum(exp(x))
  a <- max(y)
  lse <- log(sum(exp(y-a))) + a
  j <- which(!is.finite(lse))
  if (length(j) > 0) {lse[j] <- a}

  return(lse)
}

#### main function for variational DP GaussianMix (independent across dimensions)
# inputs:
  ## X: NxD data matrix
  ## M: number of initial components
  ## alpha/beta: DxM Gamma initial variational matrices for the variance of
  the Gaussians
  ## m/s2: DxM Gamma initial variational matrices for the mean of the
  Gaussians
  ## p/q: 1xM initial variational vectors for the stick-breaking point Beta
  parameter
  ## phi_0: an 1xM vector for the stick-breaking point Beta parameter
  ## alpha_0/beta_0: DxM Gamma hyperparameter matrices for the variance of
  the Gaussians
  ## m_0/s2_0: DxM Normal hyperparameter matrices for the mean of the
  Gaussians
  ## T: the temperature vector for the annealing part (pre-define)
  ## max_iterations: maximum number of VB iterations
  ## epsilon: threshold to achieve convergence
# output:
  ## alpha/beta: DxM Gamma variational matrices for the variance of the
  Gaussians
  ## m/s2: DxM Gamma variational matrices for the mean of the Gaussians
  ## p/q: 1xM variational vectors for the stick-breaking point Beta parameter
  ## r: NxD variational matrix for the latent allocation z
  ## L: ELBO values
  ## w: weight values in each iteration (for evolution purposes)
  ## printL: print the ELBO values and difference to the previous one

avb_dpgm <- function(X, M, alpha, beta, m, s2, p, q, alpha_0, beta_0,
                     m_0, s2_0, phi_0, T, max_iterations=iter, epsilon=1e-4,
printL=FALSE)
{
  # define the dimensions according to the dataset
  X <- as.matrix(X)
```

```
D <- ncol(X)
N <- nrow(X)

# initial objects to receive the variational results
# initial ELBO
L <- rep(-Inf, max_iterations)

# initial weights in each iteration
w <- matrix(1/M, ncol= M, nrow = max_iterations)

# initial weights in final iteration
pi <- rep(0, M)

  ### AVB scheme
  for (i in 2:max_iterations)
    {

    # calculation of expectations contained into the variational parameters
    E.mu <- m
    V.mu <- s2
    E.logsigma2 <- log(beta) - digamma(alpha)
    E.inv_sigma2 <- alpha/beta
    E.loglambda <- digamma(p) - digamma(p + q)
    E.log1_lambda <-  digamma(q) - digamma(p + q)

    # calculation of terms found in the variational equations
    N.E.logsigma2 <- matrix(colSums(E.logsigma2, na.rm=TRUE),  nrow=N,
    ncol=M, byrow=TRUE)
    N.E.inv_sigma2 <- matrix(colSums(E.inv_sigma2, na.rm=TRUE),  nrow=N,
    ncol=M, byrow=TRUE)
    E.log1_lambdaj <- c(0, cumsum(E.log1_lambda)[1:(M-1)])
    N.E.log1_lambdaj <- matrix(E.log1_lambdaj, nrow=N, ncol=M, byrow=TRUE)
    N.E.loglambda <- matrix(E.loglambda, nrow=N, ncol=M, byrow=TRUE)
    log.2pi <- matrix(log(2*355/113), nrow=N, ncol=M)

    inv_s2.x2 <- (X^(2)) %*% E.inv_sigma2
    inv_s2.x.m <- X %*% (E.inv_sigma2 * E.mu)
    inv_s2.s2 <- matrix(colSums(E.inv_sigma2 * V.mu), ncol=M, nrow=N, byrow
    = TRUE)
    inv_s2.m2 <- matrix( colSums(E.inv_sigma2 * (E.mu)^(2)), ncol=M,
    nrow=N, byrow = TRUE)

    D.X_m <-  inv_s2.x2 - 2 * inv_s2.x.m + inv_s2.s2 + inv_s2.m2

    log_rho <- N.E.loglambda + N.E.log1_lambdaj - (D/2) * log.2pi - (1/2) *
    N.E.logsigma2 - (1/2) * D.X_m

    # the usefulness of logsumexp function
    S <- apply(log_rho, 1, logsumexp)
```

```
log_r <- log_rho - S

# the variational parameter of z
r <- apply(log_r, 2, exp)
# trick to avoid zero values
r <- (r + 10^-9)^ Tinv[i]

# term into p,q variational parameters
Ns <- colSums(r, na.rm=TRUE)

# the stick-breaking point variational parameters
p <- 1 + Tinv[i] * Ns
q <- phi_0 + Tinv[i] * (rev(cumsum(rev(Ns))) - Ns)

# term into alpha,mu variational parameters
r.colSums <- matrix(Ns, nrow=D, ncol=M, byrow=TRUE)

# first Gamma variational parameter for the variances of the Gaussian
mixture
alpha <- alpha_0 + (1/2) * r.colSums *Tinv[i]

# calculation of terms conatined into the second Gamma variational
parameter
x2.r <- t(X^(2)) %*% r
x.r.m <- (t(X) %*% r) * E.mu
Ns.m2 <- r.colSums * (E.mu)^2
Ns.s2 <- r.colSums * V.mu
N.x_m <- x2.r - 2 * x.r.m + Ns.m2 + Ns.s2
# second Gamma variational parameter for the variances
beta <- beta_0 + (1/2) * N.x_m *Tinv[i]

# variational Gaussian variance for the means of the Gaussian mixture
s2 <- ((1/s2_0) + Tinv[i]* E.inv_sigma2 * r.colSums)^(-1)

# variational Gaussian mean for the means of the Gaussian mixture
m <- s2 * ((m_0/s2_0) +  Tinv[i]* E.inv_sigma2 * (t(X) %*% r))

# the stick-breaking point parameter
lambda <- head(p, M-1) / (head(p, M-1) + head(q, M-1))
lambda <- c(lambda, 1)

# the variational weights after the stick-breaking point computation
for (k in 1:M)
{
  pi[k] <- lambda[k] * prod(head(1-lambda, k-1))
}

# the variational weights in each iteration
w[i, ] <- round(pi,3)
```

```
    # update the expectations contained into the ELBO
    E.mu <- m
    V.mu <- s2
    E.logsigma2 <- log(beta) - digamma(alpha)
    E.inv_sigma2 <- alpha/beta
    E.loglambda <- digamma(p) - digamma(p + q)
    E.log1_lambda <-  digamma(q) - digamma(p + q)

    ## ELBO
    # each term into the ELBO has been calculated individually
    l1 <- sum (r * Tinv[i] *log_rho)
    l2 <- sum(log( phi_0) + (phi_0 - 1) * E.log1_lambda )
    l3 <- - (D * M /2) * log(2 * 355/113) - (1/2) * (sum( log(s2_0)) +
    sum(s2_0^(-1) * ((E.mu - m_0)^(2) + V.mu)))
    l4 <- sum(alpha_0 * log(beta_0)) - sum(lgamma(alpha_0)) - sum((alpha_0
    + 1) * E.logsigma2) - sum(beta_0 * E.inv_sigma2)
    l5 <- - sum(r * log(r))

    # useful names for terms in l6
    log_g.p.q <- lgamma( p + q)
    log_p <- lgamma(p)
    log_q <- lgamma(q)

    l6_1 <-  - sum(log_g.p.q) + sum(log_p) + sum(log_q)
    l6_2 <-  - sum((p -1) * E.loglambda ) - sum((q -1) * E.log1_lambda)
    l6 <- l6_1 + l6_2

    l7 <- (D * M /2) * log(2 * 355/113) + (1/2) * sum(log(V.mu)) + M*D/2
    l8 <- - sum(alpha * log(beta)) + sum(lgamma(alpha)) + sum((alpha + 1) *
    E.logsigma2) + sum(beta * E.inv_sigma2)

    # Total ELBO calculation
    L[i] <- l1 + l2 + l3 + l4 + l5 + l6 +l7 +l8

    # print ELBO value and difference with the previous one
    if (printL) { cat("Iter:\t", i, "\tELBO:\t", L[i], "\tELBO_diff:\t",
    L[i] - L[i-1], "\n")}

    # test if ELBO decreases
    if (L[i] < L[i - 1]) { message("Warning: ELBO decreases\n"); }

    # test convergence with epsilon threshold
    if (abs(L[i] - L[i - 1]) < epsilon) { break }

    # test VB needs more iteration to converge
    if (i == max_iterations) {warning("VB did not converge\n")}
  }
```

```
  object <- structure(list(alpha=alpha, beta=beta, m=m, s2=s2, p=p, q=q, r=r,
  L=L[2:i], w=w))

  return(object)
}
```