Bayesian Methods and Data Science with Health Informatics Data

by

Iliana Stanimirova Peneva

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Mathematics for Real-World Systems CDT

January 2019

Contents

List of Tables

List of Figures

Acknowledgments

There are many people who made the completion of this thesis possible. First of all, I would like to thank Rich Savage for his wonderful supervision, patience and support during my PhD. I have been extremely fortunate to be part of Team Savage.

I would like to thank my Birmingham supervisors - Keith Roberts, Felicity Evison and Paul Moss, who have been a great inspiration and have helped me understand better the biology of pancreatic cancer and the challenges of working with real-life clinical data.

None of this would have been possible without the great team from the University Hospitals Birmingham, the amazing pancreatic cancer patients from the HES study and all patients and researchers involved in The Cancer Genome Atlas project.

I am also grateful to David Rossell and Jairo Fuquene for their invaluable help with non-local priors.

I would like to thank my examiners, Prof. David Wild and Dr. Paul Kirk, for their challenging questions and helpful feedback.

Special thanks goes to Team Savage - Kat Lloyd, Jim Skinner, Matt Neal, Ale Avalos, Ayman Boustati and Nadia Jankovicova, who have been the best academic family I could ask for.

I have been very lucky to be part of the Mathematics for Real-World Systems CDT and the Centre for Complexity Science for the last four years, and I would like to thank the staff, in particular the admin team Heather Robson and Debbie Walker, and Magnus Richardson, for their help and support, and for making my PhD so enjoyable. Thank you to all the lovely MSc and PhD students, and to the 2014 MathSys cohort for sharing the experience. I have had many great office mates over the years: Joe, Sami, Roger, Chris, Cameron, Ayman, Alex, Ellen, Fede, Bernd and Ed, in the best office D2.05 - thank you for the conversations, the long lunches and much needed coffee breaks, and Jessie, Katherine, Gian Lorenzo and Michael in the infamous hot-desking office D1.13 - thank you for making the writing up more bearable and fun. A special thank you goes to Aditi for all the runs and coffee breaks that kept me sane.

There are many people not based at the University of Warwick, who helped a lot with their enthusiasm and support (and many catch-ups). Thank you to the Cov family - Amber, Alex and Catrin, who have been the best family away from home; to Clea and the lovely Zeltmann family for being there for me since I moved to UK 8 years ago; to Tash for her constant support; to Becky for her encouragement and friendship; to Evangelia for her enthusiasm; to Pooja for starting the tea club and for listening to my rants; to the Kenilworthians - Katherine, John, Cameron, Vicky, Rob, Matt and Jessie, for making me feel very welcome; to Jim, Alonso, Steph, Jack and Becky, who have been the loveliest housemates and friends.

I would like to thank my wonderful partner Jim for his love, support and patience, especially in the last few months. Finally, a huge thank you to my family for their love and support, and especially to my mum, Ganka, who has been my main supporter even from 2000 miles away and has sacrificed a lot so that I could follow my dreams.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. This thesis has not been submitted for a degree at another university.

The origin of the data for the experiments in this thesis has been explicitly mentioned in the thesis.

Parts of Chapter 6 have been published in Liu et al. [2018].

The glioblastoma case study from Chapter 5 is accepted in the proceedings of Bayesian Young Statisticians Meeting 2018.

The integrative framework and the results from Chapter 5 will be submitted shortly for publication.

Abstract

Cancer is a complex disease, driven by a range of genetic and environmental factors. Every year millions of people are diagnosed with a type of cancer and the survival prognosis for many of them is poor due to the lack of understanding of the causes of some cancers. Modern large-scale studies offer a great opportunity to study the mechanisms underlying different types of cancer but also brings the challenges of selecting informative features, estimating the number of cancer subtypes, and providing interpretative results.

In this thesis, we address these challenges by developing efficient clustering algorithms based on Dirichlet process mixture models which can be applied to different data types (continuous, discrete, mixed) and to multiple data sources (in our case, molecular and clinical data) simultaneously. We show how our methodology addresses the drawbacks of widely used clustering methods such as k-means and iClusterPlus. We also introduce a more efficient version of the clustering methods by using simulated annealing in the inference stage.

We apply the data integration methods to data from The Cancer Genome Atlas (TCGA), which include clinical and molecular data about glioblastoma, breast cancer, colorectal cancer, and pancreatic cancer. We find subtypes which are prognostic of the overall survival in two aggressive types of cancer: pancreatic cancer and glioblastoma, which were not identified by the comparison models.

We analyse a Hospital Episode Statistics (HES) dataset comprising clinical information about all pancreatic cancer patients in the United Kingdom operated during the period 2001 - 2016. We investigate the effect of centralisation on the short- and long-term survival of the patients, and the factors affecting the patient survival. Our analyses show that higher volume surgery centres are associated with lower 90-day mortality rates and that age, index of multiple deprivation and diagnosis type are significant risk factors for the short-term survival.

Our findings suggest the analysis of large complex molecular datasets coupled with methodology advances can allow us to gain valuable insights in the cancer genome and the associated molecular mechanisms.

Abbreviations

ACS-NSQUIP American College of Surgeons National Surgical Quality Improvement Program

AHRQ Agency for Healthcare Research and Quality

AIC Akaike information criterion

ARI Adjusted Rand index

ASA The American Society of Anaestethesiologists

AUC area under the curve

BCC Bayesian consensus clustering

BIC Bayesian information criterion

BMI body mass index

BPCA Bayesian principal component analysis

BRCA breast cancer

CI confidence interval

COPD chronic obstructive pulmonary disease

CRC colorectal cancer

DIC deviance information criterion

DP Dirichlet process

EM Expectation Maximisation

eMOM exponential moment prior

GBM glioblastoma multiforme

GLM generalised linear model

HES Hospital Episode Statistics

HR hazard ratio

ICGC International Cancer Genome Consortium

IMD index of multiple deprivation

iMOM inverse moment prior

JIVE joint and individual variation explained

KEGG Kyoto Encyclopedia of Genes and Genomes

LP local prior

MCMC Markov chain Monte Carlo

MDI multiple dataset integration

miRNA micro ribonucleic acid

MOFA multi-omics factor analysis

MOM moment prior

mRNA messenger ribonucleic acid

NA not applicable

NHS National Health Service

NIS Nationwide inpatient sample

NLP non-local prior

ONS Office of National Statistics

PAAD pancreatic cancer

PCA principal component analysis

PD pancreatoduodenectomy

PH proportional hazards

PPCA probabilistic principal component analysis

PSDF Patient-specific data fusion

RSF random survival forest

SNP single nucleotide polymorphism

TCGA The Cancer Genomic Atlas

WAIC Watanabe-Akaike/widely applicable information criterion

WHO World Health Organisation

Chapter 1

Introduction

In this chapter we describe the application which motivates the development of the Bayesian clustering methods presented in this thesis. We then review some of the most common inference methods employed in the following chapters. We present some concepts related to clustering and evaluation measures, which are fundamental to the thesis, and discuss the advantages of adopting a Bayesian approach when working with complex, high-dimensional data.

1.1 Motivation

Cancer is a major global health problem with 18.1 million new cases being diagnosed every year and estimated 9.6 million cancer deaths yearly [World Health Organisation, 2018]. There are more than 100 distinct types of cancer, and subtypes of tumours can be found within specific organs [Hanahan and Weinberg, 2000]. The survival prognosis for many of them is poor and there is lack of effective treatments especially for late-stage cancers due to the lack of understanding of the causes of some cancers. Despite the differences between cancer types, Hanahan and Weinberg [2000] suggested that the formation of tumours is a result of the same 6 alterations in cell physiology (Figure 1.1):

- self-sufficiency in growth signals
- insensitivity to growth-inhibitory signals
- evasion of apoptosis [1]

[1] programmed cell death

- limitless replicative potential
- sustained angiogenesis [2]
- tissue invasion and metastasis.

Later Hanahan and Weinberg [2011] add *deregulating cellular energetics* and *avoiding immune destruction* as emerging hallmarks, and define *genome instability and mutation*, and *tumour-promoting inflammation* as enabling characteristics which are associated with the acquisition of hallmark capabilities.

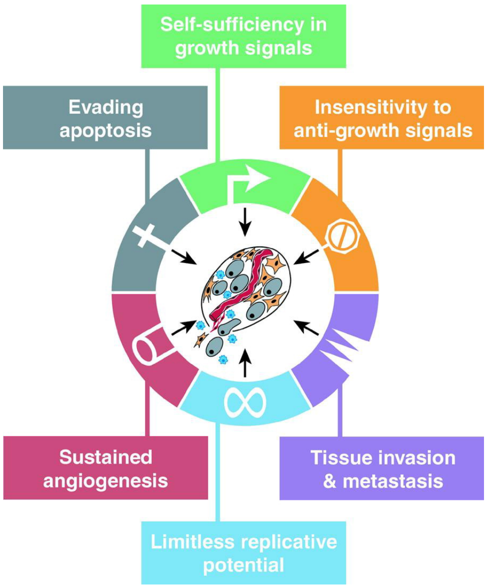

Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to anti-growth signals
Sustained angiogenesis
Tissue invasion & metastasis
Limitless replicative potential

Figure 1.1: Acquired capabilities of cancer. Not all processes are involved in the development of all tumours. Credit: Hanahan and Weinberg [2000]

Identifying the changes in the cancer genome and understanding how they interact and affect the patient should lead to earlier detection, better treatment and prevention [Verma, 2012].

Modern large-scale studies offer great opportunities to study the mechanisms underlying different types of cancer and to stratify patients into distinct subgroups that are characteristic of response to treatments and/or overall survival. For example, The Cancer Genome Atlas project [Weinstein et al., 2013] (TCGA), run jointly between the US National Cancer Institute and the National Human Genome Research Institute, has generated comprehensive, multi-dimensional maps of genomic changes of 33 different tumour types, using data from over 11000 patients, such as

[2] formation of new blood vessels

gene expression levels, methylation, single nucleotide polymorphisms (SNP) [3] and copy number variation. Another large-scale genomic study is METABRIC [Curtis et al., 2012], which was a collaboration between Canada and the UK, and aimed to classify breast tumours into further subgroups, based on molecular signatures in order to determine the optimal course of treatment for each patient. The International Cancer Genome Consortium [International Cancer Genome Consortium and others, 2010] (ICGC) is a worldwide project which coordinates a large number of research projects studying different types of cancer in a range of populations. This project similarly aims to unravel the genomic changes present in the cancer genomes.

Each of these large-scale studies involves working with a large amount of data, often from different sources. This brings the challenges of selecting informative features, estimating the number of cancer subtypes, and providing interpretative results. Identifying informative features is especially important when using genomic data as we expect only a fraction of them to contain useful information and extracting these features will improve the quality of the output. Often the expected number of cancer subtypes in a given analysis has to be pre-specified [Shen et al., 2009] which can affect the final results; however, jointly learning the cluster parameters and the number of subtypes may give better quality results that can be used to make better informed clinical decisions. Many recent studies [Shen et al., 2009; Yuan et al., 2011; Kirk et al., 2012; Mo et al., 2013, 2017; Gabasova et al., 2017] propose integrative clustering approaches to address these problems. They are based on the idea that none of the individual datasets can fully capture the complexity of cancer, but collectively, they can offer a better understanding of the true oncogenic mechanisms.

1.2 Statistical background

1.2.1 Bayesian Methods

In the Bayesian formalism of the world, probabilities capture a belief state about events. If we consider an uncertain event $\mathcal{X}$, we can encode our prior beliefs about the model $\mathcal{M}$ that generated it by a **prior** $p(\theta|\mathcal{M})$ on the model parameters θ. Once we observe the event, we can update our beliefs, in a principled manner, by using the Bayesian framework, and Bayes' theorem in particular.

[3] the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called nucleotide.

The proof of Bayes' theorem makes use of the *product rule* $p(X,Y) = p(Y|X)p(X)$, which together with the *sum rule* $p(X) = \sum_Y p(X,Y)$ form the fundamental axioms of probability, and of the symmetry property $p(X,Y) = p(Y,X)$. This means that we can express the joint probability distribution $p(\theta, \mathcal{X}|\mathcal{M})$ in two ways:

$$p(\theta, \mathcal{X}|\mathcal{M}) = p(\mathcal{X}, \theta|\mathcal{M}) \tag{1.1}$$

$$p(\mathcal{X}|\theta, \mathcal{M})p(\theta|\mathcal{M}) = p(\theta|\mathcal{X}, \mathcal{M})p(\mathcal{X}|\mathcal{M}), \tag{1.2}$$

which can be rearranged as

$$p(\theta|\mathcal{X}, \mathcal{M}) = \frac{p(\mathcal{X}|\theta, \mathcal{M})p(\theta|\mathcal{M})}{p(\mathcal{X}|\mathcal{M})}. \tag{1.3}$$

Equation (1.3) is what is more widely known as **Bayes' theorem** or **Bayes' rule**, where $p(\mathcal{X}|\theta, \mathcal{M})$ is the probability of $\mathcal{X}$ conditioned on θ and the model, also known as **likelihood**; $p(\theta|\mathcal{X}, \mathcal{M})$ is the **posterior probability** of θ after observing $\mathcal{X}$. The normalising constant $p(\mathcal{X}|\mathcal{M})$ in (1.3), also called **marginal likelihood** or **evidence**, is given by

$$p(\mathcal{X}|\mathcal{M}) = \int p(\mathcal{X}, \theta|\mathcal{M})\mathrm{d}\theta \tag{1.4}$$

$$= \int p(\mathcal{X}|\theta, \mathcal{M})p(\theta|\mathcal{M})\mathrm{d}\theta. \tag{1.5}$$

These Bayesian concepts are extremely useful not only for incorporating our prior beliefs about uncertain events, but also for performing model selection.

1.2.2 Model selection

Once we have a collection of models that could explain our data D, how do we choose the 'best' model? Multiple information criteria have been proposed to aid model selection. One of these is **Akaike Information Criterion** (AIC) [Akaike, 1974], which is based on Kullback-Leibler divergence. It does not assume that the true model is amongst the models under consideration and that the models must be nested. In AIC the number of model parameters θ is subtracted from the log density given the maximum likelihood parameter estimates $\hat{\theta}_{MLE}$ to account for how much the fitting of K parameters will increase the accuracy:

$$\mathrm{AIC} = -2\log p(D|\hat{\theta}_{MLE}) + 2K. \tag{1.6}$$

This criterion is appropriate for linear models with flat priors but if we work with hierarchical models with informative priors, it becomes less accurate [Gelman et al., 2014b].

Deviance Information Criterion (DIC) [Spiegelhalter et al., 2002] is a Bayesian version of AIC and replaces the maximum likelihood estimate $\hat{\theta}$ with the posterior mean $\hat{\theta}_{Bayes} = \mathbb{E}(\theta|y)$, and K with a data-based bias correction:

$$\text{DIC} = -2\log p(D|\hat{\theta}_{Bayes}) + 2p_{DIC}, \tag{1.7}$$

where p_{DIC} is the effective number of parameters:

$$p_{DIC} = 2\Bigg(\log p(D|\hat{\theta}_{Bayes}) - \mathbb{E}_{post}(\log p(D|\theta))\Bigg), \tag{1.8}$$

and $\mathbb{E}_{post}$ is an average of θ over its posterior distribution, which is calculated using simulations $\theta^s, s = 1, \ldots, S$ and $\mathbb{E}_{post}(\log p(y|\theta)) = \frac{1}{S}\sum_{s=1}^{S}\log p(y|\theta^s)$. DIC is easier to compute in comparison with AIC as it does not require maximising of the likelihood because it uses MCMC samples from the posterior. However, it requires the mean to be a good estimator of the posterior, which is not true for skewed distributions, for example.

Watanabe-Akaike/widely applicable criterion (WAIC) [Watanabe, 2013] is a cheaper approximation of cross-validation and is defined as follows:

$$\text{WAIC} = -2lppd + 2p_{WAIC2}, \tag{1.9}$$

where *lppd* is the log pointwise predictive density and is equal to

$$\text{computed lppd} = \sum_{i=1}^{n}\log\left(\frac{1}{S}\sum_{s=1}^{S}p(D_i|\theta^s)\right), \tag{1.10}$$

where θ_s are the draws from the posterior simulations, D_i is the ith data point, and p_{WAIC2} is a correction for the effective number of parameters to adjust for overfitting:

$$p_{WAIC2} = \sum_{i=1}^{n} var_{post}(\log p(y_i|\theta)). \tag{1.11}$$

Here var_{post} is the posterior variance of the log predictive density for each data point $V_{s=1}^{S}\log p(y_i|\theta_s)$ where $V_{s=1}^{S}$ represents the sample variance.

Bayesian Information Criterion (BIC) [Schwarz et al., 1978] adjusts for the number of fitted parameters with a penalty that increases with the sample size N and favours simpler models:

$$\text{BIC} = \log p(D|\theta_{MAP}) - \frac{1}{2}K \log N, \tag{1.12}$$

where θ_{MAP} are the parameters which maximise the posterior. BIC assumes that the sample size N is much larger than the number of parameters K [Aho et al., 2014; Kuha, 2004].

Bayesian Model Selection

Another way of comparing models is through performing Bayesian analysis where each model is given a prior probability, and then by multiplying it by the model marginal likelihood, we obtain a quantity proportional to the model posterior probability.

In Bayesian model selection, we select the model that corresponds to the most probable model $\mathcal{M}$ given $\mathcal{X}$, i.e. the model with highest posterior probability $p(\mathcal{M}|\mathcal{X})$. With the increase in the number of observations of $\mathcal{X}$, the mass of $p(\mathcal{M}|\mathcal{X})$ typically concentrates around one model. Hence, picking the model with the highest posterior probability is a reasonable choice.

In addition, if we place uniform prior over all models $\mathcal{M}$, we can express the posterior $p(\mathcal{M}|\mathcal{X})$ as follows:

$$\begin{aligned} p(\mathcal{M}|\mathcal{X}) =& \frac{p(\mathcal{X}|\mathcal{M})p(\mathcal{M})}{\int p(\mathcal{X}|\mathcal{M})p(\mathcal{M})\,\mathrm{d}\mathcal{M}} && (1.13)\\ =& p(\mathcal{X}|\mathcal{M})\frac{p(\mathcal{M})}{p(\mathcal{X})} && (1.14)\\ \propto& p(\mathcal{X}|\mathcal{M}). && (1.15) \end{aligned}$$

It turns out that instead of computing the posterior $p(\mathcal{M}|\mathcal{X})$, we can use the evidence $p(\mathcal{X}|\mathcal{M})$ to perform model selection. If we work with conjugate priors, we can compute (1.4) in closed form [Bishop, 2006]. Although this is rarely the case when we work with real, high-dimensional datasets, we can still use the evidence if we make use of certain approximations, in particular the **Laplace approximation**. We briefly summarise below the use of the Laplace approximation, which aims to find a Gaussian approximation $q(\mathbf{z})$ to a probability density $p(\mathbf{z}) = \frac{\mathbf{f(z)}}{\int \mathbf{f(z)}\mathrm{d}\mathbf{z}}$ defined over a set of M-dimensional continuous variables.

As the Gaussian distribution has the property that its logarithm is a quadratic function of the variables, we consider a Taylor expansion of $\log f(\mathbf{z})$ centred on the stationary point $\mathbf{z}_0$ where the gradient $\nabla f(\mathbf{z})$ vanishes:

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\intercal \mathbf{A}(\mathbf{z}-\mathbf{z}_0), \tag{1.16}$$

where $\mathbf{A}$ is the Hessian matrix defined by $\mathbf{A} = -\nabla\nabla \log f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0}$. Taking the exponential of both sides of (1.16), we get

$$f(\mathbf{z}) \approx f(\mathbf{z}_0)\exp\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\intercal \mathbf{A}(\mathbf{z}-\mathbf{z}_0)\} \tag{1.17}$$

which means that we can write $q(\mathbf{z})$ as

$$\begin{aligned} q(\mathbf{z}) =& \frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}}\exp\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\intercal \mathbf{A}(\mathbf{z}-\mathbf{z}_0)\} && (1.18)\\ =& \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1}) && (1.19) \end{aligned}$$

since it is proportional to $p(\mathbf{z})$.

Using this result, we can now approximate the model evidence in the following way:

$$\begin{aligned} p(\mathcal{X}|\mathcal{M}) =& \int p(\mathcal{X}|\theta,\mathcal{M})p(\theta|\mathcal{M})\mathrm{d}\theta && (1.20)\\ \approx& p(\mathcal{X}|\hat{\theta},\mathcal{M})p(\hat{\theta}|\mathcal{M})\int \exp\{-\frac{1}{2}(\theta-\hat{\theta})^\intercal \mathbf{A}(\theta-\hat{\theta})\}\mathrm{d}\theta && (1.21)\\ \approx& p(\mathcal{X}|\hat{\theta},\mathcal{M})p(\hat{\theta}|\mathcal{M})(2\pi)^{\frac{M}{2}}|A|^{-\frac{1}{2}}, && (1.22) \end{aligned}$$

where $\hat{\theta}$ is the value of θ at the mode of the posterior distribution.

When we perform statistical analysis, we usually consider multiple models $\mathcal{M}_i$ each with parameters θ_i. We can easily extend the ideas presented above to compare multiple models $\mathcal{M}_i$. If we put prior $p(\theta_i|\mathcal{M}_i)$ over the model parameters, then we can approximate the model evidence $p(D|\mathcal{M}_i)$ by the **Bayesian Information Criterion** (BIC) [Schwarz et al., 1978]

$$\begin{aligned} \log p(D|\mathcal{M}_i) \approx& \log p(D|\theta_{MAP},\mathcal{M}) + \log p(\theta_{MAP}|\mathcal{M}) + \frac{K}{2}\log(2\pi) - \frac{1}{2}|\mathbf{A}| && (1.23)\\ \approx& \log p(D|\theta_{MAP}) - \frac{1}{2}K\log N, && (1.24) \end{aligned}$$

where θ_{MAP} are the parameters which maximise the posterior, K is the number of free parameters, N is the number of data points in D and $\mathbf{A}$ is the Hessian matrix of second derivatives of the negative log posterior ($\mathbf{A} = -\nabla\nabla \log p(\theta_{MAP}|D, \mathcal{X})$). Since BIC is an approximation of the model evidence and we use it when we perform model selection in the analyses we conduct in this thesis.

1.2.3 Probability distributions

Bernoulli distribution

This is the distribution for a single binary variable $x \in \{0, 1\}$. It is a special case of the Binomial distribution for a single observation, and has the following form:

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}, \tag{1.25}$$

where μ is the probability of observing $x = 1$. The expectation and the variance of a Bernoulli distributed variable are:

$$\mathbb{E}[x] = \mu \tag{1.26}$$

$$\mathrm{var}[x] = \mu(1-\mu). \tag{1.27}$$

Binomial distribution

The binomial distribution describes the probability of observing m occurrences of $x = 1$ in a set of N samples from a Bernoulli distribution, where the probability of observing $x = 1$ is μ:

$$\mathrm{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}. \tag{1.28}$$

The expectation and the variance of a binomially distributed variable are:

$$\mathbb{E}[x] = N\mu \tag{1.29}$$

$$\mathrm{var}[x] = N\mu(1-\mu). \tag{1.30}$$

Negative Binomial distribution

The negative binomial distribution is closely related to the Bernoulli distribution. It is a discrete probability distribution of the number of successes in a sequence of independently and identically distributed Bernoulli random variables before a specified, non-random number of failures r occurs. It has two parameters r, which is the number of failures until the experiment is stopped, and p, which is the success probability in each experiment:

$$\text{Neg-Bin}(x|r,p) = \binom{x+r-1}{r-1}\left(\frac{p}{p+1}\right)^{r}\left(\frac{1}{p+1}\right)^{x}. \tag{1.31}$$

The mean and the variance of a negative binomial random variable are:

$$\mathbb{E}[x] = \frac{r}{p} \tag{1.32}$$

$$\text{var}[x] = \frac{r}{p^2}(p+1). \tag{1.33}$$

Multinomial distribution

The multinomial distribution is a multivariate generalisation of the binomial distribution and gives the distribution over counts m_k for a K-state discrete variable to be in state k given a total number of observations N:

$$\text{Mult}(m_1, m_2, \ldots, m_K|\mu, N) = \binom{N}{m_1 m_2 \ldots m_K}\prod_{k=1}^{K}\mu_k^{m_k}, \tag{1.34}$$

where μ_k is the probability that $x_k = 1$, $\mu_k \in [0,1]$ and $\sum_{k=1}^{K}\mu_k = 1$. The expectation and the variance of a multinomially distributed random variable are:

$$\mathbb{E}[m_k] = N\mu_k \tag{1.35}$$

$$\text{var}[m_k] = N\mu_k(1-\mu_k). \tag{1.36}$$

Categorical distribution

The categorical distribution, also known as multinoulli distribution, is a discrete probability distribution describing the possible results of a random variable that can take on one of K possible categories.

$$\text{Cat}(x|p_1, \ldots, p_k) = \prod_{i=1}^{k} p_i^{\mathbb{I}[x=i]}, \tag{1.37}$$

where p_i is the probability that $x = i$ such that $0 \leq p_i \leq 1$ and $\sum_{i=1}^{k} p_i = 1$, and $\mathbb{I}$ is the indicator function.

Poisson distribution

The Poisson distribution is the probability of a given number of random events occurring in a fixed interval of time. The shape of the distribution is governed by the rate of occurrence $\lambda > 0$, which is also the expected rate of occurrence:

$$\text{Poi}(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda). \tag{1.38}$$

The expectation and variance of Poisson-distributed random variable are the same:

$$\mathbb{E}[x] = \lambda \tag{1.39}$$

$$\text{var}[x] = \lambda. \tag{1.40}$$

Beta distribution

The beta distribution is conjugate to the Bernoulli and binomial distributions. It is a distribution over a continuous variable $x \in [0, 1]$. It has two shape parameters $\alpha, \beta > 0$

$$\text{Beta}(x|a, b) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \tag{1.41}$$

where $\Gamma(x)$ is defined by $\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u)\, du$. We have that

$$\mathbb{E}[x] = \frac{\alpha}{\alpha + \beta} \tag{1.42}$$

$$\text{var}[x] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{1.43}$$

Gamma distribution

The Gamma distribution is conjugate to the Poisson distribution. It is a distribution over a continuous random variable $x > 0$, and has a shape α and scale β parameters:

$$\text{Ga}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x). \tag{1.44}$$

The corresponding expectation and variance are:

$$\mathbb{E}[x] = \frac{\alpha}{\beta} \tag{1.45}$$

$$\mathrm{var}[x] = \frac{\alpha}{\beta^2}. \tag{1.46}$$

Dirichlet distribution

The Dirichlet distribution is a continuous multivariate distribution, generalising the beta distribution. It is conjugate to both the categorical and multinomial distributions. The shape of the distribution is governed by concentration parameters $\alpha_1, \ldots, \alpha_k > 0$:

$$p(\mathbf{x}) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \ldots x_k^{\alpha_k - 1}, \tag{1.47}$$

where $\mathbf{x} = (x_1, \ldots, x_k)$ is a k-dimensional vector with $0 \leq x_k \leq 1$ and $\sum_k x_k = 1$. The expectation and variance have the following forms:

$$\mathbb{E}[x_i] = \frac{\alpha_i}{\alpha_0} \tag{1.48}$$

$$\mathrm{var}[x_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \tag{1.49}$$

where $\alpha_0 = \sum_i \alpha_i$.

Gaussian distribution

The Gaussian distribution is widely used to model the distribution of continuous variables because of the **Central limit theorem**, which states that under some mild conditions, the sum of random variables has a distribution that becomes increasingly Gaussian with the increase in the number of terms in the sum [Walker, 1969]. In the case of a single variable x, the distribution is governed by a mean μ and variance σ^2:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \tag{1.50}$$

The corresponding expectation and variance are:

$$\mathbb{E}[x] = \mu \tag{1.51}$$

$$\mathrm{var}[x] = \sigma^2. \tag{1.52}$$

For a D-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution is of the form:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \tag{1.53}$$

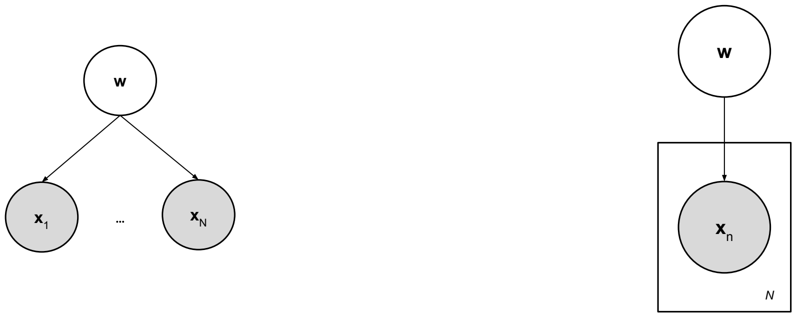
where $\boldsymbol{\mu}$ is the D-dimensional mean vector and $\boldsymbol{\Sigma}$ is the $D \times D$ covariance matrix. We have that

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \tag{1.54}$$

$$\text{cov}[\mathbf{x}] = \boldsymbol{\Sigma}. \tag{1.55}$$

1.2.4 Graphical models

Throughout this thesis we use graphical models to compactly illustrate assumptions between the variables in some of the models we consider. A **graphical model** is a way of representing a joint distribution by making conditional independence assumptions. The nodes represent the random variables. The graphical models used in this thesis are directed in order to illustrate some of the model assumptions we make.



(a) A simple directed graphical model.

(b) A compact representation of a directed graphical model.

Figure 1.2: Different ways of representing a graphical model. The shaded nodes $x_1, \ldots, x_N$ represent the observations, and the unshaded nodes $\mathbf{w}$ linked to them - any model parameters and unobserved variables.

Figure 1.2 represents a typical directed graphical model. The shaded nodes x_i in Figure 1.2b are the observed variables $X = \{x_1, \ldots, x_N\}$ generated i.i.d from a probabilistic model with parameters $\mathbf{w}$, usually represented by unshaded nodes. A directed graphical model often includes unobserved, latent variables, such as cluster indicators, component means, and this graphical representation aids their inference.

1.2.5 Inference

The methods we use to perform inference in the experiments in this thesis rely on simulated annealing. To highlight the reasons behind our choice of inference scheme, we first introduce the Markov Chain Monte Carlo methods to which simulated annealing is closely related.

Markov Chain Monte Carlo (MCMC) methods are a popular way to infer model parameters. They simulate a Markov chain whose stationary distribution is equal to a target distribution, in our case the posterior distributions, from which it is usually hard to sample. Here we present a short summary of the most commonly used MCMC techniques with a particular focus on the ones used in the thesis.

The main idea of **Gibbs Sampling** [Geman and Geman, 1987] is to approximate a distribution with a set of samples. The theory around Gibbs sampling implies that we can sample from a joint distribution by sampling sufficiently many times from the conditional distributions of each variable. We use it when we can derive the conditional distributions $p(\theta_i|\theta_{j_{j\neq i}})$ of the parameters we want to sample. For example, if we have K variables, we can update one variable at a time at iteration $t+1$ as follows:

$$\begin{aligned}
\theta_1^{(t+1)} &\sim p(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_K^{(t)}) \\
\theta_2^{(t+1)} &\sim p(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_K^{(t)}) \\
&\vdots \\
\theta_K^{(t+1)} &\sim p(\theta_K|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{K-1}^{(t+1)}).
\end{aligned}$$

Here the target distribution is the joint posterior and the conditional distributions are required to sample from it.

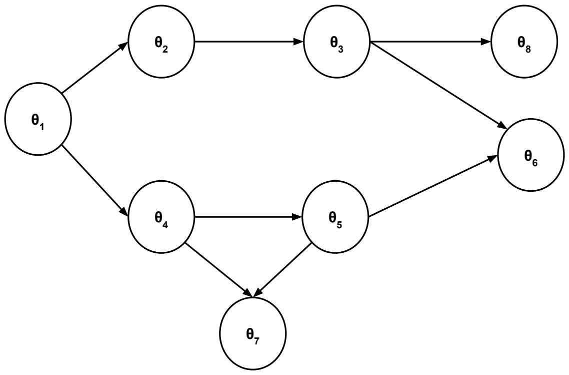


Figure 1.3: A directed graphical model

To compute the conditional distributions, we need to use only the concept of **Markov blanket** [Pearl, 2014], which is defined as the smallest set of nodes that makes a node conditionally independent of all the other nodes in the graphical model. This implies that the conditional distribution $p(\theta_i|\theta_{j_{j\neq i}})$ depends only on the nodes in the neighbourhood of θ_i, i.e. the parents of θ_i, the children of θ_i and the other parents of the children of θ_i. For example, the Markov blanket of θ_5 in the model presented on Figure 1.3 is $\text{mb}(5) = \{4\} \cup \{6, 7\} \cup \{3\} = \{3, 4, 6, 7\}$ and we get that

$$p(\theta_5|\theta_{-5}) \propto p(\theta_5|\theta_4)p(\theta_6|\theta_3, \theta_5)p(\theta_7|\theta_4, \theta_5) \tag{1.56}$$

Another popular MCMC method is **Metropolis sampling** [Metropolis et al., 1953], which is an adaptation of random walk that uses an acceptance/rejection rule to converge to the specified target distribution. The algorithm proceeds as follows:

Algorithm 1.1: Metropolis sampling

```
Draw a starting point $\theta^0$ from a starting distribution $p_0(\theta)$ ;
for $t = 1, 2, \ldots,$ do
    a) sample a proposal $\theta^*$ from a proposal distribution $J_t(\theta^*|\theta^{t-1})$, which
       must be symmetric ($J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a)$) ;
    b) Calculate $r = \frac{p(\theta^*|x)}{p(\theta^{t-1}|x)}$ ;
    c) Set $\theta^t$ to $\theta^*$ with probability $\min(r, 1)$ and to $\theta^{t-1}$ otherwise.
end
```

We have used the **Metropolis Hastings algorithm** [Robert and Casella, 1999; Hastings, 1970] as well, which generalises the Metropolis sampler by allowing the jumping rules J_t not to be symmetric and by correcting for asymmetry.

Algorithm 1.2: Metropolis Hastings sampling

```
Draw a starting point $\theta^0$ from a starting distribution $p_0(\theta)$ ;
for $t = 1, 2, \dots,$ do
    a) sample a proposal $\theta^*$ from a proposal distribution $J_t(\theta^*|\theta^{t-1})$ ;
    b) Calculate $r = \frac{p(\theta^*|x)}{p(\theta^{t-1}|x)} \frac{J_t(\theta^{t-1}|\theta^*)}{J_t(\theta^*|\theta^{t-1})}$ ;
    c) Set $\theta^t$ to $\theta^*$ with probability $\min(r, 1)$ and to $\theta^{t-1}$ otherwise.
end
```

There are a few practical issues that need to be considered when working with MCMC algorithms. If we run the chain for a sufficiently long time, we will eventually obtain samples from the posterior distribution. But how long is sufficiently long and can we predict how long it will take to equilibrate?

Another important issue is detecting convergence. Cowles and Carlin [1996] describe the practical tools available to diagnose convergence. They point out that the diagnostic tools often fail to detect convergence, could be difficult to implement and often require problem-specific coding. The authors recommend using a variety of diagnostic tools instead, and running a few parallel chains with starting points picked systematically.

As using MCMC methods often involve time-consuming intensive simulations, we would ideally want to use methods which converge faster. One such method is **simulated annealing** [Kirkpatrick et al., 1983]. It is an optimisation method inspired by statistical physics, and can be used to find an approximation to the global maximum of the posterior distribution, as opposed to the MCMC methods which try to sample from the posterior distribution.

Similarly to Metropolis Hastings, we sample a new state according to a proposal distribution $\boldsymbol{\theta'} \sim q(.|\boldsymbol{\theta}_k)$, which is often a random walk proposal $\boldsymbol{\theta'} = \boldsymbol{\theta}_k + \boldsymbol{\epsilon}_k$, where $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ for real-valued parameters. After the proposal of the new state, we compute

$$\alpha = \exp((f(\boldsymbol{\theta'}) - f(\boldsymbol{\theta}))/T_k), \tag{1.57}$$

where f is the function we want to optimise and T_k is the computational temperature. We accept the new state and set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta'}$ with probability $\min(1, \alpha)$, otherwise we stay in the current state and set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta'}$. This means that if the new state is more probable (and has lower energy), we will definitely accept it. But if it is less probable (has higher energy), we might still accept it depending on the current temperature. In this way, simulated annealing deals with the issue of getting stuck in a local minimum/maximum - it will sometimes accept worse solutions than the current one in order to explore the sample space.

The exploration of the parameter space is governed by the cooling schedule, which is the rate at which the temperature changes over time. Kirkpatrick et al. [1983] show that if the temperature is cooled sufficiently slowly, then the algorithm will probably find the global optimum. It is, however, difficult to quantify ‘sufficiently slowly’, which is the main drawback of the algorithm.

There are different cooling schedules that can be used. It is common to use the *exponential cooling schedule* of the following form $T_k = T_0 C^k$, where T_0 is the initial temperature (often $T_0 \approx 1$) and C is the cooling rate (often C is between 0.9 and 0.99). Another cooling schedule is the *logarithmic* one of the form $T_k = \frac{1}{\log(k)}$. We performed numerical experiments and the final schedule we considered in this thesis is of the form $T_k = T_0 \times 0.95^k$, with T_0 being either 100 or 1000 which we found to be good choices in order to allow the parameter space to be well explored.

1.3 Clustering methods

1.3.1 Clustering methods for continuous data

Clustering methods can be broadly divided into distance-based and model-based.

K-means clustering [Hartigan and Wong, 1979] is a method which clusters dataset $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with N observations into K clusters, where K has to be specified by the user. The objective is to minimise the sum of squares of the distances between each data point $\mathbf{x}_n$ and the assigned cluster centre,

$$J = \sum_{n=1}^{N} \sum_{\mathbf{x}_n \in C_j} (\mathbf{x}_n - \boldsymbol{\mu}_j)^2, \tag{1.58}$$

where C_j are the data clusters and $\boldsymbol{\mu}_j$ are their respective means/centres. The algorithm is initialised by randomly selecting the K centres and assigning the data points to the closest centre. Then iteratively, it recalculates the cluster centres based on the assignments and re-assigns the data points to the new nearest centre until convergence. Usually the algorithm is run for a wide range of values for K, the within-cluster sum of squares (WSS) is calculated for each K and the ‘optimal’ K is chosen with the ‘elbow method’ such that there is no significant decrease in

WSS after that (see Figure 1.4 for illustration of the elbow method). There have been different statistics proposed to replace the 'elbow method' such as the Calinski-Harabasz index [Caliński and Harabasz, 1974] and the gap statistic [Tibshirani et al., 2001].

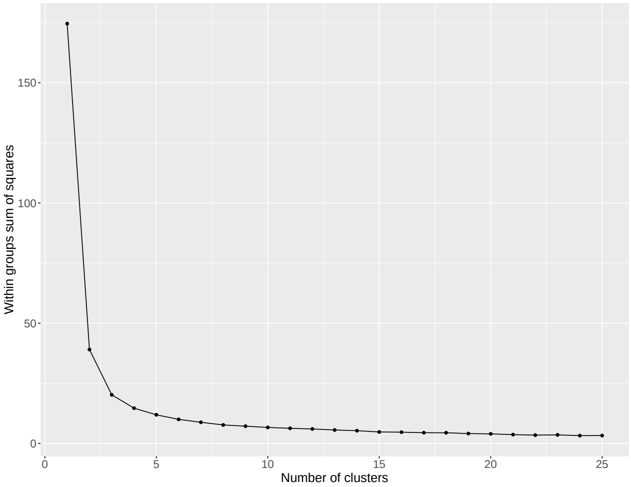


Figure 1.4: Illustration of the elbow method for selection of the number of clusters in k-means clustering. We applied k-means clustering to the iris dataset [Fisher, 1936]. The number of clusters is plotted against the total within groups sum of squares. The number of clusters is determined by the point after which there is no significant decrease in the within-cluster sum of squares. In this example, the optimal number of clusters is chosen to be 2.

There are two main approaches of performing **hierarchical clustering** [Duda et al., 1995]: bottom-up (agglomerative) and top-bottom (divisive). They both take as an input a dissimilarity matrix, which represents the pairwise distances between observations in the dataset, and produce a clustering. Agglomerative clustering starts with N groups, each initially containing one object, and at each step it merges the two most similar groups until there is a single group containing all the data. There are three common forms of agglomerative clustering based on the similarity rule used: *single linkage*, *complete linkage*, and *average linkage*. In single linkage, the two clusters with the closest pair of elements are combined, whereas in complete linkage, the two clusters separated by the furthest pair of elements are merged. In average linkage hierarchical clustering, the two nearest clusters A and B are combined, with the distance being the average of all distances between all pairs of objects in A and objects in B.

Divisive clustering starts with all the points in a single cluster and at each step,

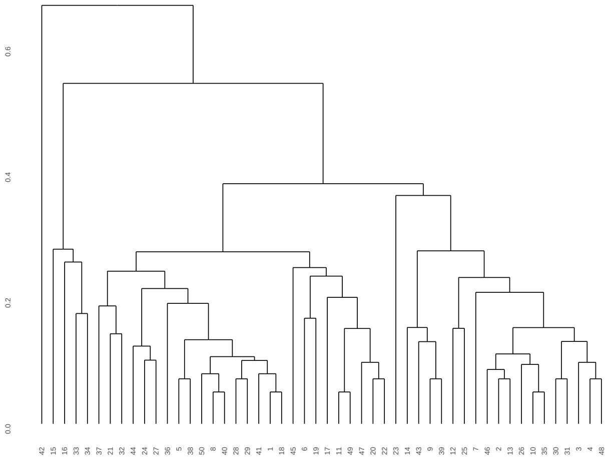


Figure 1.5: An example of a dendrogram. This was constructed by applying average linkage agglomerative clustering to 50 observations from the iris dataset.

using a splitting rule, splits a cluster into two clusters. One way of choosing the cluster to split is to pick the cluster with the largest diameter (distance between the two furthest points in the cluster) and split it into two clusters using K-means with $K = 2$, which is called bisecting K-means [Steinbach et al., 2000].

The merging/splitting process in hierarchical clustering can be represented in a dendrogram (binary tree) (see Figure 1.5 for an example of a dendrogram), where the initial objects are leaves and every time they are merged, we join them in the tree. The root node represents the group containing all the data, and the height of the branches represents the dissimilarity between the groups to be joined. The number of groups is picked in a similar way as in k-means clustering.

Mixture models [McLachlan and Peel, 2004] produce clustering, based on a probability model of the data. In finite mixture modelling, we assume that there are K clusters (with parameters θ_k) and each of observations belongs to one of them. Finite mixture models can accommodate different types of data by changing the data generating distribution and have the general form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}|\theta_k), \tag{1.59}$$

where π_k are the mixture proportions with $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$, and $f(\mathbf{x}|\theta_k)$ is the selected distribution. The clusters are usually modelled by members of the same parametric density family. We can use any distribution for the $f_k(\mathbf{X})$, with the most common mixture model being the **mixture of Gaussians** of the form:

$$p(\mathbf{X}) = \sum_{C} p(C)p(\mathbf{X}|C) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{1.60}$$

where $0 \leq \pi_k \leq 1$, is the mixture proportion for the k^{th} component, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the parameters for the k^{th} component, and $C = \{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ is the collection of n cluster labels $c_i \in \{1, \dots, K\}$, indicating the cluster membership of the ith observation x_i. **Bayesian mixture models** contain a prior over the mixing distribution and a prior over the cluster parameters.

The Expectation Maximisation (EM) algorithm [Dempster et al., 1977] is the most commonly used technique for estimating the parameters of a mixture model, for example Lee and McLachlan [2014]; O'Hagan et al. [2012]; Steiner and Hudec [2007], but one can also use MCMC methods, such as Gibbs sampling, to infer the cluster indicators C, the mixture proportions π_k and the cluster parameters (mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$). This is usually done by specifying priors for all model parameters and then estimating the posterior distributions which would become the target distributions in Gibbs sampling. A Dirichlet prior is usually placed on the mixture weights π_k, and inverse-Wishart and Gaussian priors are popular choice for priors on the cluster covariance and mean, respectively.

As we will see later in this thesis, mixture models are particularly useful when working with heterogeneous data with non-i.i.d.[4] structure. However, they face computational and statistical difficulties when the data is very high-dimensional Friedman et al. [2008]; Zhao et al. [2012]. Penalised approaches such as [Städler et al., 2017; Zhou et al., 2009] using lasso and graphical lasso penalties have been applied to mixtures in high dimensions. A recently proposed model called model-based clustering via adaptive projections (MCAP) Taschler et al. [2019] takes a different approach and instead of estimating the mixtures in the original space, it models a low-dimensional representation of the data obtained by a linear projection. The key idea behind MCAP is to achieve a bias-variance tradeoff controlled by the projection dimension q, which is set in a data-adaptive manner using a stability-based score derived from clustering subsets of the original data.

The projection dimension itself plays an important role and governs a type of bias-variance tradeoff with respect to recovery of the relevant signals. MCAP sets the projection dimension automatically in a data-adaptive manner, using a proxy for the assignment risk. Combining a full covariance formulation with the adaptive projec-

[4] independent and identically distributed

tion allows detection of both mean and covariance signals in very high dimensional problems.

Spectral clustering [Ng et al., 2002] views clustering in terms of graph cuts. It creates a weighted undirected graph $\mathbf{W}$ from a dissimilarity matrix $\mathbf{S}$ and tries to find a partition into K clusters $A_1, \ldots, A_K$ which minimises

$$cut(A_1, \ldots, A_K) = \frac{1}{2} \sum_{k=1}^{K} W(A_k, \bar{A}_k), \tag{1.61}$$

where $\bar{A}_k$ is the complement of A and $W(A, B) = \sum_{i \in A, j \in B} w_{ij}$. This, however, may lead to clusters with single points. To avoid this, the normalised cut can be minimised instead:

$$normalised\ cut(A_1, \ldots, A_K) = \frac{1}{2} \sum_{k=1}^{K} \frac{cut(A_k, \bar{A}_k)}{vol(A_k)}, \tag{1.62}$$

where $vol(A_k) = \sum_{i \in A} d_i$ and $d_i = \sum_{j=1}^{N} w_{ij}$ is the weighted degree of node i.

1.3.2 Clustering methods for categorical data

The lack of an inherent ordering or geometric distance in categorical data prohibits the application of the above deterministic clustering methods with categorical data. Most of the models for discrete data are focused on developing a measure/distance for categorical data. Here we present a short summary of the most widely used methods, which could be broadly classified into 3 groups: using **overlap-based similarity measure**, **context-based similarity measure** or **information-theoretic clustering criterion**.

The methods using overlap-based similarity measure, such as **k-modes** [Huang, 1998] and **ROCK** [Guha et al., 2000], compare the overlap between the observations. K-modes uses a matching dissimilarity measure for categorical objects, which is the total number of mismatches of the corresponding attributes of two categorical objects X and Y. The algorithm then finds the modes, which are defined as the vectors that minimise the dissimilarity measure between the vectors and the clusters, in a similar fashion to k-means. Guha et al. [2000] introduce a robust clustering algorithm for categorical data called ROCK, which is an adaptation of agglomerative hierarchical clustering for categorical data. It heuristically optimises a criterion function defined in terms of the number of 'links' between data points, which is the number of common neighbours they have in the dataset. Starting with each data point in its own cluster, the two closest clusters are merged until the required number of clusters is reached.

The methods using context-based similarity measure, such as **CACTUS** [Ganti et al., 1999] and **STIRR** [Gibson et al., 1998], compare the context in which the attributes of the observations appear. For two categorical attribute values, the context is defined as the values of other attributes with which they co-occur in the dataset. The idea behind CACTUS (Clustering Categorical Data Using Summaries) is that a summary of the entire dataset is sufficient to compute a set of 'candidate' clusters which can then be validated to determine the actual set of clusters. STIRR is an iterative algorithm based on non-linear dynamical systems, where each attribute value is represented as a weighted vertex in a graph. Multiple copies $b_1, \ldots, b_m$, called basins, of this set of weighted vertices are maintained, with b_1 being a principal basin and $b_2, \ldots, b_m$ - non-principal basins. Starting with a set of weights on all vertices, the system is iterated until a fixed point is reached. The authors argue that when the fixed point is reached, the weights in one or more of the basins $b_2, \ldots, b_m$ isolate two groups of attribute values on each attribute: the first with large positive weights and the second with small negative weights, and that these groups correspond to projections of clusters on the attribute. However, the automatic identification of such sets involves a non-trivial post-processing step, which also makes identifying the clusters a very difficult task. Algorithms such as **COOLCAT** [Barbará et al., 2002] which use an information theoretic criterion aim to generate clusters with low entropy as this implies that the clusters are homogeneous. Given a set of clusters, COOLCAT places the next point in the cluster which minimises the overall expected entropy. It acts incrementally and is able to cluster every new point without having to reprocess the entire set.

1.3.3 Clustering methods for mixed data

Clustering mixed data (e.g. datasets comprising both continuous and categorical measurements) possesses similar challenges to clustering categorical data in addition to the challenge of modelling different types of data. Here we present a short summary of the most widely used techniques.

k-prototypes [Huang, 1997], is an extension of k-means clustering to mixed data and dynamically updates the k prototypes (cluster centres) in a similar fashion to k-means in order to maximise the intra-cluster similarity of objects. The object similarity measure is derived from both numeric and categorical attributes, with the similarity measures being the squared Euclidean distance and the number of mismatches between object and cluster prototypes. This algorithm can be used with large datasets with many features. Huang et al. [2005] modify the algorithm by incorporating weights on the features, based on the importance of the features in generating the clustering result (small weights reduce the effect of noisy variables). Modha and Spangler [2003] similarly use weights on the features but use squared Euclidean distance for the continuous variables and cosine distance for the categorical features.

Another popular algorithm is **similarity based agglomerative clustering** (SBAC) [Li and Biswas, 2002]. It uses a similarity measure that gives greater weight to uncommon attribute value matches and outputs the partition using agglomerative clustering. Philip and Ottaway [1983] apply agglomerative clustering approach as well in which the similarity matrix is computed using Gower's similarity measure [5].

Another approach to clustering mixed data is to use an ensemble method. One such method is **cluster ensemble based mixed data clustering** (CEBMDC), developed by He et al. [2005]. It applies numeric and categorical clustering algorithms to the data, which is divided into 2 datasets, numeric and categorical, and then finds a final partition by using cluster ensembling, which combines several runs of different clustering algorithms to obtain a common partition. Here the cluster ensemble problem is transformed into a categorical data clustering problem, where the partitions of the continuous and categorical datasets are combined into a dataset which is then clustered with a method for clustering categorical data. The authors use the Squeezer algorithm introduced in He et al. [2002] which adds an observation to a cluster based on a similarity measure related to cluster summary statistics.

Liverani et al. [2015] present an alternative to these approaches by implementing a Dirichlet process mixture model in their R package **PReMiuM** [Liverani et al., 2015]. The model performs profile regression which nonparametrically links a (binary, categorical, count and continuous) response variable to (continuous and discrete) covariate data through the cluster membership. The model allows as well for the modelling of missing data.

[5] this involves dividing the features into two subsets: one for the categorical data, and another for the numeric features [Gower, 1971]

1.3.4 Nonparametric Bayesian methods for clustering

A common challenge with the clustering models that we introduced is that they require the choice of the number of clusters K in advance. This is why we are interested in models in which we do not specify K, but instead allow the complexity of the model to grow as more data is observed. The Bayesian nonparametric (BNP) approach to clustering achieves that by estimating the number of clusters from the data and by allowing future data to exhibit unseen clusters. The main assumption in this approach is that there is an infinite number of clusters, but only a finite number of clusters is used to generate the data. This is achieved by placing a prior $P(C)$ on the cluster indicators C that favours assigning the observations to only a small number of clusters. An example of such prior is called *Dirichlet process* [Antoniak, 1974; Escobar and West, 1995], which will present in more details as it will play an important role in the models we will develop in this thesis.

The **Dirichlet process** (DP) [Ferguson, 1973] defines a distribution on distributions. If Θ is a measurable space, then a Dirichlet process is parameterised by a base measure G_0 on Θ and a positive scalar concentration parameter α. It is characterised by the distribution it induces on finite measurable partitions of the parameter space. This means that if we divide the parameter space into measurable partitions, we want the probability distribution G on Θ to follow a Dirichlet distribution on each partition. We formalise this in the following theorem:

Theorem 1. *Let G_0 be a probability distribution on a measurable space Θ, and α be a positive scalar. Consider a finite partition $(T_1, \dots, T_K)$ of Θ, with $\cup_{k=1}^{K} T_k = \Theta$ and $T_k \cap T_l = \emptyset$. A random probability distribution G on Θ is drawn from a Dirichlet process if its measure on every finite partition $(T_1, \dots, T_k)$ of Θ follows a Dirichlet distribution*

$$(G(T_1), \dots, G(T_K)) \sim \mathit{Dir}(\alpha G_0(T_1), \dots, \alpha G_0(T_K)) \tag{1.63}$$

Proof. The characterisation in (1.63) is true if the probabilities add appropriately when a partition's cells are combined. This is guaranteed by the aggregation properties of finite Dirichlet distribution (1.64). Another way of proving the existence of Dirichlet process is by using Kolmogorov's consistency conditions [Ferguson, 1973] which are satisfied for any stochastic process. □

The **base measure** G_0 in a Dirichlet process specifies the mean of $\mathrm{DP}(\alpha, G_0)$ as $\mathbb{E}[G(T)] = G_0(T)$. The Dirichlet process draws distributions around the base distribution in the same way the normal distribution draws real numbers around its mean.

The **concentration parameter** α can be interpreted as an inverse variance since $\mathrm{Var}[G(T)] = \frac{G_0(T)(1-G_0(T)}{\alpha+1}$ [Teh, 2011]. The larger α, the smaller the variance and the mass of the DP concentrates around the mean.

This can be seen from the plots below (Figure 1.6) which show realisations of the Dirichlet distribution for different concentration parameter α. Higher values of α result in more spread out draws, whereas lower values of α give more concentrated draws.

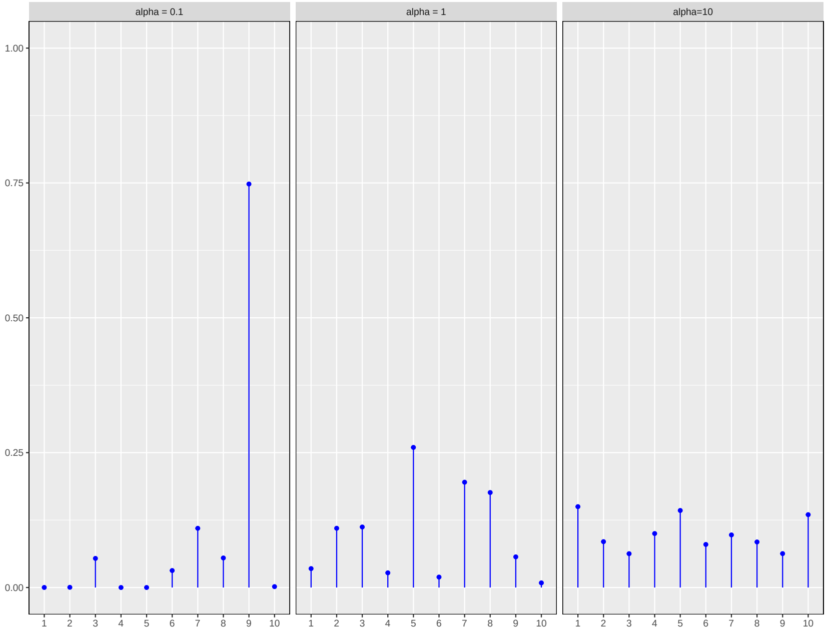


Figure 1.6: Samples from a Dirichlet distribution for $\alpha = \{0.1, 1, 10\}$.

One of the useful properties of the Dirichlet distribution is the *aggregation property*, which is particularly helpful for deriving the posterior and predictive distributions for the Dirichlet process. If $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha)$, then the multinomial parameters attained by aggregation also follow Dirichlet distribution. For example, adding the first two parameters results in

$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K). \tag{1.64}$$

Aggregation of any subset of the categories results in a Dirichlet distribution.

Posterior distribution

If the prior on G is a Dirichlet process ($p(G) = \text{DP}(\alpha, G_0)$), and we observe θ ($p(\theta|G) = G(\theta)$), then the posterior distribution is also a Dirichlet process because of the conjugacy of Dirichlet distribution:

$$(G(T_1), \ldots, G(T_K)|\theta \in T_k) \sim \text{Dir}(\alpha G_0(T_1), \ldots, \alpha G_0(T_k) + 1, \ldots, \alpha G_0(T_k)). \quad (1.65)$$

This can be extended to N observations:

$$p(G|\theta_1, \ldots, \theta_n) = \text{DP}(\alpha + n, \frac{\alpha}{\alpha + N} G_0 + \frac{1}{\alpha + N} \sum_{i=1}^{N} \delta_{\theta_i}), \quad (1.66)$$

which can be shown to be true using the conjugacy of finite Dirichlet distributions [Teh, 2011].

Other representations

The preceding subsections provided implicit representation of the Dirichlet process and outlined some of its properties but they did not provide a scheme for sampling from a Dirichlet process or an expression for its predictive distribution. We will now describe three different representations of the Dirichlet process: the **Chinese Restaurant Process** [Hjort et al., 2010],the **stick-breaking process** [Sethuraman, 1994] and the **Pólya Urn Model** [Blackwell and MacQueen, 1973], which have been popular in the Bayesian nonparamterics literature and play an important role in the computational methods for Dirichlet processes.

The implicit data partition property of the Dirichlet process evokes a comparison with the idea of the never-ending tables in San Francisco's Chinatown, named **Chinese restaurant process** (CRP) [Pitman et al., 2002; Hjort et al., 2010]. Let us imagine an initially empty restaurant with an infinite number of tables in it $K = 1, \ldots$, where only a finite number of them are going to be occupied. Customer 1 (with value ϕ_1) comes and by sitting down starts a group and sets the group/table parameter θ_1 for the rest of the group. After that customer 2 comes and joins customer 1 with probability $\frac{1}{\alpha+1}$ or sits on a new table with probability $\frac{\alpha}{\alpha+1}$. Similarly, the $N + 1$st customer sits down at a new table with probability $\frac{\alpha}{\alpha+N}$ or joins table k with probability $\frac{N_k}{N+\alpha}$, where N_k denotes the number of people already at table k and N is the total number of customers so far. In this way, we obtain a procedure for drawing samples from DP.

We summarise the algorithm for generating samples from a Chinese restaurant process below:

Algorithm 1.3: Generating samples from Chinese restaurant process

```
Customer 1 enters the restaurant and sits at table 1
$\phi_1 = \theta_1, K = 1, N = 1, N_k = 1$ ;
for $N = 2, \ldots,$ do
    customer $N$ sits at table $k$ with probability $\frac{N_k}{N-1+\alpha}$ and at table $K+1$ (new
      table) with probability $\frac{\alpha}{N-1+\alpha}$ ;
    if new table was chosen then
        $K \leftarrow K+1, \theta_{K+1} \sim G_0$ ;
    end
    set $\phi_N$ to $\theta_k$ of the table $k$ at which customer $N$ sat ;
    set $N_k \leftarrow N_{k+1}$
end
```

We can represent the same generation of samples in two analogous ways by using the concept of **Pólya Urn Model** [Blackwell and MacQueen, 1973] and the **stick-breaking process** [Sethuraman, 1994].

In the **Pólya Urn model**, we start with an urn containing $\alpha G_0(x)$ balls of colour x for each possible colour of x, with G_0 denoting the base distribution. At each step, we draw a ball from the urn, note its colour and then return it back in the urn together with another ball from the same colour. This generates samples from a Dirichlet process without having to construct the underlying mixture $G \sim \text{DP}(\alpha, G_0)$.

The following theorem summarises the model and the derivation of the predictive distribution.

Theorem 2. *Let $G \sim DP(\alpha, G_0)$ be distributed according to a Dirichlet process, where the base measure G_0 has corresponding density $g(\theta)$. If we consider a set of N observations $\bar{\theta}_i \sim G$ taking K distinct values $\{\theta\}_{k=1}^{K}$, then the predictive distribution of the next observation is equal to*

$$p(\bar{\theta}_{N+1} = \theta | \bar{\theta}_1, \ldots, \bar{\theta}_N, \alpha, G_0) = \frac{1}{\alpha + N}\left(\alpha g(\theta) + \sum_{k=1}^{K} N_k \delta(\theta, \theta_k)\right) \tag{1.67}$$

where N_k is the number of previous observations of θ_k.

The **stick-breaking process** provides another way of generating samples from a Dirichlet process. We will provide first the theorem which outlines the process.

Theorem 3. *Let $\pi = \{\pi_k\}_{k=1}^{\infty}$ be an infinite sequence of variables derived from the following stick-breaking process, with a parameter $\alpha > 0$:*

$$\beta_k \sim Beta(1, \alpha) \quad k = 1, 2, \dots \tag{1.68}$$

$$\pi_k = \beta_k \prod_{l=1}^{k} (1 - \beta_l) = \beta_k (1 - \sum_{l=1}^{k-1} \pi_l). \tag{1.69}$$

Given a base measure G_0 on Θ, consider the following discrete random measure:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k) \qquad \theta_k \sim G_0. \tag{1.70}$$

This construction guarantees that $G \sim DP(\alpha, G_0)$. Conversely, samples from a Dirichlet process are discrete and have a representation as in (1.9)

This construction is illustrated in Figure 1.7. Variables π_k partition the unit length stick, with the k^{th} variable π_k being a random proportion β_k of the remaining stick. We use $\pi \sim$ GEM(α) [Ishwaran and Zarepour, 2002; Pitman et al., 2002] to indicate the set of variables sampled from this process, where GEM is named after Griffiths, Engen and McCloskey.

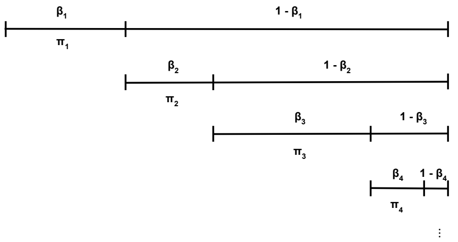


Figure 1.7: Diagram representing the stick-breaking process for sampling from a Dirichlet process

This construction provides an alternative representation of the concentration parameter α. As $\beta_k \sim \text{Beta}(1, \alpha)$, then we have that

$$\mathbb{E}[\beta_k] = \frac{1}{1 + \frac{\alpha}{1}}. \tag{1.71}$$

larger α.

These three different representations of the Dirichlet process illustrate not only the discreteness property of draws from a DP, but also the clustering property of DP.

The unique values of $\theta_1, \ldots, \theta_N$ induce a partitioning of $\{1, \ldots, N\}$ into clusters, with each cluster taking the same value θ_k^*.

Exchangeability

It is important to note that the cluster assignments under the CRP distributions are exchangeable as this enables inference in models which use DPs. The property follows immediately from de Finetti's theorem [De Finetti, 1937; Diaconis, 1977] and means that the joint distribution

$$p(c_1, c_2, \ldots, c_N) = p(c_1)p(c_2|c_1)p(c_3|c_1, c_2) \ldots p(c_N|c_1, c_2, \ldots, c_{N-1})$$

is independent of the order of which the observations are assigned to clusters. We will now show why this is the case.

If I_k denotes the set of indices of customers assigned to the kth group, K is the number of occupied groups and N_k is the number of customers assigned to the k^{th} group, then the product of terms that correspond to the customers in the k^{th} group $\frac{\alpha.1.2...(N_k-1)}{(I_{k,1}-1+\alpha)(I_{k,2}-1+\alpha)...(I_{k,N_k}-1+\alpha)}$ can be derived as follows: the first customer in group k contributes $\frac{\alpha}{I_{k,1}-1+\alpha}$ as he/she starts a new cluster, the second customer contributes $\frac{1}{I_{k,2}-1+\alpha}$, the third one - $\frac{2}{I_{k,3}-1+\alpha}$ and so on. Hence, we can rewrite the joint distribution

$$\begin{aligned} p(c_{1:N}) &= \prod_{k=1}^{K} \frac{\alpha(N_k - 1)!}{(I_{k,1} - 1 + \alpha)(I_{k,2} - 1 + \alpha) \ldots (I_{k,N} - 1 + \alpha)} \\ &= \frac{\alpha^K \prod_{k=1}^{K}(N_k - 1)!}{\prod_{i=1}^{N}(i - 1 + \alpha)}, \end{aligned} \tag{1.72}$$

which implies the exchangeability.

In addition to the exchangeability, the probability of starting a new group depends on the concentration parameter α - lower α corresponds to fewer clusters (a priori), whereas the higher α leads to more clusters. It can be shown that the number of occupied tables K almost surely approaches $\alpha \log(N)$ as $N \to \infty$ [Petrone and Raftery, 1997; Pitman et al., 2002; Müller and Mitra, 2013].

1.3.5 Clustering methods for data integration

The advances of measurement technologies such as sequencing the human genome [National Human Genome Research Institute, 2018] (Figure 1.8) has led to the availability of very detailed and precise genomic information about large cohorts of cancer patients. This data has provided many insights in the changes occurring in different types of cancer and in the different cancer subtypes [Cancer Genome Atlas Network and others, 2012a,b; Shen et al., 2009; Gabasova et al., 2017; Argelaguet et al., 2018].

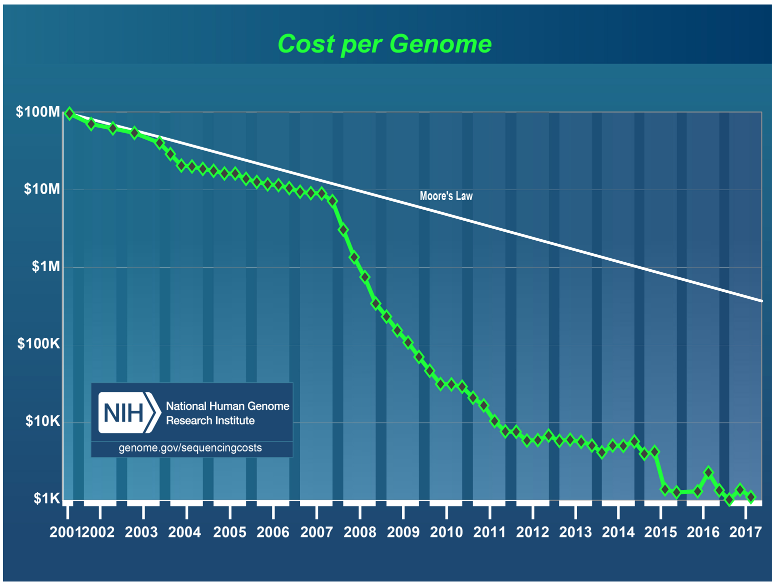


Figure 1.8: The cost of sequencing human genome. Credit: National Human Genome Research Institute [2018].

In order to model the high dimensionality and complexity of omics data appropriately, many novel methods have been developed, for example Shen et al. [2009]; Savage et al. [2010]; Yuan et al. [2011]; Lock et al. [2013]; Lock and Dunson [2013]; Kirk et al. [2012]. One of the most commonly used approaches for modelling multi-source input is to separately cluster each data type and then manually integrate the results [Hoadley et al., 2014]. This, however, does not model the interactions between the different data types and leads to inconsistent clustering. Here we are interested in the development of integrative approaches which allow joint inference (across all datasets) and which identify a single clustering structure. There are two major challenges to the development of such approach. To capture both concordant and unique alterations across data types, separate modelling of the covariance between data types and the variance-covariance structure within data types is required. The second challenge is incorporating dimensionality reduction, which is a key to the feasibility and performance of integrative clustering approaches when modelling high-dimensional data.

Shen et al. [2009] address these challenges using the connection between k-means clustering and latent variable models. Their model, **iCluster**, is based on a joint latent variable model, which models tumour subtypes as clusters in an unobserved (latent) space which can be estimated simultaneously from all available data type.

iCluster jointly estimates the latent variables $\mathbf{Z}$ from, for example, copy number variation data $\mathbf{X}_1$, methylation data $\mathbf{X}_2$, gene expression data $\mathbf{X}_3$ and other continuous genomic datasets. The mathematical form of the model is as follows:

$$\begin{aligned}\mathbf{X}_1 &= \mathbf{W}_1\mathbf{Z} + \boldsymbol{\varepsilon}_1\\ \mathbf{X}_2 &= \mathbf{W}_2\mathbf{Z} + \boldsymbol{\varepsilon}_2\\ &\vdots\\ \mathbf{X}_t &= \mathbf{W}_t\mathbf{Z} + \boldsymbol{\varepsilon}_t,\end{aligned}$$

where t is the number of different datasets, $\mathbf{Z}$ denotes the latent variables which induce dependencies across all data types and represent the underlying driving factors than can be used for disease subtype assignment, $\mathbf{W}_.$ are the loading matrices which project the data onto a lower dimension space, and $\boldsymbol{\varepsilon}_.$ are the independent error terms which represent any unaccounted variances. An Expectation-Maximisation algorithm, which alternates between computing the expected value of the complete-data log-likelihood with respect to $\mathbf{Z}$ given $\mathbf{X}$ and the current estimates of $\mathbf{W}.$ and $\boldsymbol{\varepsilon}_.$ in the E-step, and updating $\mathbf{W}.$ and $\boldsymbol{\varepsilon}_.$ in the M-step is used to infer the model parameters. Once convergence is reached, the posterior mean $\mathrm{E}[\mathbf{Z}|\mathbf{X}]$ is computed and the final partition is found by applying k-means to $\mathrm{E}[\mathbf{Z}|\mathbf{X}]$. The model selection is performed using a proportion of deviance (POD) metric, defined in terms of the cluster separability. The model closest to having perfectly separated clusters, and thus with the lowest POD, is chosen as the final model. Shen et al. [2009] deal with the high dimensionality of the data by using a lasso penalty [Tibshirani, 1996] to perform variable selection. This also reduces the variance of the model and leads to better clustering performance.

To address the limitation of iCluster to modelling only continuous data, **iClusterPlus** [Mo et al., 2013] was developed to integrate different data types (binary, categorical, continuous).

In iClusterPlus, the genomic variables x_{ijt} (i^{th} sample, j^{th} genomic feature, t^{th} data type) are connected to the latent process via a parametric joint model in which different genomic features are correlated through $\mathbf{z}_i$. If x_{ijt} is a binary variable, for example, mutation or gender, it is modelled by a logistic regression

$$\log \frac{p(x_{ijt} = 1|\mathbf{z}_i)}{1 - p(x_{ijt} = 1|\mathbf{z}_i)} = \alpha_{jt} + \beta_{jt}\mathbf{z}_i, \tag{1.73}$$

where $p(x_{ijt} = 1|\mathbf{z}_i)$ is the probability of gene j being mutated in patient i given the value of the latent factor $\mathbf{z}_i$, α_{jt} is an intercept term, and β_{jt} is a length-k row vector of coefficients that determine the weights genomic variable j contributes to the latent variables.

If x_{ijt} is a multicategory variable, a multilogit regression is used to model it:

$$P(x_{ijt} = c|\mathbf{z}_i) = \frac{\exp(\alpha_{jct} + \beta_{jct}\mathbf{z}_i)}{\sum_{l=1}^{C} \exp(\alpha_{jlt} + \beta_{jlt}\mathbf{z}_i)}, \tag{1.74}$$

where the coefficients have similar interpretation to the logit case.

If x_{ijt} is a continuous variable, then it is modelled by a linear regression to follow a Normal distribution:

$$x_{ijt} = \alpha_{jt} + \beta_{jt}\mathbf{z}_i + \epsilon_{ijt}, \tag{1.75}$$

where the error terms are uncorrelated.

Finally, if x_{ijt} is a count variable, it is modelled by a Poisson regression:

$$\log(\lambda(x_{ijt}|z_i)) = \alpha_{jt} + \beta_{jt}\mathbf{z}_i \tag{1.76}$$

where $\lambda(x_{ijt}|\mathbf{z}_i)$ is the conditional mean of the count given $\mathbf{z}_i$.

Similarly to iCluster, the lasso (L_1-norm) penalty is applied in iClusterPlus to identify the genomic variables which make important contributions to the latent process. A modified Monte Carlo Newton-Raphson algorithm has been used to address the intractable joint log-likelihood when different types of data are modelled. In addition, the number of clusters is determined by a deviance ratio metric that can be interpreted as the percentage of total variation explained by the correct model. The optimal number of clusters is determined by the point of transition after which there is no significant change in the deviance ratio. Although iClusterPlus deals well with the challenge of modelling different types of data, the statistical inference is not very straightforward. This is due to the fact that for each number of principal components, an extensive parameter search needs to be performed in order to determine the optimal penalty parameter values.

The authors of **iCluster** and **iClusterPlus** address the disadvantages of the models by developing **iClusterBayes** [Mo et al., 2017], which not only jointly models omics data of continuous and discrete data types but also improves on the inference and computational speed. It models a continuous variable y_{ijt} in the tth dataset by a linear regression

$$y_{ijt} = \mathbf{x}_i \mathbf{\Gamma}_{jt} \boldsymbol{\beta}_{jt} + \boldsymbol{\varepsilon}_{ijt}, \tag{1.77}$$

where $\boldsymbol{\beta}_{jt} = (\beta_{0jt}, \ldots, \beta_{kjt})^{\intercal}$ is the coefficient vector associated with the jth feature in the tth data set with $\boldsymbol{\beta}_{jt} \sim \mathcal{N}(\boldsymbol{\beta}_{0t}, \mathbf{\Sigma}_{0t})$, $\mathbf{\Gamma}_{jt} = \text{diag}(1, \gamma_{jt}, \ldots, \gamma_{jt})$ with $\gamma_{jt} \sim \text{Be}(q_t)$, $\mathbf{x}_i = (1, \mathbf{z}_i)$, where $\mathbf{z}_i$ is the ith latent variable and $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$ where k is the number of latent dimensions, and $\boldsymbol{\varepsilon}_{ijt} \sim \mathcal{N}(0, \sigma^2_{jt})$ is a random error term with $\sigma^2_{jt} \sim \text{IG}(\nu_0/2, \nu_0\sigma^2_0/2)$. Gibbs sampling is used to sample from the posterior distributions of σ^2_{jt} and $\boldsymbol{\beta}_{jt}$, where Metropolis Hastings is used to infer γ_{jt} and $\mathbf{z}_i$.

If y_{ijt} is a binary variable, indicating , for example, the presence or absence of a mutation, it is modelled by logistic regression:

$$\log \frac{p(y_{ijt} = 1|\mathbf{z}_i)}{1 - p(y_{ijt} = 1|\mathbf{z}_i)} = \mathbf{x}_i \mathbf{\Gamma}_{jt} \boldsymbol{\beta}_{jt}, \tag{1.78}$$

where $\mathbf{z}_i, \mathbf{\Gamma}_{jt}$ and $\boldsymbol{\beta}_{jt}$ have the same interpretation and priors as in the continuous case. If y_{ijt} is a count variable, then it is modelled by Poisson regression

$$\log(\lambda(y_{ijt}|\mathbf{z}_i)) = \mathbf{x}_i \mathbf{\Gamma}_{jt} \boldsymbol{\beta}_{jt}, \tag{1.79}$$

where $\mathbf{z}_i, \mathbf{\Gamma}_{jt}$ and $\boldsymbol{\beta}_{jt}$ have the same interpretation and priors as in the continuous and binary cases. As the posterior distributions of $\mathbf{\Gamma}_{jt}$ and $\boldsymbol{\beta}_{jt}$ have no closed form expressions in the binary and count data case, they are inferred via Metropolis Hastings.

Savage et al. [2010] propose a data fusion model which infers transcriptional modules by integrating gene expression and transcription factor binding data. The model extends the hierarchical Dirichlet process model of Teh et al. [2005] to allow the data fusion on gene-by-gene basis. It introduces an indicator variable for each gene to determine whether the gene should join a cluster based on both data sources or if it should be clustered separately for each source.

Patient-specific data fusion model (PSDF) proposed by Yuan et al. [2011] is an extension of the model proposed by Savage et al. [2010]. It similarly uses a two-level hierarchy of Dirichlet processes, where each patient has a binary state r_i that defines whether their data are concordant across the data types and are either fused (allocated to the same cluster across all datasets, $r_i = 1$) or unfused ($r_i = 0$). PSDF requires discretisation of the data, which is then modelled by a naive Bayes data model. It incorporates feature selection by using another indicator variable I_a, which denotes whether a feature is on or off. Similarly to Savage et al. [2010], the inference is performed by MCMC methods, which also require considerable computational costs.

Another Bayesian method for unsupervised integrative clustering of multiple datasets is **Multiple Dataset Integration** (MDI) [Kirk et al., 2012]. Each dataset is modelled as a finite approximation to a Dirichlet process mixture model [Ishwaran and Zarepour, 2002], which has the following form:

$$p(x) = \sum_{c=1}^{N} \pi_c f(x|\theta_c). \tag{1.80}$$

In (1.80) $f(x)$ is the probability density model for the data, π_c are the mixing proportions, f is a parametric density and θ_c are the parameters of component c. To aid inference, latent component allocation variables $c_j \in \{1, \dots, N\}$ are introduced, with c_i being the component responsible for $\mathbf{x}_i$. The full model specification is as follows:

$$\begin{aligned}
\mathbf{x}_i|c_i, \theta &\sim F(\theta_{c_i}) \\
c_i|\pi &\sim Mult(\pi_1, \dots, \pi_N) \\
\pi_1, \dots, \pi_N &\sim Dir(\alpha/N, \dots, \alpha/N) \\
\theta_c &\sim G^{(0)},
\end{aligned}$$

where F is the distribution corresponding to density f, $\pi = (\pi_1, \dots, \pi_N)$ is the N mixture proportions, α is a concentration parameter, and $G^{(0)}$ is the prior for the component parameters. MDI can be applied to any type of data; for example, Kirk et al. [2012] use Gaussian process models for gene expression time series data and multinomial model for categorical data. The inference of the model parameters is performed using a Gibbs sampling scheme.

MDI links the models for the datasets via the conditional prior on the component allocation variables

$$p(c_{i1}, c_{i2}, \dots, c_{iK}|\phi) \approx \prod_{k=1}^{K} \pi_{c_{ik}k} \prod_{k=1}^{K-1} \prod_{l=k+1}^{K} (1 + \phi_{kl}\mathbb{I}(c_{ik} = c_{il})), \tag{1.81}$$

where $\mathbb{I}$ is the indicator function, $\phi_{kl} \in \mathbf{R}_{\geq 0}$ is a parameter that controls the strength of association between datasets k and l and ϕ is the collection of all ϕ_{kl}s. Note that c_{ik} is the component allocation variable associated with gene i in model k, and that $\pi_{c_{ik}k}$ is the mixture proportion associated with component c_{ik} in model k. The larger ϕ_{kl}, the more likely it is that c_{ik} and c_{il} will be the same, and hence the greater similarity between the clustering structure of dataset k and dataset l. As MDI only looks at the pairwise relations between datasets, this can limit the interpretability of the results. Although the authors use the model to find groups of genes that cluster together in gene expression and ChIP-chip data, it can similarly be applied to find groups of patients that cluster together in different genomic data sources [Savage et al., 2013].

Other data integration methods similarly use Dirichlet process mixture model since it offers scalable inference and learns the number of clusters from the data. One example is **Bayesian Consensus Clustering** (BCC) proposed by Lock and Dunson [2013]. This model extends a Dirichlet process mixture model to accommodate data from M sources $\mathbf{X}_1, \dots, \mathbf{X}_M$. Each dataset is available for a common set of N objects and requires a probability model $f_m(X_n|\theta_m)$ parameterised by θ_m. There is a separate clustering of the objects for each data type, but they adhere loosely to an overall clustering based on a parameter α_m.

The source specific clusterings $\mathcal{L}_m$ are connected to the overall clustering $\mathcal{C}$ via a dependence function ν:

$$P(L_{mn} = k|C_n) = \nu(k, C_n, \alpha_m), \tag{1.82}$$

where $L_{mn} \in \{1, \dots, K\}$ is the component corresponding to object n in dataset m and $C_n \in \{1, \dots, K\}$ is the overall mixture component for object n. Since the datasets are independent of $\mathcal{C}$ conditional on the source-specific clusterings, $\mathcal{C}$ serves to unify the source-specific clusterings.

BCC differs from traditional consensus clustering, often used to combine multiple realisations from the same algorithm, in that the clusterings are modelled in a statistical way that allows for uncertainty in all parameters, and both source-specific and consensus clusterings are estimated simultaneously and the strength of association is learned from the data. However, as the authors point out, the model tends to select a large number of clusters even when the Dirichlet process concentration parameter is very small, which is unrealistic from biological point of view. Thus they consider an alternative heuristic measure that selects the number of clusters which give maximum adherence to an overall clustering. The use of this measure coupled with a Gibbs sampling inference scheme, leads to more computationally demanding and less straightforward inference.

Some of the methods described above consider only the shared structure between the molecular datasets. However, the individual data structure can be informative. To account for both shared and individual structure, Lock et al. [2013] develop the **Joint and Individual Variation Explained** (JIVE) model. It uses the biological relation between different types of molecular data, which motivates the idea of shared patterns between these types of data, referred to as the joint structure. JIVE decomposes a dataset into a sum of three terms: a low-rank approximation capturing joint structure between data types, low-rank approximations capturing structure individual to each data type, and residual noise.

For example, if we want to integrate data from multiple data sources $\mathbf{X}_1, \dots, \mathbf{X}_k$, where $k \geq 2$, then each of the matrices is decomposed as follows:

$$\begin{aligned}\mathbf{X}_1 &= \mathbf{J}_1 + \mathbf{A}_1 + \boldsymbol{\varepsilon}_1\\ &\vdots\\ \mathbf{X}_k &= \mathbf{J}_k + \mathbf{A}_k + \boldsymbol{\varepsilon}_k,\end{aligned}$$

where $\mathbf{A}_i$ is the matrix representing the individual structure of $\mathbf{X}_i$, $\mathbf{J}_i$ is the submatrix of the joint structure matrix associated with $\mathbf{X}_i$, and $\boldsymbol{\varepsilon}_i$ are $p_i \times n$ error matrices of independent entries. The joint structure matrix $\mathbf{J}$ has all the $\mathbf{J}_i$ matrices stacked together and the rows of the joint and individual patterns are orthogonal: $\mathbf{J}\mathbf{A}_i^\intercal = 0$. This implies that the sample patterns responsible for joint structure between data types are unrelated to sample patterns responsible for individual structure.

The joint and individual structures are estimated by minimising the sum of squared error. $\mathbf{R}$ is the $p \times n$ matrix of residuals after accounting for joint and individual structure. The matrices $\mathbf{J}, \mathbf{A}_1, \dots, \mathbf{A}_K$ are estimated by minimising $||\mathbf{R}||^2$ under the given ranks. This is accomplished by iteratively estimating the joint and individual structure:

- Given $\mathbf{J}$, find $\mathbf{A}_1, \dots, \mathbf{A}_k$ to minimise $||\mathbf{R}||$.

- Given $\mathbf{A}_1, \ldots, \mathbf{A}_k$ find $\mathbf{J}$ to minimise $||\mathbf{R}||$.
- Repeat until convergence.

We finish our review of clustering algorithms for data integration with two of the most recently developed algorithms: **Clusternomics** [Gabasova et al., 2017] and a **multi-omics factor analysis** model, created by Argelaguet et al. [2018]. Clusternomics is a probabilistic clustering method which models clusters on the level of individual datasets using hierarchical Dirichlet process mixture models [Teh et al., 2005], whilst also extracting global structure that arises from the local cluster assignments. The method makes the assumptions that the clustering structure in one of the datasets should influence the clustering in other, and that different degrees of dependence should be allowed between clusters across datasets. However, Clusternomics requires setting up the number of global clusters to a specific value, and derives the global clusters as a combination of local clusters, which often results in prohibitively many combinations to compute in timely fashion. Although the authors provide a way of reducing the number of combinations required to define the global clusters, the model inference which uses Gibbs sampling is still slow.

The multi-omics factor analysis (MOFA) model proposed by Argelaguet et al. [2018] is a generalisation of principal component analysis to multi-omics data. The method builds on group factor analysis [Virtanen et al., 2012; Khan et al., 2014; Klami et al., 2015; Bunte et al., 2016; Leppäaho et al., 2017] and models each dataset $\mathbf{Y}^m$ for $m = 1, \ldots, M$ as follows

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{m\intercal} + \boldsymbol{\varepsilon}^m, \tag{1.83}$$

where $\mathbf{Z}$ are the latent factors, $\mathbf{W}^m$ is the loadings matrix associated with the mth dataset and $\boldsymbol{\varepsilon}^m$ is the residual error associated with the mth dataset. It can handle missing data and model different types of data, and performs inference using variational approximations. Although the inference scheme is faster than in other methods, it does not provide a full posterior of the model parameters.

We summarise the most important features of the integrative clustering methods in Table 1.1.

Model	Data types	Estimate number of clusters	Inference	Reference
iCluster	continuous	proportion of deviance	Expectation Maximisation	Shen et al. [2009]
iClusterPlus	continuous, binary, categorical, count	deviance ratio	modified Monte Carlo Newton-Raphson	Mo et al. [2013]
iClusterBayes	continuous, binary, categorical, count	deviance ratio	random walk MH	Mo et al. [2017]
MDI	continuous, binary, categorical, count	DPMM	Gibbs Sampling	Kirk et al. [2012]
PSDF	continuous, binary, categorical, count	DPMM	Gibbs Sampling	Yuan et al. [2011]
BCC	continuous, binary, categorical, count	Max adherence to an overall clustering	Gibbs Sampling	Lock and Dunson [2013]
JIVE	continuous	Using scores	Optimisation	Lock et al. [2013]
Clusternomics	continuous, binary, categorical, count	Using DIC	Variational inference/Gibbs sampling	Gabasova et al. [2017]
MOFA	continuous, binary, count	Using scores	Variational inference	Argelaguet et al. [2018]

Table 1.1: Important features of the data integration clustering methods used in the analyses summarised in this chapter.

1.3.6 Variable selection

When we perform variable selection for a clustering task, we ideally want to keep only the relevant features, which contain essential information to perform the task, and discard any uninformative variables that offer no discriminative power. There are two main types of methods that aim to achieve that: **filter** and **wrapper** [Fop et al., 2018].

The filter techniques assess the relevance of features by looking only at the statistics of the data such as variance, correlation or mutual information. These techniques are often used to preselect variables in the analyses of datasets with too many variables. The inferred classification is then used to assess the quality of the selected variables. Since the variable selection and model estimation are decoupled,this approach can miss important and relevant information.

In contrast, the wrapper methods perform learning and variable selection at the same time. They have become increasingly popular because they can provide a better representation of the data generating process and lead to a more accurate classification Dy and Brodley [2004]; Law et al. [2004]. The wrapper methods are split into 3 categories based on the statistical approach used to select the variables. The **Bayesian** approach assumes that there is a latent variable indicating whether an observed variable is informative or not. The **model selection** approach reformulates the task of variable selection as a model selection problem. The features are selected by comparing the models using a predefined criterion. The third approach is performed using **a penalisation term** which shrinks the estimates for the model parameters towards an overall common value, which implies the irrelevance of the corresponding features.

1.3.7 Comparing two clustering partitions

In order to compare the outputs from the models summarised above and developed in this thesis, we need a meaningful ‘measure’, which should ideally tell us how similar two partitions are, how close to the ground truth a certain partition is and whether the algorithm is susceptible to small perturbations and to the order of the data. We provide below a short summary of the most commonly used measures, which serves as a justification for our choice of measure.

Consider a finite dataset X with cardinality $|X| = N$, and two clustering partitions of the dataset $C = \{C_1, \ldots, C_k\}$ and $C' = \{C'_1, \ldots, C'_l\}$, with $C_1, \ldots, C_k$ being the

k clusters forming partition C and $C'_1, \ldots, C'_l$ being the l clusters forming partition C'. We can measure how similar two partitions are based on **counting pairs**, on **mutual information** and on **set overlaps**. We will focus on the first two types of measure as the measures based on set overlaps are difficult to use because of their asymmetry [Wagner and Wagner, 2007].

The measures based on counting pairs count the number of pairs of objects classified in the same way in both clusterings.

One such measure is the **Chi-squared coefficient** [Mirkin, 2001], which is defined as follows:

$$\chi(C, C') = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(m_{ij} - E_{ij})^2}{E_{ij}}, \tag{1.84}$$

where $m_{ij} = |C_i \bigcap C_j|$ and $E_{ij} = \frac{|C_i||C'_j|}{n}$. Although this measure is easy to use, it requires strong assumptions such as the independence of the two clusterings.

The **Rand Index** [Rand, 1971] was motivated by classification problems where the ground truth is known. It counts the number of correctly classified pairs of elements and is defined as follows:

$$RI(C, C') = \frac{2(n_{11} + n_{00})}{N(N-1)}, \tag{1.85}$$

where n_{11} is the number of pairs in the same cluster under C and C' and n_{00} is the number of pairs in different clusters under C and C'. The value of the Rand index ranges from 0 to 1 but is highly dependent on the number of clusters as shown by Morey and Agresti [1984].

The expected value of the Rand Index of two random partitions does not take a constant value. To address this issue, Hubert and Arabie [1985] propose an adjustment which assumes a generalised hypergeometric distribution as the null hypothesis. This means that two clusterings are drawn randomly with a fixed number of clusters and a fixed number of elements in each cluster. The proposed **adjusted Rand Index** measure [Kuncheva and Hadjitodorov, 2004] is the normalised difference of the Rand index and its expected value under the null hypothesis:

$$ARI(C, C') = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} \binom{m_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}, \tag{1.86}$$

where $t_1 = \sum_{i=1}^{k} \binom{|C_i|}{2}$, $t_2 = \sum_{j=1}^{l} \binom{|C'_i|}{2}$ and $t_3 = \frac{2t_1t_2}{N(N-1)}$. This index has expected value zero for independent clusterings and maximum value of 1 for identical clusterings. We should take into account the strong assumptions about the model of randomness this index makes when we use it.

Another widely used measure is the **Jaccard index**, which is very similar to the Rand Index but it disregards the pairs of elements that are in different clusters for both clusterings. It is defined as follows:

$$\mathcal{J}(C, C') = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}, \tag{1.87}$$

where n_{10} is the number of pairs that are in the same cluster under C but in different ones under C', and n_{01} is the number of pairs that are in different clusters under C but the same under C'.

Another popular approach to comparing two partitions is based on **mutual information**.

When applied to clustering, the entropy associated with clustering C [Meilă, 2007] is

$$\mathcal{H}(C) = -\sum_{i=1}^{k} P(i) \log_2 P(i), \tag{1.88}$$

where $P(i) = \frac{|C_i|}{n}$ is the probability that this element is in cluster C_i. This implies that the entropy of a clustering is a measure for uncertainty about the cluster of a randomly picked element. This notion of entropy can be extended to that of mutual information, which describes how much we can reduce the uncertainty about the cluster of a random element when knowing its cluster in another clustering of the same set of elements. It is defined as follows

$$\mathcal{I}(C, C') = \sum_{i=1}^{k} \sum_{j=1}^{l} P(i, j) \log_2 \frac{P(i, j)}{P(i)P(j)}, \tag{1.89}$$

where $P(i, j)$ is the probability that an element belongs to cluster i in C and to cluster j in C'. The mutual information is a metric on the space of all clusterings but it is not bounded by a constant value which makes it difficult to interpret.

Strehl and Ghosh [2002] try to address this issue by developing a **normalised mutual information** metric which is defined below

$$NMI_1(C, C') = \frac{\mathcal{I}(C, C')}{\sqrt{\mathcal{H}(C)\mathcal{H}(C')}} \tag{1.90}$$

and is bounded between 0 and 1.

Another proposal for **normalised mutual information** is made by Ana and Jain [2003]. Their proposed metric takes the following form

$$NMI_2(C, C') = \frac{2\mathcal{I}(C, C')}{\mathcal{H}(C) + \mathcal{H}(C')}, \tag{1.91}$$

which is similarly bounded between 0 and 1. However, NMI has been shown to favour partitions with more clusters and to overestimate the number of true clusters [Amelio and Pizzuti, 2015].

Based on the advantages and disadvantages of the measures outlined above we have picked to work with the Adjusted Rand Index (1.86) when the ground truth is available.

1.4 Latent variable models

A **latent variable model** aims to express the distribution of N d-dimensional observations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ with a set of p-dimensional latent (unobserved) variables $\mathbf{Z}$, where $p < d$. We assume that we can factorise the joint distribution $p(\mathbf{X}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\mathbf{Z})p(\mathbf{Z})$. Under a latent variable model, the conditional distribution $p(\mathbf{X}|\mathbf{Z})$ can be expressed in terms of a mapping $\mathbf{W}$ from the latent variables to the observations

$$\mathbf{X} = f(\mathbf{Z}; \mathbf{W}) + \boldsymbol{\varepsilon} \tag{1.92}$$

where f is the mapping with parameter $\mathbf{W}$ and $\boldsymbol{\varepsilon}$ is noise, independent from the latent variables. The definition of the latent variable model is completed by specifying a prior on the latent variables $\mathbf{Z}$, a prior on the noise $p(\boldsymbol{\varepsilon})$ and a mapping $f(\mathbf{Z}; \mathbf{W})$ [Bishop, 1998].

1.4.1 Latent variable models for continuous data

We start by describing **principal component analysis** (PCA) [Hotelling, 1933; Jolliffe, 1986], which although not a probabilistic model, can be extended to a probabilistic latent variable model. It is a well-established technique for dimensionality reduction that is widely used for applications such as image processing, feature extraction, exploratory data analysis, visualization, and pattern recognition.

There are two commonly used definitions of PCA which give rise to the same model.

The most common derivation of PCA is in terms of orthogonal projection of the data onto a lower-dimensional linear space, also called principal subspace, so that the retained variance under the projection is maximised. It can be shown that for a set of observed d-dimensional data $\mathbf{x}_n, n \in \{1, \ldots, N\}$, the p ($p < d$) *principal axes* $\mathbf{w}_j$ are those orthonormal axes which satisfy the maximal variance condition. The p *principal components* of the observed vectors $\mathbf{x}_n$ are given by the vectors $\mathbf{z}_n = \mathbf{W}^\intercal(\mathbf{x}_n - \bar{\mathbf{x}})$ where $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p\}$.

An alternative derivation of PCA is based on the minimisation of error projection. Bishop [2006] shows that of all orthogonal linear projections $\mathbf{z}_n$, the principal component projections minimise the reconstruction error $\sum_n |\mathbf{x}_n - \hat{\mathbf{x}}_n|^2$, and that the optimal linear reconstruction of $\mathbf{x}_n$ is given by $\hat{\mathbf{x}}_n = \mathbf{W}\mathbf{z}_n + \bar{\mathbf{x}}$.

A probabilistic extension of PCA, known as **probabilistic PCA** (PPCA) has been proposed by both Roweis [1998] and Tipping and Bishop [1999] independently. PPCA has certain practical advantages, including the existence of a computationally efficient Expectation Maximisation algorithm, a principled approach for dealing with missing data, and a likelihood function which allows for the comparison with other probabilistic models.

We can formulate **probabilistic principal component analysis** by introducing first the latent variables $\mathbf{Z}$ with a Gaussian prior

$$p(\mathbf{Z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1.93}$$

which correspond to the principal component subspace. Next, we define a Gaussian conditional distribution of the observations $\mathbf{X}$ conditioned on $\mathbf{Z}$

$$p(\mathbf{X}|\mathbf{Z}) \sim \mathcal{N}(\mathbf{X}|\mathbf{W}\mathbf{Z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}), \tag{1.94}$$

where $\mathbf{W}$ is the loadings matrix and the mapping function in this case, whose columns span a linear subspace within the data space that corresponds to the principal subspace, $\boldsymbol{\mu}$ is the mean vector, and σ^2 governs the variance.

If we want to determine the likelihood estimates for the model parameters $\mathbf{W}, \sigma^2$, we need to derive first the marginal distribution of the observations $p(\mathbf{X})$. We have that the marginal distribution is

$$p(\mathbf{X}) = \int p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})\mathrm{d}\mathbf{Z}, \tag{1.95}$$

which by using (1.93) and (1.94) means that the marginal distribution $p(\mathbf{X})$ is also

Gaussian and is given by

$$p(\mathbf{X}) = \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \mathbf{C}), \tag{1.96}$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^\intercal + \sigma^2\mathbf{I}$ is the covariance matrix.

We can now derive maximum likelihood estimates for the model parameters $\mathbf{W}$, σ and $\boldsymbol{\mu}$ by differentiating the likelihood function with respect to each of the model parameters:

$$\log p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{i=1}^{N} \log p(\mathbf{x}_i|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) \tag{1.97}$$

$$= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})^\intercal \mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \tag{1.98}$$

where D is the dimensionality of the data, and N is the number of observations.

Setting the derivative of (1.97) with respect to $\boldsymbol{\mu}$ equal to 0 gives that $\boldsymbol{\mu} = \bar{\mathbf{X}}$, which is the data mean. Using this, we can rewrite (1.98) as follows:

$$\log p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{N}{2}\{D\log(2\pi) + \log|\mathbf{C}| + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})\}, \tag{1.99}$$

where S is the data covariance matrix.

Tipping and Bishop [1999] show that the maximum likelihood estimate of $\mathbf{W}$ can be written as

$$\mathbf{W}_{ML} = \mathbf{U}_M(\mathbf{L}_M - \sigma^2\mathbf{I})^{\frac{1}{2}}\mathbf{R} \tag{1.100}$$

where $\mathbf{U}_M$ is the $D \times M$ matrix whose columns are the M eigenvectors of the data covariance matrix $\mathbf{S}$ with the largest eigenvalues λ_i, $\mathbf{L}$ is $M \times M$ diagonal matrix with entries equal to the eigenvalues λ_i, and $\mathbf{R}$ is an arbitrary rotation $M \times M$ orthogonal matrix, usually set to be the identity matrix.

The corresponding maximum likelihood solution for σ^2 is given by

$$\sigma^2_{ML} = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i, \tag{1.101}$$

which is the average variance associated with the discarded dimensions.

We can also derive an expression for the posterior distribution of the latent variables in order to learn them as well. Using Bayes theorem and the results for conditional distribution of a Gaussian variable in [Bishop, 2006], we have that:

$$p(\mathbf{Z}|\mathbf{X}) = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^{\intercal}(\mathbf{X} - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}) \tag{1.102}$$

where $\mathbf{M} = \mathbf{W}^{\intercal}\mathbf{W} + \sigma^2\mathbf{I}$.

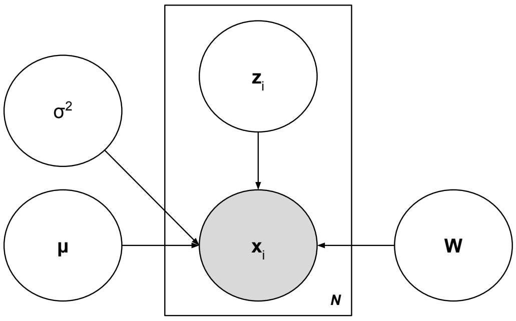


Figure 1.9: Graphical model of PPCA - the shaded node $\mathbf{x}_i$ represents the observations, $\mathbf{z}_i$ - the latent variables, $\mathbf{W}$ - the loadings matrix, $\boldsymbol{\mu}$ - the mean vector and σ^2 - the error term.

The number of latent dimensions can be found by using crossvalidation since PPCA has a well-defined likelihood function, and selecting the model corresponding to the largest log likelihood on a validation dataset. However, this can often be computationally costly.

Alternative approaches are to use model selection techniques such as Bayesian information criterion, or **Bayesian principal component analysis** (BPCA) [Bishop, 1999], which determines the latent dimensionality in an efficient way. BPCA uses the idea of evidence approximation, which Bishop [1999] points out is a suitable choice in the case of many data points and tightly peaked posterior. The only way in which BPCA differs from PPCA is in the prior with hyperparameter α over the columns of the mapping $\mathbf{W}$ (see the graphical model on Figure 1.10), that allows extra dimensions in $\mathbf{W}$ to be excluded.

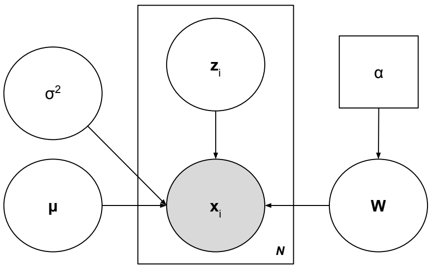


Figure 1.10: Graphical model of BPCA - the shaded node $\mathbf{x}_i$ represents the observations, $\mathbf{z}_i$ - the latent variables, $\mathbf{W}$ - the loadings matrix, $\boldsymbol{\mu}$ - the mean vector and σ^2 - the error term.

The final linear latent variable model we will consider is **factor analysis** which is closely related to PPCA. Its definition differs from that of PPCA in the conditional distribution of the observed variables $\mathbf{X}$ given the latent variables $\mathbf{Z}$

$$p(\mathbf{X}|\mathbf{Z}) = \mathcal{N}(\mathbf{X}|\mathbf{W}\mathbf{Z} + \boldsymbol{\mu}, \boldsymbol{\Psi}) \tag{1.103}$$

where $\boldsymbol{\Psi}$ is a diagonal matrix, and the columns of $\mathbf{W}$ capture the correlations between the observations (Figure 1.11). If $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$, then we recover PPCA. However, unlike PPCA, there is no closed-form maximum likelihood solution for $\mathbf{W}$. We can instead use an EM algorithm proposed by Rubin and Thayer [1982].

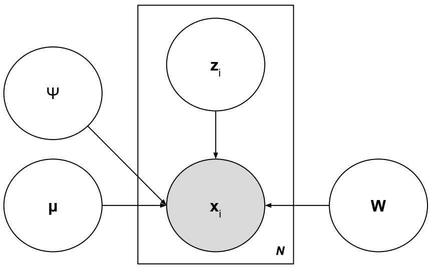


Figure 1.11: Graphical model of factor analysis - the shaded node $\mathbf{x}_i$ represents the observations, $\mathbf{z}_i$ - the latent variables, $\mathbf{W}$ - the loadings matrix, $\boldsymbol{\mu}$ - the mean vector and $\boldsymbol{\Psi}$ - the error matrix.

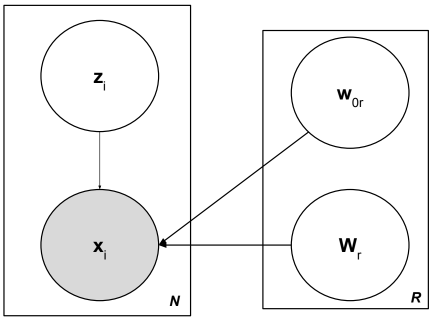


Figure 1.12: A graphical model representing the categorical PCA model.

1.4.2 Latent variable models for discrete data

Latent variable models can be used with discrete data as well. Such models are often used in text and image analysis, information retrieval, bioinformatics and social sciences.

Some examples include **categorical PCA** [Murphy, 2012], **multinomial PCA** [Buntine and Jakulin, 2004; Buntine, 2002], **Latent Dirichlet allocation** [Blei et al., 2003].

In categorical PCA (see Section 12.4 [Murphy, 2012]), the observations have the form $x_{ij} \in \{1, \dots, R_j\}$, where j is the number of features and R_j is the number of categories that the jth feature can take (Figure 1.12). Each x_{ij} is generated from a latent variable $\mathbf{z}_i \in \mathbb{R}^L$, which is passed through a softmax function:

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{1.104}$$

$$p(\mathbf{x}_i|\mathbf{z}_i, \theta) = \prod_{r=1}^{R} \mathrm{Cat}(x_{ir}|\mathcal{S}(\mathbf{W}_r^\intercal \mathbf{z}_i + \mathbf{w}_{0r})), \tag{1.105}$$

where $\mathbf{W}_r \in \mathbb{R}^{L \times M}$ is the factor loading matrix for the r^{th} feature, $\mathbf{w}_{0r}$ is the offset term for the r^{th} feature and R_j is the number of different categories. The softmax function transforms a K-dimensional real-valued vector η into a K-dimensional vector $\mathcal{S}(\eta)$ of real values, where each entry is in $(0, 1)$ and all entries add up to 1:

$$\mathcal{S}(\eta)_j = \frac{\exp(\eta_j)}{\sum_{k=1}^{K} \exp(\eta_k)}. \tag{1.106}$$

The corresponding distribution of the observations does not have closed-form solution:

$$p(\mathbf{x}_{i,1:R}) = \int \left[\prod_{r=1}^{R} p(x_{ir}|\mathbf{z}_i, \mathbf{W}_., \mathbf{w}_{0.}) \right] \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\mathbf{z}_i. \tag{1.107}$$

However, we can fit the model using a modified version of EM [Khan et al., 2010], for example. In the E-step, a Gaussian approximation to the posterior distribution $p(\mathbf{z}_i|\mathbf{x}_i, \{\mathbf{W}_r\}_{r=1}^{R}, \{\mathbf{w}_{0r}\}_{r=1}^{R})$ is inferred, and in the M-step, the model parameters $(\{\mathbf{W}_r\}_{r=1}^{R}, \{\mathbf{w}_{0r}\}_{r=1}^{R})$ are maximised.

We can model count data by using a Poisson model

$$p(\mathbf{x}_i|\mathbf{z}_i) = \prod_{v=1}^{V} \mathrm{Poi}(x_{iv}|\exp(\mathbf{w}_{v,:}^{\intercal}\mathbf{z}_i)). \tag{1.108}$$

We can fit the model in a similar manner to Categorical PCA.

We can model count vectors whose total sum is known with multinomial PCA [Buntine, 2002; Buntine and Jakulin, 2004] (Figure 1.13). This is similar to the multinomial model above but instead of using the softmax function, we use a matrix $\mathbf{B}$ with entries $0 \leq b_{v,k} \leq 1$ and $\sum_v b_{v,k} = 1$, and a vector $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha)$ such that

$$p(\mathbf{x}_i|L_i, \boldsymbol{\pi}) = \mathrm{Mult}(\mathbf{x}_i|L_i, \mathbf{B}\pi_i). \tag{1.109}$$

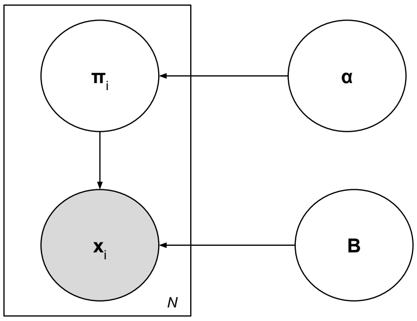


Figure 1.13: A graphical model representing the multinomial PCA model.

If we have a variable length sequence of known length, we can use instead

$$p(\mathbf{x}_{i,1:L}|\boldsymbol{\pi}) = \prod_{i=1}^{L} \mathrm{Cat}(x_{il}|\mathbf{B}\pi_i), \tag{1.110}$$

which corresponds to the latent Dirichlet allocation model [Blei et al., 2003].

1.5 Thesis outline

The rest of this thesis proceeds as follows:

Chapter 2 introduces a novel Bayesian nonparametric method for clustering mixed data, which is based on Dirichlet process mixtures, and highlights its advantages over traditional clustering approaches. The model, which we call BayesCluster, is implemented in Chapter 3, where we demonstrate its useful properties with synthetic and real datasets and compare its performance with other commonly used methods.

In Chapter 4, we present several extensions to BayesCluster, based on the ideas of non-local priors, split-merge and cluster-size priors, and illustrate how they lead to stronger model parsimony and the identification of more interpretable clusters.

In Chapter 5, we extend BayesCluster to a combined data integration and clustering model. The core idea being the integrative framework we adopt is that the model learns a common set of latent features jointly from multiple heterogenous data types. We consider the application of the data integration model to identify cancer subtypes indicative of overall survival and present the results from four studies, involving genomic datasets from different projects part of The Cancer Genome Atlas.

Chapter 6 explores the impact of different clinical factors on the short- and long-term survival of pancreatic cancer patients following pancreatic cancer resection. We present the results from a study using the Hospital Episode Statistics database, which aims to assess the impact of centralisation of surgeries on the patients' survival.

Finally, we summarise the main contributions of this thesis in Chapter 7, and outline directions for future work.

Chapter 2

Mixed Data Clustering: Theory

The datasets we work with in this thesis require the development of clustering algorithms that can model different types of data - continuous, discrete, and binary. As we want to combine the information about the same patient cohort from different high-dimensional datasets, we also need a model that can provide efficient dimensionality reduction and hence, a more concise description of the data. Many latent variable models provide a framework for the accomplishment of these tasks and we make use of them in the development of a Bayesian framework to clustering mixed data. We outline the theory behind the novel clustering algorithm, called **BayesCluster**, which can be applied to different types of data. BayesCluster combines Bayesian nonparametric clustering with latent variable representations of the data. We focus on clustering a single, potentially mixed, dataset in this chapter - we extend the model to multiple datasets in Chapter 5.

2.1 BayesCluster - a model for mixed data clustering

We introduce BayesCluster, a method for clustering mixed data. It offers the advantages of both latent variable models, which offer a lower-dimensional representation of the data, and Bayesian nonparametric models, which require few assumptions about the data and are relatively insensitive to outliers since their approach is to fit a single model that can adapt its complexity to the data instead of specifying the number of components in advance [Hollander et al., 2013].

2.1.1 Model specification for continuous data

We assume that the dataset $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ that we want to model is normalised, i.e. has mean zero and unit variance, and that $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independent and identically multivariate distributed Gaussian variables.

We model each D-dimensional continuous observation $\mathbf{x}_i$ by a Gaussian likelihood with unknown mean and variance

$$p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}, \boldsymbol{\varepsilon}) = \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}), \tag{2.1}$$

where $\mathbf{z}_i$ is the corresponding P-dimensional latent variable, $\mathbf{W}$ is the $D \times P$ loadings matrix, which is the projection matrix that maps the data to a lower-dimensional space, and $\boldsymbol{\varepsilon}$ is the error term, representing the residual variance. We assume that $\boldsymbol{\varepsilon}$ is $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ Gaussian noise.

We place a Normal prior on the means of the clusters of latent variables $\boldsymbol{\mu}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and assume that the latent variables $\mathbf{z}_i$ are independent Normally distributed random variables with $p(\mathbf{z}_i|c_i = k, \boldsymbol{\mu}_k) = \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I})$, where c_i is the cluster indicator for the ith latent variable.

The latent variables $\mathbf{Z}$ are modelled using an infinite mixture model [Rasmussen, 1999]:

$$p(\mathbf{Z}) = \sum_{k=1}^{\infty} \pi_k f(\mathbf{Z}|\boldsymbol{\theta}_k), \tag{2.2}$$

where we place a Dirichlet process prior $\mathrm{Dir}(\pi|\alpha)$ on the mixing propotions π and use a Normal distribution $\mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I})$ to model the latent variables. We assume that the clusters of latent variables have the same covariance to simplify the model and infer only the cluster means.

Using this approach, we cluster the lower-dimensional latent variables $\mathbf{Z}$ rather high-dimensional observations $\mathbf{X}$, and we do not need to specify the number of clusters as we learn this from the data. We can summarise the probabilistic model and the assumptions we make as follows:

$$\begin{aligned} p(\pi|\alpha) =& \mathrm{Dir}(\alpha) \\ p(c_i|\boldsymbol{\pi}) =& \mathrm{Mult}(\boldsymbol{\pi}) \\ p(\mathbf{z}_i|c_i = k, \boldsymbol{\mu}_k) =& \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \\ p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}, \boldsymbol{\varepsilon}) =& \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}) \\ p(\boldsymbol{\mu}_k) =& \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I}) \\ p(\mathbf{W}) =& \prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_d|\mathbf{0}, \mathbf{I}) \\ p(\boldsymbol{\varepsilon}) =& \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0}, \sigma^2\mathbf{I}). \end{aligned}$$

The graphical model below (Figure 2.1) presents the generative model and dependencies between the parameters and the observations, and we have listed the parameters and their interpretation in the Notation table 2.1 below.

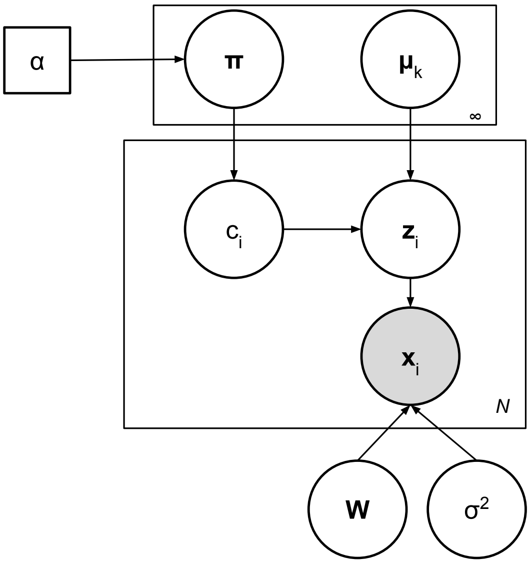


Figure 2.1: A graphical model representing the independence assumptions for the BayesCluster model applied to continuous data.

Inference

A standard approach to performing inference in Dirichlet process mixture models involves using MCMC methods. For example, Neal [2000] outlines different inference

Parameter	Domain	Interpretation
N	$\mathbb{Z}^+$	number of observations
D	$\mathbb{Z}^+$	dimension of data
P	$\mathbb{Z}^+$	number of principal components
$\mathbf{x}_i$	$\mathbb{R}^D$	the i^{th} observation
$\mathbf{x}_i^C$	$\mathbb{R}^{D-R}$	the continuous part of the i^{th} mixed observation
$\mathbf{x}_i^D$	$\mathbb{R}^R$	the discrete part of the i^{th} mixed observation
N_k	$\mathbb{Z}^+$	number of observations in cluster k
$N_{k,-i}$	$\mathbb{Z}^+$	number of observations in cluster k excluding the ith observation
K	$\mathbb{Z}^+$	number of occupied clusters
c_i	$\{1, \ldots, K\}$	i^{th} cluster indicator variable
C	$\{c_1, \ldots, c_N\}$	the collection of indicator variables
π_k	$[0, 1]$	mixing proportion for the k^{th} cluster
$\mathbf{z}_j$	$\mathbb{R}^P$	j^{th} latent factor
$\mathbf{Z}$	$\mathbb{R}^{N \times P}$	the collection of all latent factors
$\mathbf{W}$	$\mathbb{R}^{D \times P}$	loadings matrix (continuous variables)
$\mathbf{W}_r^D$	$\mathbb{R}^{R \times P}$	r^{th} loadings matrix (discrete variables)
$\mathbf{W}^D$	$\{\mathbf{W}_1^D, \ldots \mathbf{W}_r^D\}$	the collection of all loadings matrices (discrete variables)
$\mathbf{w}_{0r}$	$\mathbb{R}^R$	r^{th} offset term
$\mathbf{w}_0$	$\mathbb{R}^R$	the collection of all offset terms
$\boldsymbol{\varepsilon}$	$\mathbb{R}^D$	residual noise (continuous variables)
$\boldsymbol{\theta}_k$	$\mathbb{R}^P, \mathbb{R}^{P \times P}$	the model parameters for cluster k $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
$\boldsymbol{\mu}_k$	$\mathbb{R}^P$	mean of the kth cluster
α	$\mathbb{R}^+$	concentration parameter

Table 2.1: Parameters in the BayesCluster model for mixed type data, their domains and interpretation

schemes for models with conjugate and non-conjugate priors, whereas Ishwaran and James [2001] propose a blocked Gibbs sampler. However, latent variable models are often optimised with respect to the model parameters.

We will first outline the MCMC steps for updating the model parameters, which we then use to define a simulated annealing optimiser.

Since the mixing proportions $\boldsymbol{\pi}$ have a symmetric Dirichlet process prior with concentration parameter α which is conjugate to the multinomial prior on the cluster indicator c_i, we can integrate out the mixing proportions and thus have fewer parameters to learn:

$$p(c_{1:N}|\alpha) = \int p(c_{1:N}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)\mathrm{d}\boldsymbol{\pi} \tag{2.3}$$

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha+N)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\frac{\alpha}{K})}. \tag{2.4}$$

In this model specification, we assume that the hyperparameter α is fixed but later on we will adopt a Bayesian approach and infer α. Hence, the model parameters we have left to infer are the cluster indicators c_i, the latent factors $\mathbf{Z}$, the cluster means $\boldsymbol{\mu}_k$, the loadings matrix $\mathbf{W}$ and σ^2.

To infer the cluster partition C, we follow the procedure outlined by Neal [2000]. It states that the posterior marginal distribution of the cluster indicators, given all the other model parameters and data, is fully specified by the computing the probability that $c_i = k$, where k is an existing occupied cluster, and the probability that $c_i = k^*$, where k^* is a new cluster.

Using Bayes' rule, we can express the probability of assigning the ith observation to an existing cluster k as follows

$$p(c_i = k|c_{-i}, \mathbf{Z}, \boldsymbol{\pi}, \alpha) \propto p(c_i = k|c_{-i}, \boldsymbol{\pi}, \alpha)p(\mathbf{z}_i|\mathbf{z}_{-i}, c_i = k, c_{-i}) \tag{2.5}$$

$$= \frac{p(c_{1:N}|\alpha)}{p(c_{-i}|\alpha)} p(\mathbf{z}_i|\mathbf{z}_{-i}, c_i = k, c_{-i}), \tag{2.6}$$

where c_{-i} denotes all cluster indicators of the latent variables excluding the ith one, and $\mathbf{z}_{-i}$ denotes all latent variables without the ith one. We have omitted the normalising constant.

By exchangeability we assume that c_i is the cluster indicator variable of the last data point. Following Antoniak [1974], we can derive a simpler expression for the numerator in the first term in (2.6) as follows:

$$p(c_1, \ldots, c_N|\alpha) = \int p(c_1, \ldots, c_N|\pi)p(\pi|\alpha)\mathrm{d}\pi \tag{2.7}$$

$$= \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})}, \tag{2.8}$$

where N_k is the number of data points allocated to cluster k and we have made use of the relation $\Gamma(x+1) = x\Gamma(x)$. $\Gamma(x)$ is the gamma function defined as follows for

$x > 0$ [Abramowitz and Stegun, 1965]:

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) \mathrm{d}u \tag{2.9}$$

Hence,

$$\begin{aligned} p(c_i = k | c_{-i}, \alpha) &= \frac{p(c_{1:N}|\alpha)}{p(c_{-i}|\alpha)} && (2.10) \\ &= \frac{\frac{1}{\Gamma(N+\alpha)}}{\frac{1}{\Gamma(N+\alpha-1)}} \times \frac{\Gamma(N_k + \frac{\alpha}{K})}{\Gamma(N_{k,-i} + \frac{\alpha}{K})} && (2.11) \\ &= \frac{N_{k,-i} + \frac{\alpha}{K}}{N + \alpha - 1}. && (2.12) \end{aligned}$$

The second term in (2.6) is equal to $\mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I})$ using Rasmussen [1999].

Hence, the probability of assigning $\mathbf{z}_i$ to an existing cluster k is equivalent to

$$p(c_i = k | c_{-i}, \mathbf{Z}, \boldsymbol{\pi}, \alpha) = \frac{N_{k,-i} + \frac{\alpha}{K}}{N + \alpha - 1} \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}). \tag{2.13}$$

This derivation assumes that we are working with a finite mixture model when we know/have fixed the number of occupied components K. Instead, we can work with an infinite Dirichlet mixture model and learn the number of occupied clusters K from the data which is a more realistic scenario when we work with real-world datasets and do not know the ground truth. If we take $K \to \infty$, then Equation (2.13) is equal to

$$p(c_i = k | c_{-i}, \mathbf{Z}, \boldsymbol{\pi}, \alpha) \propto \frac{N_{k,-i}}{N + \alpha - 1} \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}). \tag{2.14}$$

To derive the probability of starting a new cluster, we follow Rasmussen [1999] and Neal [2000], and get that it is equivalent to

$$p(c_i = k^* | c_{-i}, \mathbf{Z}, \pi, \alpha) \propto \frac{\alpha}{N + \alpha - 1} \int \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I}) \mathrm{d}\boldsymbol{\mu}_k. \tag{2.15}$$

We have omitted the normalising constant here as well.

After we have computed the probabilities of allocating the i^{th} observation to any of the existing clusters and of starting a new cluster, one would normally perform a Gibbs sampling step. However, we instead choose to convert this to an optimiser by assigning the i^{th} item to the most probable cluster. Throughout the whole algorithm, we work with log probabilities as they offer improved numerical stability. In addition, maximising the log probability is equivalent to maximising the probability.

After we have reallocated all the data points, we can update the other model parameters. We present below the derivations of the conditional distributions, which we obtain using the model specification in 2.3 and the graphical model, presented on Figure 2.1.

The conditional distribution of the mean $\boldsymbol{\mu}_k$ of cluster k is equal to:

$$\begin{aligned} p(\boldsymbol{\mu}_k|\mathbf{Z}, C) &\propto p(\boldsymbol{\mu}_k) \prod_{i:c_i=k} p(\mathbf{z}_i|c_i = k, \boldsymbol{\mu}_k) \\ &= \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I}) \prod_{i:c_i=k} \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \\ &= \mathcal{N}\Big(\boldsymbol{\mu}_k|\frac{1}{N_k+1} \sum_{i:c_i=k} \mathbf{z}_i, \frac{1}{N_k+1}\mathbf{I}\Big), \end{aligned}$$

where N_k is the number of observations allocated to the kth cluster.

Using the identities related to computing conditional distributions of Gaussian variables presented in Chapter 2 of Bishop [2006], we find that the conditional distribution of the latent factors $\mathbf{Z}$ have the following form

$$\begin{aligned} p(\mathbf{z}_i|c_i, \boldsymbol{\mu}_k, \mathbf{x}_i, \mathbf{W}, \boldsymbol{\varepsilon}) &\propto p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}, \boldsymbol{\varepsilon}) p(\mathbf{z}_i|c_i, \boldsymbol{\mu}_k) \\ &= \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})\mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \\ &= \mathcal{N}(\mathbf{z}_i|(\sigma^2\mathbf{W}^\intercal\mathbf{W} + \mathbf{I})^{-1}(\mathbf{W}^\intercal(\sigma^2\mathbf{I}\mathbf{x}_i + \boldsymbol{\mu}_k), (\sigma^2\mathbf{W}^\intercal\mathbf{W} + \mathbf{I})^{-1}). \end{aligned}$$

We are not able to obtain a closed-form expression for the conditional distribution of the loadings matrix $\mathbf{W}$. However, we can use the approximation $\mathbf{X} \approx \mathbf{WZ}$ to update $\mathbf{W}$ after updating the latent factors $\mathbf{Z}$. We do that in the following way: let $\mathbf{Z}^*$ denote the updated latent variables. Then we obtain an update for $\mathbf{W}$ by solving $\mathbf{X} = \mathbf{W}\mathbf{Z}^*$. We have that

$$\mathbf{X}\mathbf{Z}^{*\intercal} = \mathbf{W}\mathbf{Z}^*\mathbf{Z}^{*\intercal} \tag{2.16}$$

which implies that we can use the following expression to update $\mathbf{W}$ as follows

$$\mathbf{W} = \mathbf{X}\mathbf{Z}^{*\intercal}(\mathbf{Z}^*\mathbf{Z}^{*\intercal})^{-1}. \tag{2.17}$$

Finally, we can find the conditional distribution for σ^2 using the results in Murphy

[2007]:

$$p(\sigma^2|\mathbf{X},\mathbf{Z},\mathbf{W},C,\boldsymbol{\mu}) = \mathrm{IG}\Big(\frac{1+N}{2}, \frac{1+(\mathbf{X}-\mathbf{WZ})^\intercal(\mathbf{X}-\mathbf{WZ})}{2}\Big). \tag{2.18}$$

Since we are going to work with different types of data, for example discrete data where we can not obtain closed-form expressions for most of the posterior distributions of the model parameters, we would like to use an inference scheme which can be applicable for all data types. One such method is random walk Metropolis Hastings.

We can apply it in the case of continuous BayesCluster as follows:

- Reassign the cluster indicators C using (2.14) and (2.15);
- At iteration t, propose a move from the current state $\mathcal{S} = \{\mathbf{Z}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\varepsilon}^{(t)}, \mathbf{W}^{(t)}\}$ to a new state $\mathcal{S}^*$, where the latent factors $\mathbf{Z}$, cluster means $\boldsymbol{\mu}$, and error term $\boldsymbol{\varepsilon}$ have been updated using the following proposal distributions:
  - $\mathbf{z}_j^* = \mathbf{z}_j^{(t)} + \mathcal{N}(\mathbf{0}, 0.01 \times \mathbf{I})$
  - $\boldsymbol{\mu}_k^* = \boldsymbol{\mu}_k^{(t)} + \mathcal{N}(\mathbf{0}, 0.01 \times \mathbf{I})$
  - $\sigma^* = \sigma^{(t)} + \mathcal{N}(0, 0.001)$

  and the update of the loadings matrix $\mathbf{W}$ has been obtained using the approximation $\mathbf{X} \approx \mathbf{WZ}$.
- Accept the move to $\mathcal{S}^*$ with probability $\min(1, r)$, where $r = \exp(f(S^*) - f(S))$ and f is the model log posterior calculated as outlined in Appendix B.1; otherwise remain in $\mathcal{S}$.

The MCMC methods, however, are mainly useful when we work with datasets of small or moderate size and become infeasible in large-scale data analyses due to the computation costs. In addition, working with Metropolis Hastings requires manually tuning the proposal distributions in order to explore the posterior space efficiently (Chapter 1). There are numerous alternatives to using MCMC methods in Dirichlet process mixture models: predictive recursion [Newton and Zhang, 1999; Martin et al., 2009], variational Bayes [Blei et al., 2006; Kurihara et al., 2007], weighted Chinese restaurant sampling [Ishwaran and Takahara, 2002; Ishwaran and James, 2003] and sequential importance sampling [Bush and MacEachern, 1996; Quintana and Newton, 2000]. However, the predictive recursion method requires the approximation of a normalising constant at each update step, whereas the variational methods involve more parameters to be tuned and are often sensitive to the starting values.

We propose instead an inference scheme based on simulated annealing since it is a stochastic optimiser and can avoid getting stuck in local maxima. In addition, using simulated annealing will provide us with a near optimal solution much faster than the MCMC methods. The main drawback of simulated annealing is that we do not get the full posterior of the model parameters but since our main interest will be in finding a single summary clustering partition, we believe this to be an acceptable trade-off.

Section 1.2.5 highlights the close relation between random walk Metropolis Hastings and simulated annealing, and shows that we can easily adapt the inference procedure in BayesCluster to perform simulated annealing instead. We proceed in the following way: after we reassign all cluster indicators C using (2.14) and (2.15), we propose a move from the current state $\mathcal{S}$ to a state $\mathcal{S}^*$, where the latent factors $\mathbf{Z}$, cluster means $\boldsymbol{\mu}$, and error term $\boldsymbol{\varepsilon}$ have been updated using the same proposal distributions as in Section 2.1.1, and the update of the loadings matrix $\mathbf{W}$ has been obtained using the approximation $\mathbf{X} \approx \mathbf{WZ}$. After the proposal of the new state $\mathcal{S}^*$, we compute

$$r = \exp\Big(\frac{(f(\mathcal{S}^*) - f(\mathcal{S})}{T_k}\Big), \tag{2.19}$$

where f is the model log posterior and T_k is the temperature at iteration k of the cooling schedule. We accept the new state $\mathcal{S}^*$ and update the model parameters with probability $\min(1, r)$, otherwise we stay in the current state $\mathcal{S}$ and do not update model parameters. We finish this iteration by increasing the iteration counter from k to $k + 1$.

We use an exponential cooling schedule $T_k = T_0 C^k$, with $T_0 = 100$ and $C = 0.95$, which have been picked based on numerical experiments. At the end of each iteration, we check for convergence using the log posterior, with the stopping criterion being $|f_k - f_{k-1}| < 0.0001$, and continue until either a certain number of iterations has been reached (we set the maximum number of iterations to 1000, which our experiments have shown to be sufficient to explore the parameter space) the model has converged.

Initialisation and model selection

Simulated annealing is often initialised with a random starting point but its performance can be improved by using a heuristic strategy such as a k-means solution [Van Laarhoven and Aarts, 1987]. We adopt this initialisation approach in our experiments. Since the k-means solution is sensitive to the initial choice of cluster centres [Baswade and Nalwade, 2013; Bradley and Fayyad, 1998], we initialise the model from a few different random starting points (5 in the experiments in this thesis) and choose the final clustering partition to be the one corresponding to the maximum a posteriori for each of the random initialisations. We apply PPCA to initialise the latent variables $\mathbf{Z}$ and the loadings $\mathbf{W}$, and sample σ^2 from $\mathrm{IG}(1, 1)$.

We run BayesCluster for a range of number of latent dimensions ($P = 2, \ldots, 10$) and we use the Bayesian information criterion to select P. An alternative approach to model selection is to set a prior over the columns of $\mathbf{W}$ as in Bayesian PCA and infer the number of latent dimensions automatically. This is straightforward when we are working with continuous data. However, when we work with discrete data, we would need to introduce a prior over the columns of each loadings matrix, which increases the model complexity. Since we want to select the final model in a similar manner for the different types of data (continuous, discrete and mixed), we choose to use the Bayesian information criterion for model selection in BayesCluster.

We can summarise the workflow of continuous BayesCluster as follows:

Algorithm 2.1: Continuous BayesCluster

```
Perform PPCA to initialise the latent variables $\mathbf{Z}$ and the loadings matrix $\mathbf{W}$,
  and sample $\sigma^2$ from $\mathrm{IG}(1, 1)$ ;
Initialise the cluster partition by using k-means clustering on the latent factors
  $\mathbf{Z}$;
while $t < num_{iterations}$ & not converged do
    Sample a random permutation $\tau$ of $1, \dots, N$ ;
    for $j \in \tau$ do
        Remove the $i^{th}$ observation from its current cluster and update the
          cluster's sufficient statistics ;
        Compute the probabilities of joining an existing cluster and of starting a
          new cluster using (2.14) and (2.15);
        Set $c_i = \arg\max_{1,\dots,K,k^*} \log p(c_i = k | c_{-i}, \mathbf{Z}, \boldsymbol{\pi}, \alpha)$ and update the
          cluster's sufficient statistics ;
    end
    Update the model parameters using simulated annealing ;
    Compute the model log posterior ;
    Check for convergence.
end
```

2.1.2 Model specification for discrete data

BayesCluster can be applied to different types of discrete data after making certain adjustments. We focus on modelling categorical data, which is the one we most commonly encounter in the applications we consider in this thesis.

The data which we model has the form $x_{ij} \in \{1, \dots, r_j\}$, where j is the number of features and r_j is the number of categories for the jth feature. We assume that each x_{ij} is generated from a latent variable $\mathbf{z}_i \in \mathbb{R}^P$, which has a Gaussian prior, and is passed through the softmax function as follows

$$p(\mathbf{z}_i | \boldsymbol{\mu}_k, c_i = k) = \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{I}) \tag{2.20}$$

$$p(\mathbf{x}_i | \mathbf{z}_i, \theta) = \prod_{r=1}^{R} \mathrm{Cat}(x_{ir} | \mathcal{S}(\mathbf{W}_r^T \mathbf{z}_i + \mathbf{w}_{0r})) \tag{2.21}$$

where $\mathbf{W}_r \in \mathbb{R}^{P \times M}$ is the loadings matrix for the r^{th} feature, and $\mathbf{w}_{0r} \in \mathbb{R}^M$ is the

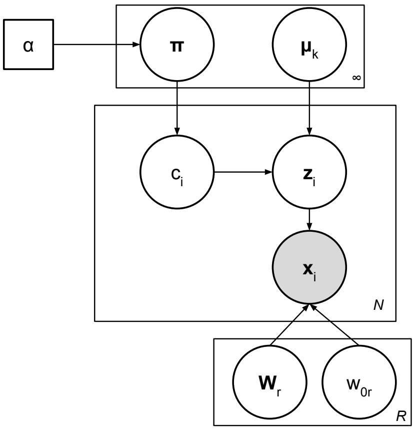


Figure 2.2: A graphical model, representing the independence assumptions of Categorical BayesCluster.

offset for the r^{th} feature. We place a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ on each row of of each loadings matrix $\mathbf{W}_r$, and similarly a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ on the offsets $\mathbf{w}_{0r}$ as suggested by [Khan et al., 2010]. Similarly to the continuous case of BayesCluster, we place a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ on the means of the clusters of latent variables $\boldsymbol{\mu}_k$. The graphical representation of the model is shown on Figure 2.2.

We can summarise the probabilistic model and the assumptions we make as follows:

$$\begin{aligned}
p(\pi|\alpha) =& \mathrm{Dir}(\alpha) \\
p(c_i|\boldsymbol{\pi}) =& \mathrm{Mult}(\boldsymbol{\pi}) \\
p(\mathbf{z}_i|c_i = k, \mu_k) =& \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \\
p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}_r, \mathbf{w}_{0r}) =& \prod_{r=1}^{R} \mathrm{Cat}(x_{ir}|\mathcal{S}(\mathbf{W}_r^\intercal \mathbf{z}_i + \mathbf{w}_{0r})) \\
p(\boldsymbol{\mu}_k) =& \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I}) \\
p(\mathbf{W}_r) =& \prod_{j=1}^{J} \mathcal{N}(\mathbf{w}_j|\mathbf{0}, \mathbf{I}) \\
p(\mathbf{w}_{0r}) =& \mathcal{N}(\mathbf{w}_{0r}|\mathbf{0}, \mathbf{I}).
\end{aligned}$$

Inference

The inference procedure for the discrete BayesCluster is very similar to the model for continuous data. We briefly summarise it below, highlighting the differences.

After we apply logistic PCA [Landgraf and Lee, 2015] to initialise the latent factors

$\mathbf{Z}$, we use k-means to obtain the initial cluster membership.

The probabilities of the i^{th} observation joining an existing cluster or starting a new cluster are the same as in the model for continuous data and are given by (2.14) and (2.15). We then allocate the observation to the cluster with the highest log probability, and after we have reallocated all of them, we update the model parameters simulated annealing and check for convergence. We repeat until either convergence or we have reached a certain number of iterations.

We infer the model parameters using simulated annealing as we can not derive their conditional posteriors in closed form. This is because the marginal likelihood of the observed variables is

$$p(\mathbf{x}_{i,1:R}) = \int [\prod_{r=1}^{R} p(x_{ir}|\mathbf{z}_i, \mathbf{W}_r, \mathbf{w}_{0r})] \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \mathrm{d}\mathbf{z}_i, \tag{2.22}$$

which can not be computed because of the lack of conjugacy.

We use similar settings to the continuous BayesCluster ones to perform simulated annealing. The proposal distributions for $\mathbf{Z}$ and $\boldsymbol{\mu}$ are given in Section (2.1.1). In the case of modelling discrete data, we need to infer all model parameters $\mathbf{Z}$, $\mathbf{W}^D$, $\mathbf{w}_0^D$, $\boldsymbol{\mu}$, C as we cannot use the approximation $\mathbf{X} \approx \mathbf{WZ}$ to find the loadings matrices $\mathbf{W}^D$. The proposal distributions for the loading matrices $\mathbf{W}^D$ and the offset terms $\mathbf{w}_0^D$ are as follows:

$$\mathbf{W}_{.r}^{D*} = \mathbf{W}_{.r}^{D(t)} + \mathcal{N}(\mathbf{0}, 0.001 \times \mathbf{I}) \tag{2.23}$$

$$\mathbf{w}_{0r}^{D*} = \mathbf{w}_{0r}^{D(t)} + \mathcal{N}(\mathbf{0}, 0.001 \times \mathbf{I}). \tag{2.24}$$

We accept moving to the new state $\mathcal{S}^*$ with probability $\min(1, r)$, where r is found using (2.19) and f is the model log posterior calculated using the derivation in Appendix B.2.

We initialise the model and determine the final partition and number of occupied clusters in the same manner as in the model for continuous data.

We can summarise the steps involved in discrete BayesCluster as follows:

Algorithm 2.2: Discrete BayesCluster

```
Perform Categorical PCA to initialise the latent variables $\mathbf{Z}$, sample the
  loadings matrices $\mathbf{W}^D$ and offsets $\mathbf{w}_0$ from the corresponding priors ;
Initialise the cluster partition by using k-means clustering on the latent
  variables $\mathbf{Z}$;
while $t < num_{iterations}$ & not converged do
    Sample a random permutation $\tau$ of $1, \dots, N$ ;
    for $i \in \tau$ do
        Remove $i^{th}$ observation from its current cluster and update cluster's
          sufficient statistics ;
        Compute the probabilities of joining an existing cluster and of starting a
          new cluster using (2.14) and (2.15);
        Set $c_i = \arg\max_{1,\dots,K,k^*} \log p(c_i = k|c_{-i}, \mathbf{Z}, \pi, \alpha)$ and update cluster's
          sufficient statistics ;
        Update model parameters using simulated annealing ;
        Compute the model log posterior using B.2;
        Check for convergence.
    end
end
```

Modelling other types of discrete data

We can model other types of discrete data by using the same modelling framework where we reduce the dimensionality and cluster the latent factors $\mathbf{Z}$ with Dirichlet process mixture model. For example, if we have count data, then we can use a Poisson model [Murphy, 2012]

$$p(\mathbf{x}_i|\mathbf{z}_i) = \prod_{v=1}^{V} \text{Poi}(x_{iv}|\exp(\mathbf{w}_{v,:}^T \mathbf{z}_i)) \tag{2.25}$$

to model the dataset appropriately. This model is an example of exponential family PCA, developed by Collins et al. [2002] and Mohamed et al. [2009].

If we have have ordinal data, we can use item response theory [Johnson and Albert, 2006; Fox, 2010] for example, which assumes that the observed variables $\mathbf{x}_i$ are the categorical manifestation of the latent variables $\mathbf{z}_i$. For each ordinal variable x_{ij} with K_j levels, we assume that its value is determined by the value of $\mathbf{z}_i$ in relation to $K_j + 1$ vector of thresholds λ_j, $(-\infty = \lambda_{j,0} \leq \lambda_{j,1} \leq \dots \lambda_{j,Kj} \leq \infty)$. If, for example, the latent $\mathbf{z}_i$ is such that $\lambda_{j,k-1} < \mathbf{z}_i < \lambda_{j,k}$, then the value of $\mathbf{x}_i$ is k. In addition, we can express the probability of observing a variable of level k as the difference between two Gaussian cumulative distributions:

$$p(\mathbf{x}_i = k) = \Phi\Big(\frac{\delta_k - \boldsymbol{\mu}_i}{\sigma_i}\Big) - \Phi\Big(\frac{\delta_{k-1} - \boldsymbol{\mu}_i}{\sigma_i}\Big), \tag{2.26}$$

where δ_k is the proportion of observed values of variable i which are less or equal to k, and $\mathbf{z}_i \sim \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_i, \sigma_i)$.

2.1.3 Model specification for mixed data

We can combine the models presented in Sections 2.1.1 and 2.1.2 to model mixed data. To ease the task of having to cluster mixed type data, we consider the dataset to be the result of the integration of two or more smaller datasets. For example, if there is a dataset with real-valued and categorical features, we treat the dataset as the integration of a real-valued and a categorical dataset, which are used to jointly infer a single set of latent variables, and thereby a single clustering partition.

As the data we consider consists of both continuous and categorical observations, we denote by $\mathbf{x}_i^C$ the continuous vector corresponding to the i^{th} observation, and by $\mathbf{x}_i^D$ - the categorical vector. We assume that each $\mathbf{x}_i^C$ is Normally distributed $\mathcal{N}(\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})$, and each discrete variable x_{ir}^D is multinomially distributed with parameters achieved through the softmax transformation of $\mathbf{W}_r^D\mathbf{z}_i + \mathbf{w}_{0r}$. We can summarise the model as follows:

$$p(\mathbf{z}_i|\boldsymbol{\mu}_k, c_i = k) = \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{I}) \tag{2.27}$$

$$p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}^C, \boldsymbol{\varepsilon}, \mathbf{W}^D, \mathbf{w}_0^D) = \mathcal{N}(\mathbf{x}_i^C|\mathbf{W}^C\mathbf{z}_i, \sigma^2\mathbf{I}) \prod_{r=1}^{R} \mathrm{Cat}(x_{ir}^D|\mathcal{S}(\mathbf{W}_r^{D\intercal}\mathbf{z}_i + \mathbf{w}_{0r})). \tag{2.28}$$

where $\mathbf{W}$ and $\boldsymbol{\varepsilon}$ are the loadings matrix and error, respectively, for the continuous observations, $\mathbf{W}^D = \{\mathbf{W}_1^D, \dots, \mathbf{W}_R^D\}$ and $\mathbf{w}_0^D = \{\mathbf{w}_{01}^D, \dots, \mathbf{w}_{0R}^D\}$ are the loadings matrices and offset terms, respectively, for the discrete variables. We place a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ on each row of each loadings matrix $\mathbf{W}_r^D$ and of $\mathbf{W}$, and similarly a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ on the offsets $\mathbf{w}_{0r}^D$ as suggested by Khan et al. [2010]. Similarly to the continuous and discrete cases, we place a Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ on the cluster mean $\boldsymbol{\mu}_k$ and IG$(1, 1)$ prior on σ^2. The plate diagram of the model is

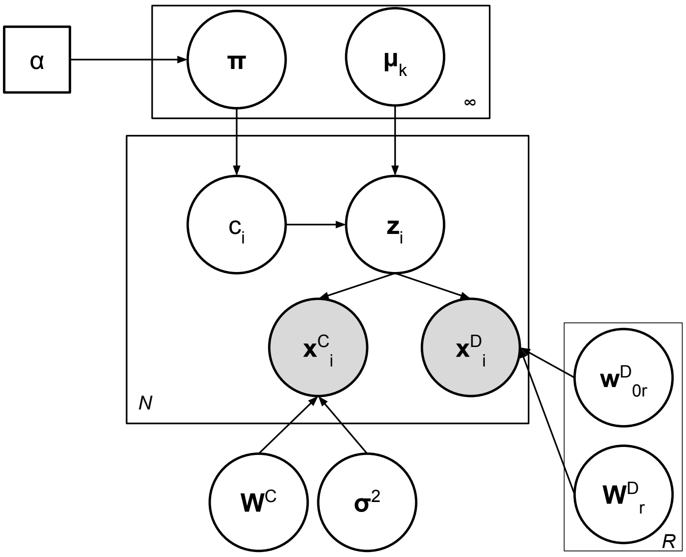


Figure 2.3: A graphical model, representing the independence assumptions of Mixed BayesCluster.

presented on Figure 2.3.

We can summarise the probabilistic model and the assumptions we make as follows:

$$p(\pi|\alpha) = \mathrm{Dir}(\alpha) \tag{2.29}$$

$$p(c_i|\boldsymbol{\pi}) = \mathrm{Mult}(\boldsymbol{\pi}) \tag{2.30}$$

$$p(\mathbf{z}_i|c_i = k, \boldsymbol{\mu}_k) = \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \tag{2.31}$$

$$p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}^C, \mathbf{W}^D, \mathbf{w}_{0r}, \boldsymbol{\varepsilon}) = \mathcal{N}(\mathbf{x}_i|\mathbf{W}^C\mathbf{z}_i, \sigma^2\mathbf{I}) \prod_{r=1}^{R} \mathrm{Cat}(x_{ir}|\mathcal{S}(\mathbf{W}_r^{D\intercal}\mathbf{z}_i + \mathbf{w}_{0r}^D)) \tag{2.32}$$

$$p(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I}) \tag{2.33}$$

$$p(\mathbf{W}^C) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_d^C|\mathbf{0}, \mathbf{I}) \tag{2.34}$$

$$p(\boldsymbol{\varepsilon}) = \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0}, \sigma^2\mathbf{I}) \tag{2.35}$$

$$p(\mathbf{W}_r^D) = \prod_{j=1}^{J} \mathcal{N}(\mathbf{w}_j^D|\mathbf{0}, \mathbf{I}) \tag{2.36}$$

$$p(\mathbf{w}_{0r}^D) = \mathcal{N}(\mathbf{w}_{0r}^D|\mathbf{0}, \mathbf{I}). \tag{2.37}$$

Inference

We assume that the continuous and discrete parts of the dataset share the same latent variables. This follows from our aim of trying to learn a single coherent latent representation of all our data. We perform jointly the lower dimensional projection and the learning of the clustering structure by applying categorical PCA and PPCA to the discrete and continuous parts of the dataset respectively and by using a Dirichlet process mixture to model the cluster membership.

We reassign the data points to new clusters using (2.14) and (2.15) in the same way as for the continuous and discrete data. Khan et al. [2010] suggest using variational Expectation-Maximisation to infer the other model parameters. This, however, increases the already high number of parameters we need to learn and requires the use of approximations. Thus, we use similar inference scheme to the ones employed in the continuous and discrete BayesCluster models. We use simulated annealing to infer the latent variables $\mathbf{Z}$, the loadings matrices $\mathbf{W}^D$, offsets $\mathbf{w}_0$ and residual error $\boldsymbol{\varepsilon}$, and use the approximation $\mathbf{X} \approx \mathbf{W}^C\mathbf{Z}$ to find the loadings matrix $\mathbf{W}^C$. The proposal distributions for the parameters in the new state $\mathcal{S}^*$ are given in (2.1.1) and (2.23). We accept the move to the new state with probability $\min(1, r)$, where r is found using (2.19) and f is the model log posterior, calculated using the derivation in Appendix B.3.

We initialise the model and determine the final partition and number of occupied clusters in the same manner as in the model for continuous data.

We can summarise the steps involved in mixed BayesCluster as follows:

Algorithm 2.3: Mixed BayesCluster

```
Perform PPCA or Categorical PCA to initialise the latent variables $\mathbf{Z}$, sample
  the loadings matrices $\mathbf{W}^C$, $\mathbf{W}^D$, offsets $\mathbf{w}_0^D$ and residual error $\boldsymbol{\epsilon}$ from the
  corresponding priors ;
Initialise the cluster partition using k-means clustering;
while $t < num_{iterations}$ & not converged do
    Sample a random permutation $\tau$ of $1, \dots, N$ ;
    for $i \in \tau$ do
        Remove the $i^{th}$ observation from its current cluster and update cluster's
          sufficient statistics ;
        Compute the probabilities of joining an existing cluster and of starting a
          new cluster using (2.14) and (2.15);
        Set $c_j = \arg\max_{1,\dots,K,k^*} \log p(c_j = k | c_{-j}, \mathbf{Z}, \boldsymbol{\pi}, \alpha)$ and update the
          cluster's sufficient statistics ;
    end
    Update model parameters using simulated annealing ;
    Compute the model log posterior ;
    Check for convergence.
end
```

2.2 Conclusions

We have presented a novel method for Bayesian clustering based on Dirichlet process mixtures and linear latent variable models to handle mixed data types, called BayesCluster. It has several advantages over traditional approaches, which we have highlighted throughout this chapter, and it can be applied to continuous, discrete and mixed data. In Chapter 3 we explore the applicability of BayesCluster with synthetic and real-world datasets, and in Chapter 4, we propose several extensions which lead to the identification of more interpretable and well-defined clusters.

Chapter 3

Mixed Data Clustering: Numerical Experiments

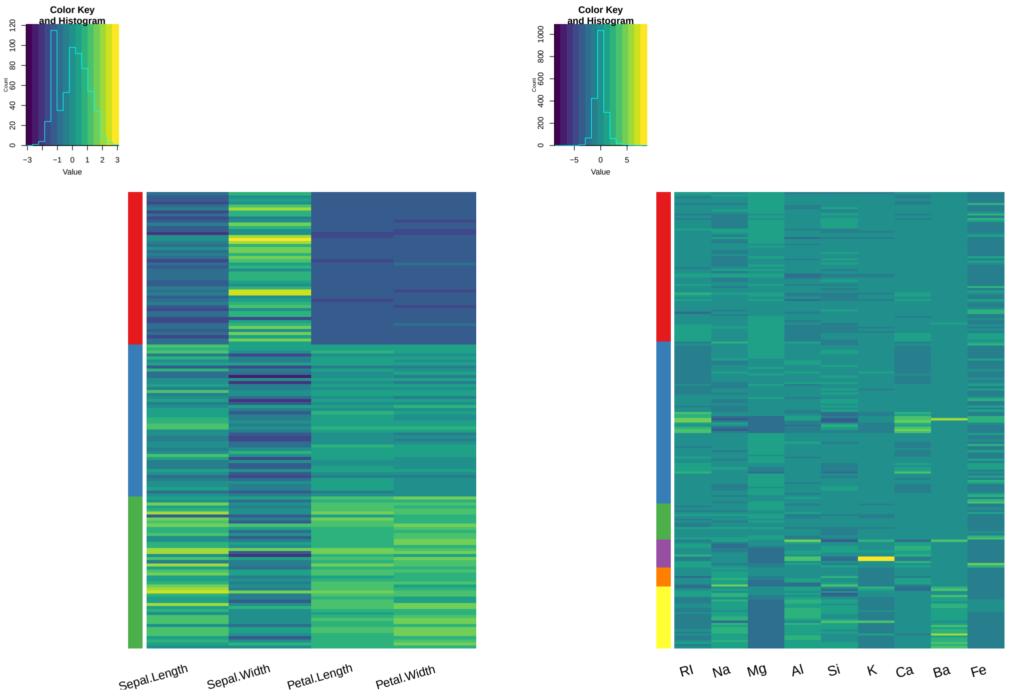
We introduced BayesCluster in Chapter 2, which can be applied to continuous, discrete or mixed data to identify meaningful clusters. In this chapter, we demonstrate its useful properties with synthetic and real datasets. We compare its performance with other commonly used methods for clustering continuous, discrete and mixed data.

Code to perform the experiments in this chapter is available at: `https://github.com/ilianapeneva`.

3.1 Continuous data

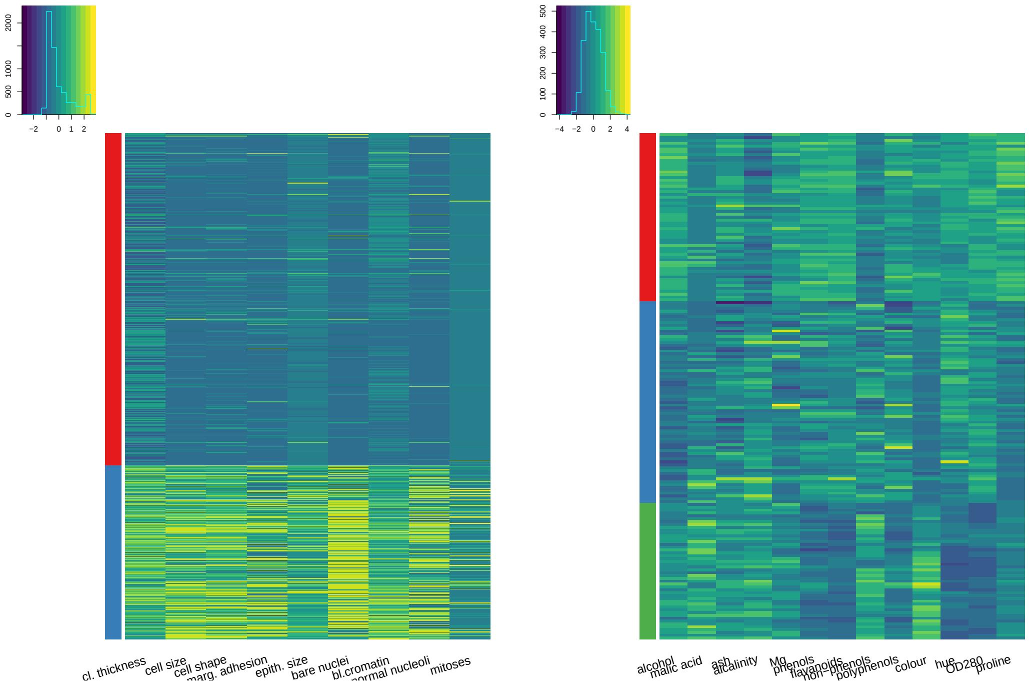
We compared continuous BayesCluster with k-means clustering, Gaussian mixture model (GMM) and iClusterPlus over 5 datasets (4 real and 1 synthetic). Another frequently used clustering approach involves first projecting the dataset down to a P-dimensional space using PPCA or PCA, with P chosen using cross-validation, and then using GMM or k-means to perform clustering in this lower dimensional space. Some examples of this approach can be found in [Holter et al., 2000; Alter et al., 2000]. However, this procedure is very inflexible as once the number of dimensions is selected, it remains fixed throughout the clustering process. This means that if the data distribution is different from the assumed one, the dimensions selected using PPCA or PCA might deviate from the optimal and hence, the quality of the clustering output could be poor. We ideally want to do the dimensionality reduction and clustering jointly because this allows us to adapt the approach more easily to working with mixed data types and with multiple datasets.

We selected real datasets from the University of California Irvine (UCI) machine learning repository which are well studied and for which the ground truth is known. These datasets are often not high-dimensional but the availability of the ground truth enables the comparison with other clustering methods. We investigate the performance of BayesCluster on high-dimensional datasets with synthetic datasets. The four real datasets we used are the iris dataset (150 observations, 3 classes (setosa, versicolor, virginica), 4 attributes), Wisconsin breast cancer diagnostic dataset (569 observations, 2 classes (benign, malignant), 10 attributes), glass identification dataset (214 observations, 7 classes, 10 features), and wine dataset (178 observations, 3 classes, 12 features) which are available on the UCI machine learning repository `https://archive.ics.uci.edu/ml/datasets/`. The heatmaps of the normalised datasets (zero mean and unit variance) are presented on Figure 3.1. We used the generative model of BayesCluster to create 10 synthetic datasets, with each having 150 observations with 3 classes (50 observations in each class) and 100 features.



(a) iris

(b) glass



(c) Wisconsin breast cancer

(d) wine

Figure 3.1: Heatmaps of the continuous datasets. The observations are on the y-axis and are ordered according to the ground truth cluster membership, and the features are on the x-axis.

We used the following R packages in the experiments:

- ‘stats’ [R Core Team, 2018] for the implementation of k-means clustering;
- ‘mclust’ [Scrucca et al., 2017] for the implementation of Gaussian mixture models and the computation of the adjusted Rand index;
- ‘iClusterPlus’ [Mo and Shen, 2016] for the implementation of iClusterPlus.

We implemented k-means using the ‘stats’ package for number of clusters from 1 to 20 and maximum number of iterations set to 50. We used the within-group sum of squares metric and the elbow method to select the final number of clusters K (Figure C.1 in Appendix C). We implemented the Gaussian mixture model using the ‘mclust’ package with all possible covariance structures (spherical, ellipsoidal and diagonal) and for number of clusters from 1 to 9. We used BIC to select the final model (Figure C.2 in Appendix C). We implemented iClusterPlus with the default options in the package - we set the number of MCMC burn-in steps to 100, the total number of MCMC draws to 200, the number of maximum iterations for the EM algorithm to 20 and the threshold for convergence set to 1e-4. We chose to implement iClusterPlus rather than iCluster for consistency reasons as iClusterPlus can be applied to continuous, discrete and mixed data whereas iCluster can be applied to continuous data only. We ran ClusterPlus for number of latent dimensions between 1 and 10 and used the deviance ratio metric to pick the final P (Figure C.3 in Appendix C). We ran BayesCluster with 5 random initialisations for 1000 iterations and for number of latent dimensions P between 2 and 10. The threshold for convergence was set to 1e-4, and model parameters were initialised as outlined in Chapter 2. We used BIC to select the final number of latent dimensions P (Figure C.4 in Appendix C).

3.1.1 Synthetic data

We generated 10 synthetic real-valued datasets in the following way: for each dataset, we chose the first two principal components to capture most of the variation, and generated three clusters with 50 observations each, with the assumption that the datasets were normalised and have zero mean and unit variance. We sampled the three cluster means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$ as follows:

$$\boldsymbol{\mu}_1 \sim \mathcal{N}((0,0), \mathbf{I}) \tag{3.1}$$

$$\boldsymbol{\mu}_2 \sim \mathcal{N}((2,2), \mathbf{I}) \tag{3.2}$$

$$\boldsymbol{\mu}_3 \sim \mathcal{N}((4,4), \mathbf{I}). \tag{3.3}$$

We then generated the latent variables $\mathbf{Z}$ by sampling from the following Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I})$, $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{I})$ and $\mathcal{N}(\boldsymbol{\mu}_3, \mathbf{I})$. After that, we generated the loadings matrix $\mathbf{W}$ by sampling from the priors on the rows of the loadings matrices $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the error terms $\boldsymbol{\varepsilon}$ by sampling from its prior $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, where $\sigma \sim \mathrm{IG}(1,1)$. We finally generated the dataset $\mathbf{X}$ using

$$p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}, \boldsymbol{\varepsilon}) = \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}). \tag{3.4}$$

Figure 3.2 presents an example of a continuous dataset, generated using BayesCluster as a generative model.

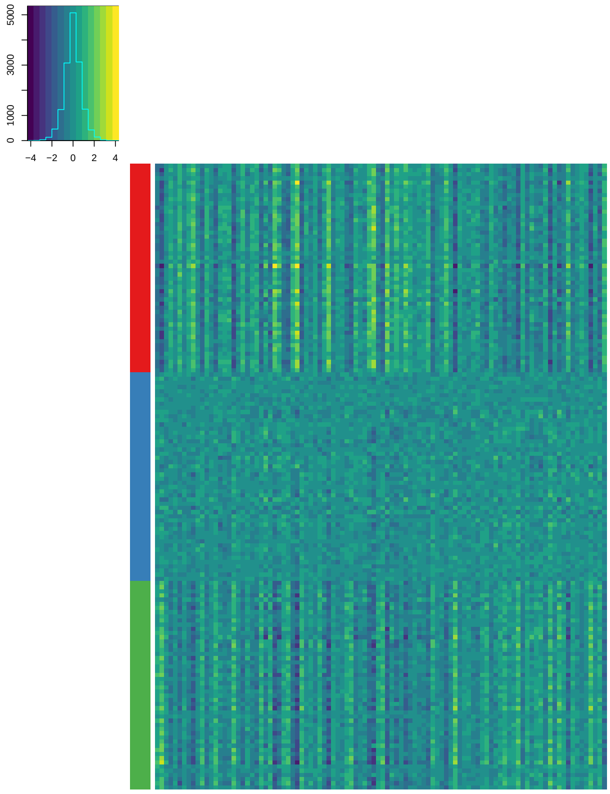
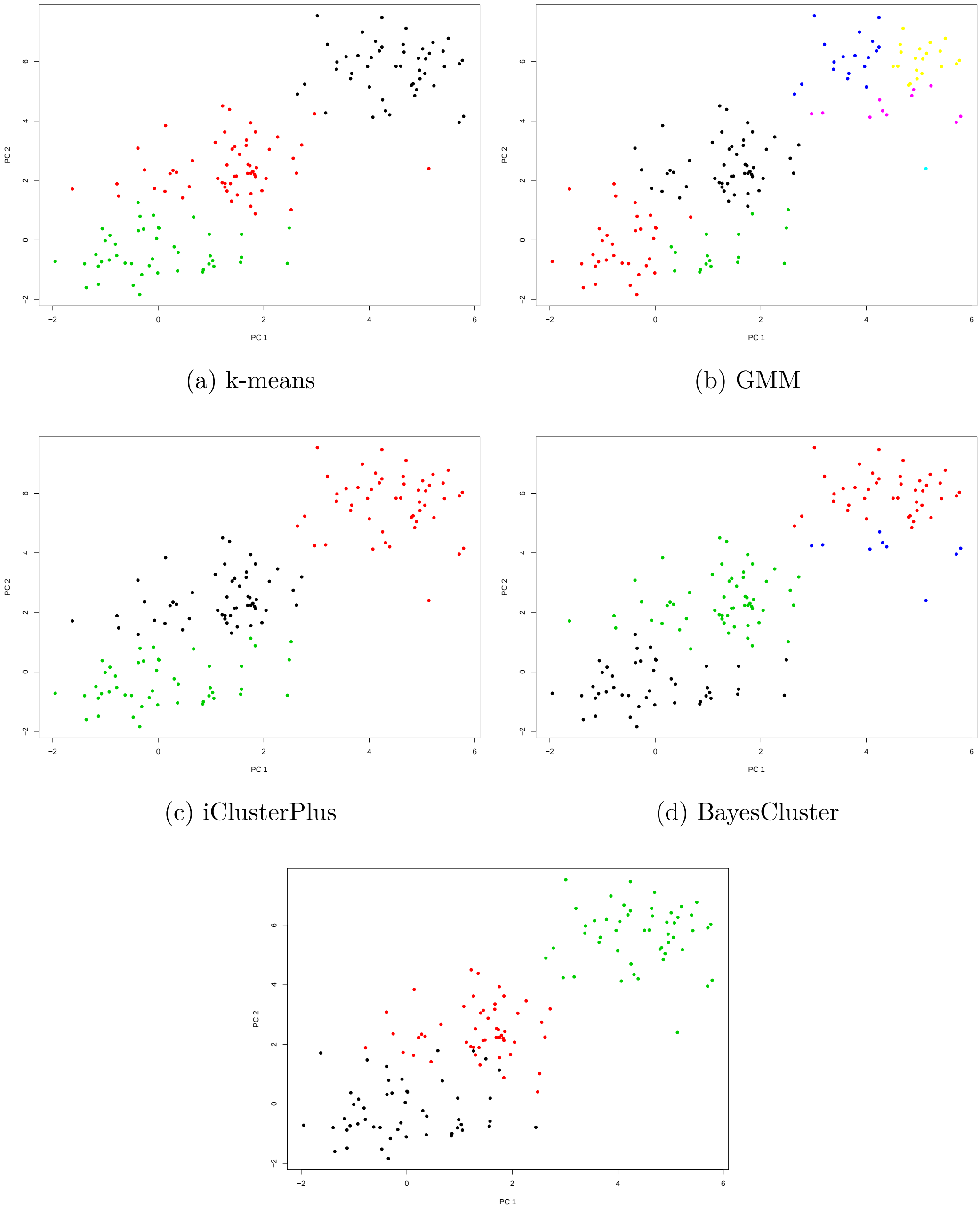


Figure 3.2: An example of a synthetic continuous dataset (zero mean and unit variance), generated using the BayesCluster generative model. The observations are on the y-axis and are ordered according to the ground truth cluster membership, and the features are on the x-axis.

The experiments with synthetic data allow us to test the model in different scenarios. With these datasets we aim to investigate how well BayesCluster can model overlapping clusters. Figure 3.3 shows the final partitions for one of the synthetic datasets. In this case, none of the methods have modelled well the overlap between the black and the red clusters: for example, GMM has created an additional cluster to model some of the overlap. iClusterPlus and k-means have identified the correct number of clusters ($K = 3$), whereas BayesCluster and GMM have overestimated it by creating additional clusters in the cases which are hard to model.



(a) k-means

(b) GMM

(c) iClusterPlus

(d) BayesCluster

(e) ground truth

Figure 3.3: Comparison between the clustering partitions of the dataset presented on Figure 3.2 with k-means clustering, Gaussian mixture model, iClusterPlus and BayesCluster, and the ground truth. The latent variables **Z** which have been used for the generation of the data are plotted and coloured according to cluster membership.

BayesCluster outperforms the other methods in regards with mean adjusted Rand index (Table 3.1). It also managed to identify the correct number of clusters ($K = 3$) in most of the datasets although it did not always identify correctly the latent dimensionality ($P = 2$).

Model	mean ARI (± st. error)	Est. K (prop.)	Est. P (prop.)	p-value	Comp. time
k-means	0.513 (0.339,0.686)	2, 3, 7 (0.5,0.4,0.1)	-	0.7496	0.53sec
GMM	0.356 (0.114,0.598)	1-7 (0.2,0.1)	-	0.1854	3.53 sec
iClusterPlus	0.446 (0.225,0.666)	3, 7, 8 (0.6,0.2,0.2)	2, 6, 7 (0.6,0.2,0.2)	0.4718	1.13 min
BayesCluster	0.52 (0.343,0.657)	3, 4 (0.6,0.4)	2 , 6 (0.7,0.3)	-	18.84 min

Table 3.1: Comparison of the results on the synthetic data in terms of adjusted Rand index (ARI) averaged over the 10 synthetic datasets, ± st. error ARI, estimates of the number of clusters K, estimates of the latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run. The proportion of times a certain value for the number of clusters K or for the number of principal components P is estimated is put into brackets after the value.

3.1.2 Real datasets

We compared the accuracy of the final partitions in terms of mean (and ± standard error) adjusted Rand Index (ARI), estimated number of clusters K, estimated latent dimensionality P and computation time. We summarise the results in Tables 3.2, 3.3, 3.4 and 3.5. In terms of computation time, BayesCluster is slower than all comparison methods, which is due to using multiple different starting positions, and to the R code for BayesCluster not being as well optimised as the other methods.

Iris dataset

All methods apart from iClusterPlus have similar performance on the iris dataset in terms of the adjusted Rand index. iClusterPlus was the only method that managed to identify the true number of clusters ($K = 3$) correctly but it did not allocate all versicolor and virginica observations in the correct clusters. K-means, the Gaussian mixture model and BayesCluster merged the observations from the versicolor and virginica groups into one cluster which could be because the observations from these two iris types have similar sepal width values.

Model	mean ARI (± st.error)	Estimated K	Estimated P	p-value	Comp. time
k-means	0.568 (0.568,0.568)	2	-	0.1679	0.07 s
GMM	0.568 (0.568,0.568)	2	-	0.1679	1.31 s
iClusterPlus	0.482 (0.413,0.552)	3	2	0.00388	2.75 s
BayesCluster	0.567 (0.566,0.568)	2	2	-	29.87 s

Table 3.2: Comparison of the results on the iris dataset in terms of mean adjusted Rand index (ARI), ± st. error ARI, estimates of the number of clusters K, estimates of the latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run.

Wisconsin breast cancer dataset

K-means clustering outperforms BayesCluster on the Wisconsin breast cancer dataset, which could be because we initialise the BayesCluster partition with the k-means solution, and then the inference scheme gets stuck in a local maximum with worse partition. The Gaussian mixture model overestimated the number of clusters and found 5 patient groups, whereas iClusterPlus found 3 clusters.

Model	mean ARI (± st.error)	Estimated K	Estimated P	p-value	Comp. time
k-means	0.839 (0.839,0.839)	2	-	7.417e-07	0.40 s
GMM	0.357 (0.357,0.357)	5	-	1.726e-11	5.41 s
iClusterPlus	0.659 (0.633,0.685)	3	2	2.797e-05	43.68 s
BayesCluster	0.728 (0.699,0.757)	2	2	-	21.45 min

Table 3.3: Comparison of the results on the Wisconsin breast cancer dataset in terms of mean adjusted Rand index (ARI), ± st. error ARI, estimates of number of clusters K, estimates of latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run.

Glass dataset

The glass dataset is the only continuous dataset where BayesCluster did not produce competitive results. The poor result may be a consequence of inappropriate assumption of Gaussianity or uninformative data (see Figure 3.1). None of the methods correctly estimated the true number of clusters ($K = 7$) and they all had low adjusted Rand indices.

Model	mean ARI (± st.error)	Estimated K	Estimated P	p-value	Comp. time
k-means	0.183 (0.155,0.211)	13	-	0.8794	0.16 s
GMM	0.156 (0.156,0.156)	3	-	0.0001	1.58 s
iClusterPlus	0.24 (0.218,0.262)	6	5	2.373e-06	14.84 s
BayesCluster	0.182 (0.168,0.195)	4	3	-	7.02 min

Table 3.4: Comparison of the results on the glass dataset in terms of mean adjusted Rand index (ARI), ± st. error ARI, estimates of the number of clusters K, estimates of the latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run.

Wine dataset

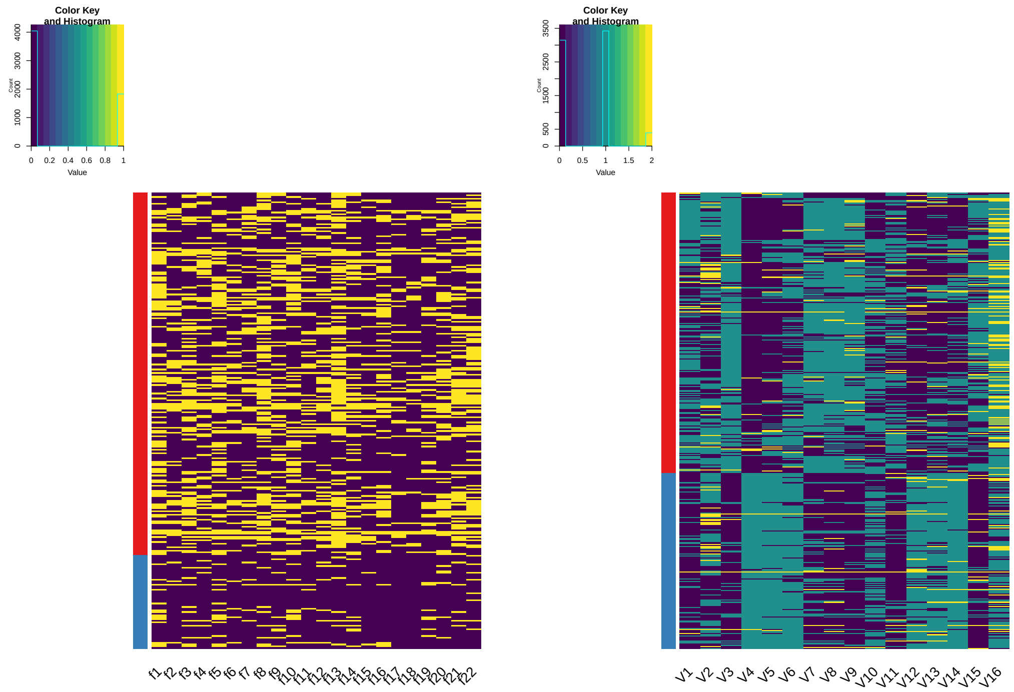
All methods apart from iClusterPlus performed well on the wine dataset in regards with both adjusted Rand index and estimated number of clusters. They all found that there were $K = 3$ clusters and clustered incorrectly only a few observations.

Model	mean ARI (± st.error)	Estimated K	Estimated P	p-value	Comp. time
k-means	0.897 (0.897,0.897)	3	-	0.0029	0.14 s
GMM	0.929 (0.929,0.929)	3	-	6.692e-06	6.16 s
iClusterPlus	0.437 (0.369,0.505)	5	4	1.51e-09	19.49 s
BayesCluster	0.872 (0.853,0.892)	3	2	-	8.94 min

Table 3.5: Comparison of the results on the wine dataset in terms of average adjusted Rand index (ARI), ± st. error ARI, estimates of the number of clusters K, estimates of the latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run.

3.2 Discrete data

We compared discrete BayesCluster with k-modes and iClusterPlus over 3 datasets (2 real and 1 synthetic). We selected real datasets from the UCI machine learning repository as the ground truth is known for them. These datasets are often not high-dimensional but the availability of the ground truth enables the comparison with other clustering methods. We investigate the performance of BayesCluster on high-dimensional datasets with synthetic datasets. The two datasets we used are Spect heart dataset (267 observations, 2 classes, 22 attributes), and congressional voting records (435 observations, 2 classes, 16 attributes), which are available on the UCI Machine Learning Repository `https://archive.ics.uci.edu/ml/datasets/`. The voting dataset consists of the votes for each of the US House Representatives congressmen on 16 key votes in 1984, and the votes are recorded as ‘yes’, ‘no’, ‘NA’. We set the missing values to be another category, which represents abstained from voting. The heatmaps of the datasets are presented on Figure 3.4.



(a) Spect heart

(b) Congressional votes

Figure 3.4: Heatmaps of the discrete datasets. The observations are on the y-axis and are ordered according to the ground truth cluster membership, and the features are on the x-axis.

We used the following R packages in the experiments:

- 'klaR' [Weihs et al., 2005] for the implementation of k-modes clustering;
- 'iClusterPlus' [Mo and Shen, 2016] for the implementation of iClusterPlus.

We implemented k-modes using the 'klaR' package with the following parameters: we set the maximum number of iterations to 20 and we did not use the weighted version of the distance between the clusters. We implemented k-modes for number of clusters from 1 to 20 and used the within-group sum of squares and the elbow method to select the number of clusters K (Figure C.5 in Appendix C). We implemented iClusterPlus with the default options in the package - we set the number of MCMC burn-in steps to 100, the total number of MCMC draws to 200, the number of maximum iterations for the EM algorithm to 20 and the threshold for convergence set to 1e-4. We ran iClusterPlus for number of latent dimensions P between 1 and 11 and used the deviance ratio metric to pick the final P (Figure C.6 in Appendix C). We ran BayesCluster for 1000 iterations with 5 random initialisations and for number of latent dimensions P between 2 and 10. The threshold for convergence set to 1e-4, and model parameters initialised as outlined in Chapter 2. We used BIC to select the final number of latent dimensions P (Figure C.7 in Appendix C).

3.2.1 Synthetic data

We used the generative model of discrete BayesCluster to create 10 synthetic datasets, each with 150 observations, 2 classes and 100 attributes. For each discrete dataset, we chose the first two principal components to capture the most of the variation, and generated two clusters with 50 observations and 100 observations respectively. We sampled the two cluster means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ as follows:

$$\boldsymbol{\mu}_1 \quad \sim \mathcal{N}((0,0),\mathbf{I}) \tag{3.5}$$

$$\boldsymbol{\mu}_2 \quad \sim \mathcal{N}((20,20),\mathbf{I}) \tag{3.6}$$

We then generated the latent variables $\mathbf{Z}$ by sampling from the following Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1,\mathbf{I})$ and $\mathcal{N}(\boldsymbol{\mu}_2,\mathbf{I})$. After that we generated the loading matrices $\mathbf{W}_1,\ldots,\mathbf{W}_R$ by sampling from the priors on the rows $\mathcal{N}(\mathbf{0},\mathbf{I})$, and the offset terms $\mathbf{w}_{0r},\ldots,\mathbf{w}_{0R}$ by sampling from their prior $\mathcal{N}(\mathbf{0},\mathbf{I})$. We generated the dataset $\mathbf{X}$ by sampling from a multinomial distribution with parameters derived from the softmax transformation of the linear mapping of the latent variables onto a higher

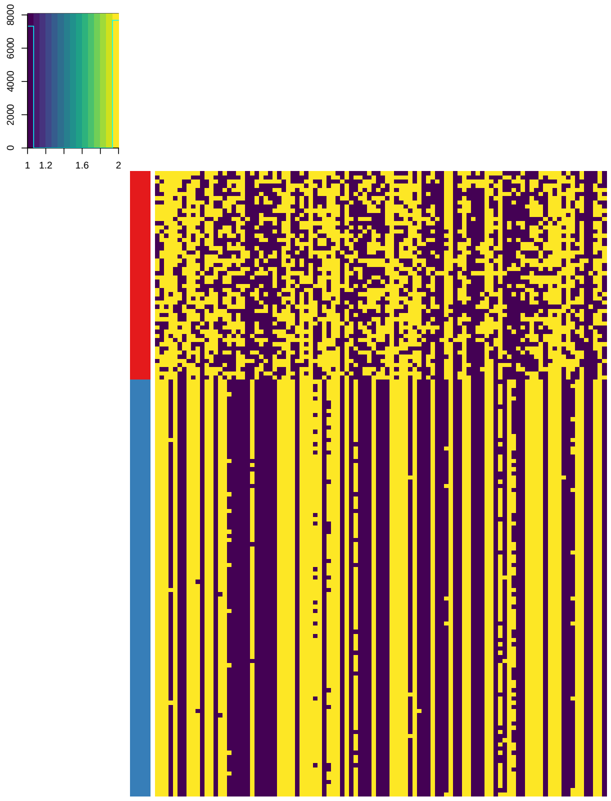


Figure 3.5: An example of a synthetic discrete dataset, generated using the BayesCluster generative model. The observations are on the y-axis sorted by cluster membership and the features are on the x-axis.

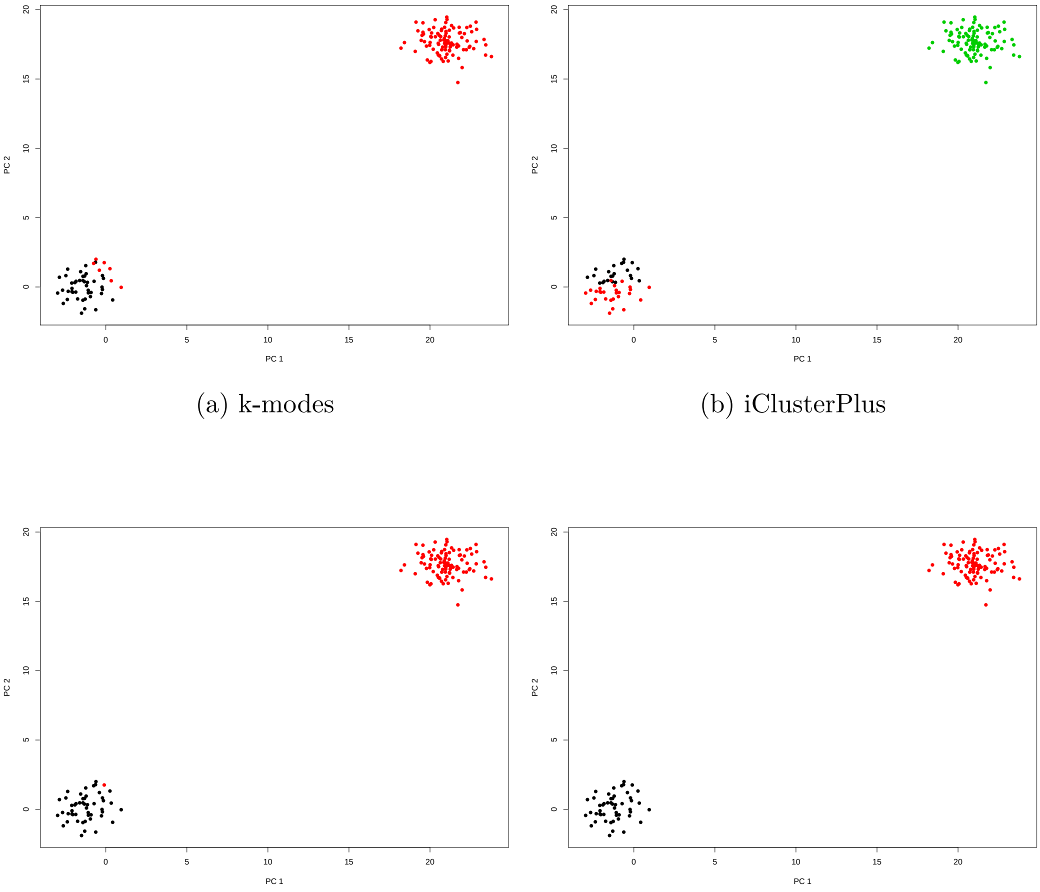
dimensional space:

$$p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}_{1:R}, \mathbf{w}_{01:0R}) = \prod_{r=1}^{R} \mathrm{Cat}(x_{ir}|\mathcal{S}(\mathbf{W}_r\mathbf{z}_i + \mathbf{w}_{0r})). \tag{3.7}$$

Figure 3.5 presents an example of a discrete dataset, generated using BayesCluster as a generative model.

The results, presented in Table 3.6, show that both latent variable models performed better than k-modes. The worse performance of k-modes was because it identified 6 clusters in one of datasets. BayesCluster was able to estimate both the latent dimensionality ($P = 2$) and the number of clusters ($K = 2$) accurately, and when it failed to do so, it was usually because it created a small third cluster. iClusterPlus estimated the latent dimensionality correctly but found 3 clusters in the datasets due to the procedure it uses to determine the number of clusters (it sets $K = P + 1$).

Figure 3.6 shows the final partitions for the synthetic dataset on Figure 3.5 obtained with k-modes, iClusterPlus and BayesCluster. Both k-modes and BayesCluster have identified correctly the number of clusters, though they have not allocated all the observations from the black cluster to the correct group.



(a) k-modes

(b) iClusterPlus

(c) BayesCluster

(d) ground truth

Figure 3.6: Comparison between the clustering partitions of the synthetic dataset presented on Figure 3.5) with k-modes clustering, iClusterPlus and BayesCluster, and the ground truth. The latent variables $\mathbf{Z}$ which have been used for the generation of the data are plotted and coloured according to cluster membership.

Model	mean ARI (± st. error)	Est. K (prop.)	Est. P (prop.)	p-value	Comp. time
k-modes	0.875 (0.78,0.96)	2, 3, 6 (0.5,0.3,0.2)	-	0.0002	27.21 s
iClusterPlus	0.789 (0.61,0.97)	3 (1.00)	3 (1.00)	0.0162	1.36 min
BayesCluster	0.948 (0.91,0.98)	2, 3 (0.5, 0.5)	2, 3, 4 (0.7,0.1,0.2)	-	18.98 min

Table 3.6: Comparison of the results on the synthetic data in terms of adjusted Rand index (ARI) averaged over the 10 datasets, ± 1 st. error ARI, estimates of the number of clusters K, estimates of the latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods, and computation time per run. The proportion of times a certain value for the number of clusters K or for the number of principal components P is estimated is put into brackets after the value.

3.2.2 Real datasets

We compared the accuracy of the final partitions in terms of mean (and ± standard error) adjusted Rand Index (ARI), estimated number of clusters K, estimated latent dimensionality P and computation time. We summarise the results in Tables 3.7 and 3.8 below. In all of the experiments, BayesCluster found the partition with the highest adjusted Rand Index.

Spect heart

Interestingly, all models performed poorly on the Spect heart dataset which might be due to the preprocessing involved to obtain binary features. The observations were first obtained from database of Spect image sets, which were then processed to extract 44 continuous features which summarise the images. These continuous patterns were further processed to get 22 binary features. BayesCluster was the only method that managed to identify correctly the true number of clusters ($K = 2$).

Model	mean ARI (± st.error)	Estimated K	Estimated P	p-value	Comp. time
k-modes	-0.025 (-0.048,-0.002)	4	-	0.0293	4.28 s
iClusterPlus	0.032 (0.013,0.051)	3	2	0.1762	47.15 s
BayesCluster	0.019 (-0.004,0.044)	2	4 or 5	-	23.57 min

Table 3.7: Comparison of the results on the Spect heart dataset in terms of mean adjusted Rand index (ARI), ± st. error ARI, estimates of the number of clusters K, estimates of the latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run.

Congressional voting dataset

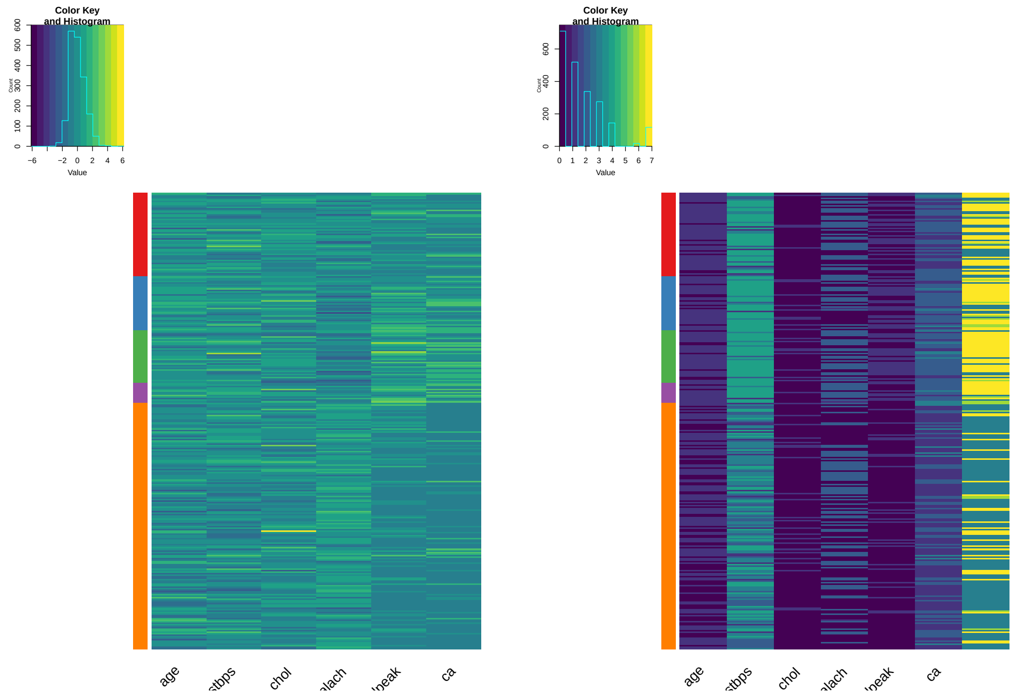
In the voting dataset, BayesCluster identified the correct number of clusters ($K = 2$) and found the partition with the highest mean ARI of 0.543. K-modes also identified the correct number of parties and produced similar results to BayesCluster, while iClusterPlus overestimated it and found 4 parties.

Model	mean ARI (± st. error)	Estimated K	Estimated P	p-value	Comp. time
k-modes	0.523 (0.505,0.541)	2	-	0.04769	3.34 s
iClusterPlus	0.268 (0.202,0.334)	4	3	6.187e-08	3.06 min
BayesCluster	0.543 (0.519,0.567)	2	2 or 5	-	30.15 min

Table 3.8: Comparison of the results on the voting dataset in terms of mean adjusted Rand index (ARI), ± st. error ARI, estimates of the number of clusters K, estimates of the latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run.

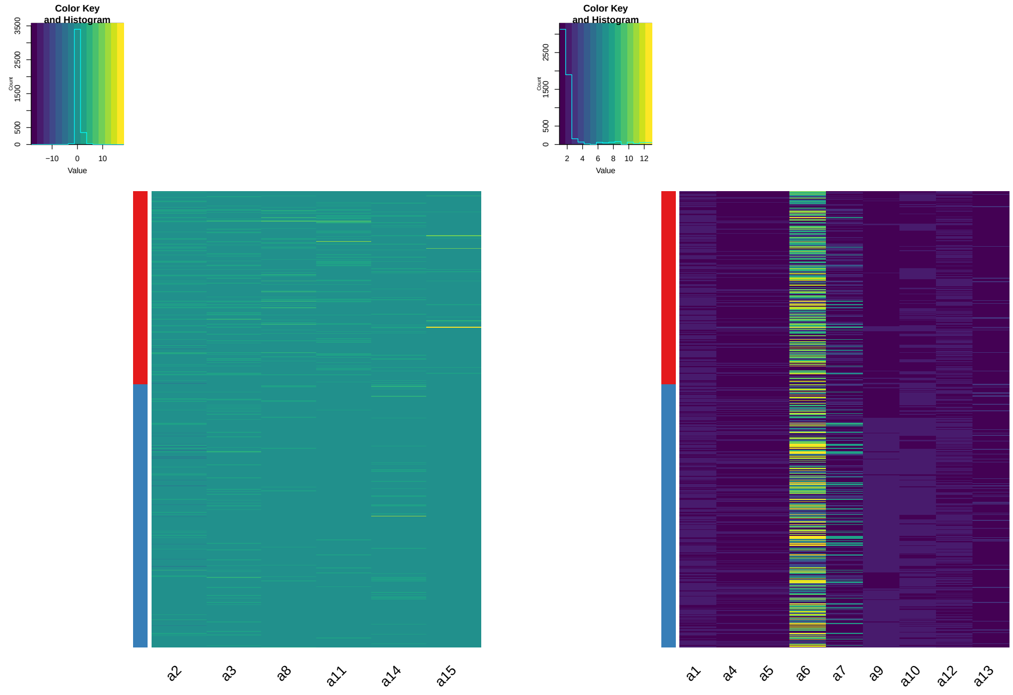
3.3 Mixed data

We compared mixed BayesCluster with k-prototypes and iClusterPlus over 3 datasets (2 real and 1 synthetic). We selected real datasets from the UCI machine learning repository which are well studied and for which the ground truth is known. We investigate the performance of BayesCluster on high-dimensional datasets with synthetic datasets. Since iClusterPlus cannot be used to cluster mixed data, we treat the mixed datasets as a result of the integration of a continuous and a discrete dataset, similarly to BayesCluster. The two real datasets we used are a credit approval dataset (690 observations, 2 classes, 15 attributes) and a heart disease data (270 observations, 5 classes, 13 attributes) available on the UCI machine learning repository `https://archive.ics.uci.edu/ml/datasets/`. The heatmaps of the datasets are presented on Figure 3.7. We have split the datasets into continuous and discrete parts to distinguish between the different type features.



(a) heart disease (continuous)

(b) heart disease (discrete)



(c) credit approval (continuous)

(d) credit approval (discrete)

Figure 3.7: Heatmaps of the mixed datasets. The observations are on the y-axis and are ordered according to the ground truth cluster membership, and the features are on the x-axis.

We used the following R packages in the experiments:

- 'clustMixType' [Szepannek, 2018] for the implementation of k-prototypes clustering;
- 'iClusterPlus' [Mo and Shen, 2016] for the implementation iClusterPlus.

We implemented k-prototypes with the following parameters: we ran the algorithm for number of clusters between 1 and 20, set the number of maximum iterations to 100 and did not use the parameter α, which describes the trade off between the Euclidean distance between numeric variables and a matching coefficient metric between categorical variables. We used the within-cluster sum of squares metric and the elbow method to determine the number of clusters K (Figure C.8 in Appendix C). We implemented iClusterPlus with the default options in the package - we set the number of MCMC burn-in steps to 100, the total number of MCMC draws to 200, the number of maximum iterations for the EM algorithm to 20 and the threshold for convergence set to 1e-4. We ran iClusterPlus for the number of latent dimensions P between 1 and 10 and selected the final P using the deviance ratio metric (Figure

C.9 in Appendix C). We ran BayesCluster with 5 random initialisations for 1000 iterations and for number of latent dimensions P between 2 and 10.We set the threshold for convergence to 1e-4 and initialised the model parameters as outlined in Chapter 2. We chose the final number of latent dimensions using BIC (Figure C.10 in Appendix C).

3.3.1 Synthetic data

We used the generative model of mixed BayesCluster to create 10 synthetic mixed datasets. For each dataset, we chose the first two principal components to capture the most of the variation, and generated three clusters with 50 observations each with 100 continuous and 100 discrete features (each with 2 categories), with the assumption that the continuous observations were normalised and have zero mean and unit variance. We sampled the three cluster means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$ as follows:

$$\boldsymbol{\mu}_1 \sim \mathcal{N}((0,0), \mathbf{I}) \tag{3.8}$$

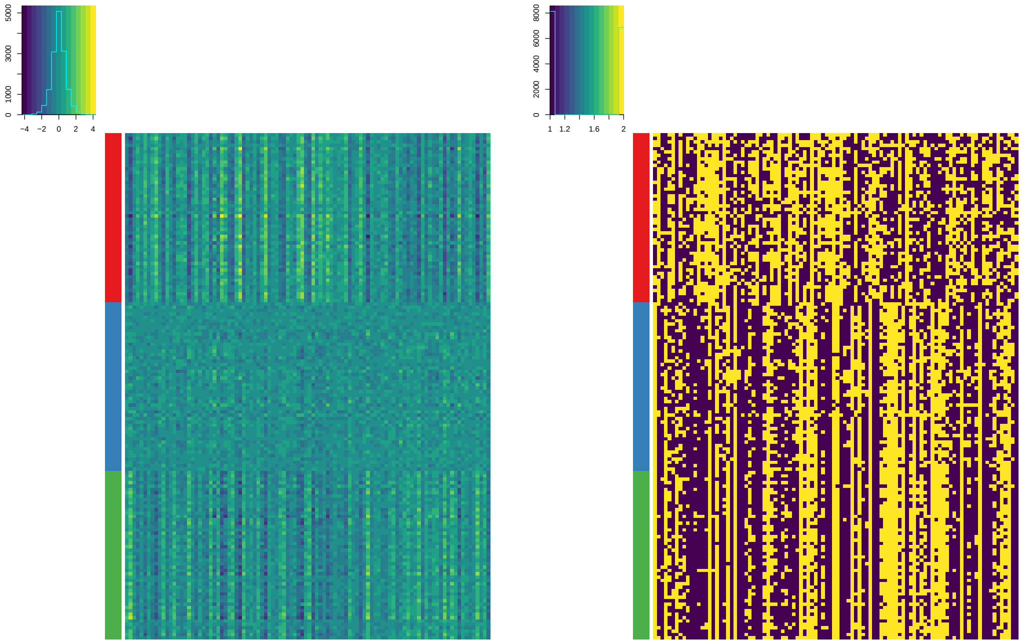
$$\boldsymbol{\mu}_2 \sim \mathcal{N}((2,2), \mathbf{I}) \tag{3.9}$$

$$\boldsymbol{\mu}_3 \sim \mathcal{N}((4,4), \mathbf{I}). \tag{3.10}$$

We then generated the latent variables $\mathbf{Z}$ by sampling from the following Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}), \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{I})$ and $\mathcal{N}(\boldsymbol{\mu}_3, \mathbf{I})$. After that, we generated the loadings matrices $\mathbf{W}^C$ and $\mathbf{W}_1^D, \dots, \mathbf{W}_R^D$ for the continuous and discrete variables, respectively, by sampling from the priors on the rows $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the offset terms $\mathbf{w}_{01}^D, \dots, \mathbf{w}_{0R}^D$ from their prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the error term ϵ from sampling from its prior $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ where $\sigma \sim \text{IG}(1,1)$. We finally generated the dataset $\mathbf{X}$ using

$$p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}^C, \mathbf{W}^D, \mathbf{w}_0^D, \epsilon) = \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}) \prod_{r=1}^{R} \text{Cat}(x_{ir}|\mathcal{S}(\mathbf{W}_r^\intercal \mathbf{z}_i + \mathbf{w}_{0r}^D)) \tag{3.11}$$

Figure 3.8 presents an example of a mixed dataset (split into continuous and discrete), generated using BayesCluster as a generative model.

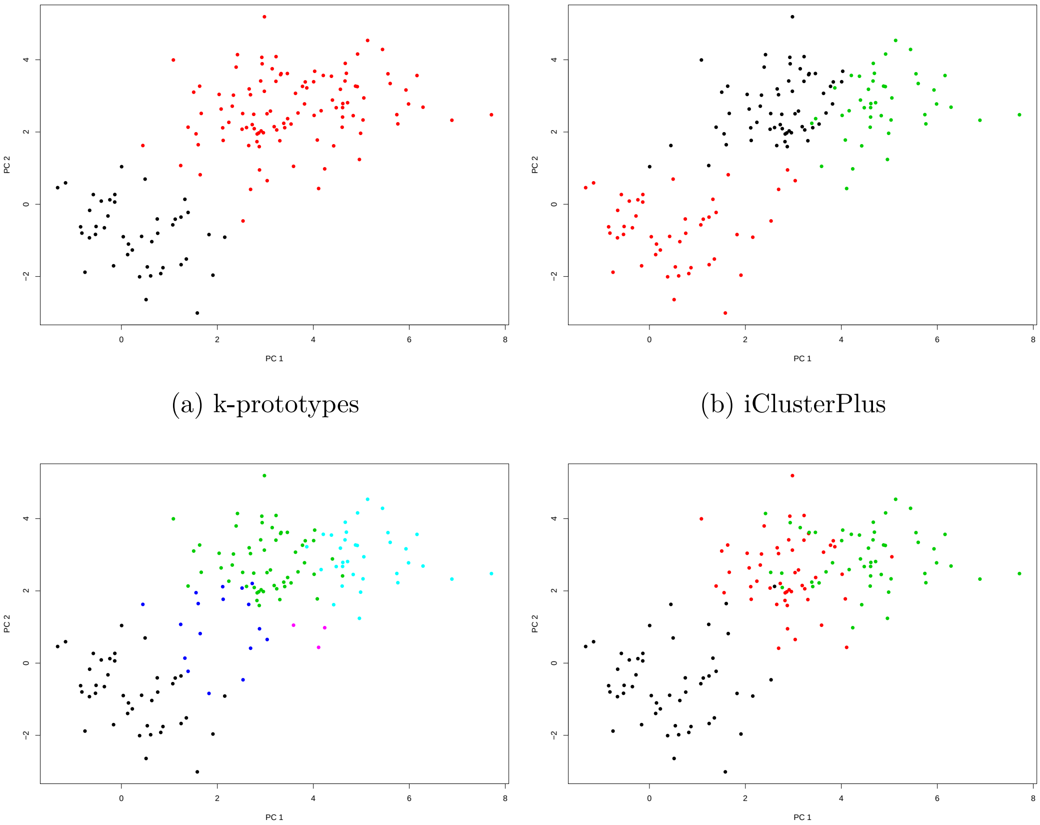


(a) continuous

(b) discrete

Figure 3.8: An example of a synthetic mixed dataset, generated using the BayesCluster generative model, and split into continuous and discrete subsets. The observations are on the y-axis and the features are on the x-axis. The heatmap for the discrete dataset shows that it is hard to distinguish between two of the clusters, which could be due to the softmax transformation used to generate the data.

With the synthetic datasets, we tested how well each of the methods could identify overlapping clusters. Figure 3.9 shows the results for the synthetic dataset presented on Figure 3.8. iClusterPlus was the only model to correctly identify the number of components ($K = 3$) but it did not model well the overlap between the green and red clusters. k-prototypes merged the overlapping clusters, whereas BayesCluster modelled the overlap by introducing new clusters.



(a) k-prototypes

(b) iClusterPlus

(c) BayesCluster

(d) ground truth

Figure 3.9: Comparison between the clustering partitions of the synthetic dataset with k-prototypes clustering, iClusterPlus and BayesCluster, and the ground truth. The latent variables $\mathbf{Z}$ which have been used for the generation of the data are plotted and coloured according to cluster membership.

Both latent variable models outperformed k-prototypes in regards with adjusted Rand index (Table 3.9). iClusterPlus managed to estimate the latent dimensionality ($P = 2$) and number of clusters ($K = 3$) correctly in all datasets which is due to the assumptions we made when creating the datasets ($P = 2$ and $K = 3$) and to the method iClusterPlus uses to find the number of clusters ($K = P + 1$).

Model	mean ARI (± st. error)	Est. K (prop.)	Est. P (prop.)	p-value	Comp. time
k-prototypes	0.43 (0.43,0.43)	2, 3 (0.9 , 0.1)	-	3.99e-08	3.29 min
iClusterPlus	0.533 (0.504,0.561)	3 (1.00)	2 (1.00)	0.05418	2.25 min
BayesCluster	0.558 (0.538,0.578)	2 - 5 (0.1, 0.1)	2, 3 (0.5, 0.5)	-	1hr 2min

Table 3.9: Comparison of the results on the mixed data in terms of adjusted Rand index (ARI) averaged over the 10 datasets, ± st. error ARI, estimates of the number of clusters K, estimates of the latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run. The proportion of times a certain value for the number of clusters K or for the number of principal components P is estimated is put into brackets after the value.

3.3.2 Real datasets

We compared the accuracy of the final partitions in terms of mean (± standard error) adjusted Rand index (ARI), estimated number of clusters K, estimated latent dimensionality P and computation time. The results from all experiments are summarised in Tables 3.10 and 3.11 below.

Heart disease dataset

All methods performed poorly on the heart disease dataset - BayesCluster and iClusterPlus tended to underestimate the true number of clusters ($K = 5$), whereas k-prototypes overstimated them. Combining the four classes which indicate the presence of heart disease into one, and trying to cluster the patients into either disease-free or disease-present might lead to more accurate results.

Model	mean ARI	Estimated	Estimated	p-value	Comp.
	(± st. error)	K	P		time
k-prototypes	0.1 (0.068,0.131)	6 or 7	-	3.26e-08	6.41 min
iClusterPlus	0.12 (0.065,0.175)	3	2	0.0001	47.65 s
BayesCluster	0.219 (0.194,0.245)	2	2	-	29.87 s

Table 3.10: Comparison of the results on the heart disease dataset in terms of mean adjusted Rand index (ARI), ± st. error ARI, estimates of the number of clusters K, estimates of the latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run.

Credit approval dataset

In the case of the credit approval data, both iClusterPlus and BayesCluster had low adjusted Rand indices, which might be due to the models' assumptions. Although BayesCluster identified the correct number of clusters ($K = 2$), it allocated a large proportion of the 'non-approved' observations to the 'approved' cluster. Looking at the continuous part of the credit approval dataset (Figure (c) 3.7), we notice that the observations have similar values across the six continuous features. In addition, the features included in this dataset have been anonymised and we can not assess their relevance to the clustering task. Hence, the outputs of BayesCluster and iClusterPlus could have been affected by uninformative data.

Model	mean ARI (± st. error)	Estimated K	Estimated P	p-value	Comp. time
k-prototypes	0.336 (0.307,0.365)	3	-	1.65e-11	8.06 min
iClusterPlus	0.082 (0.049,0.114)	4	3	0.0164	2.32 min
BayesCluster	0.055 (0.049,0.062)	2	2	-	43.06 min

Table 3.11: Comparison of the results on the credit approval dataset in terms of average adjusted Rand index (ARI), ± st. error ARI, estimates of number of clusters K, estimates of latent dimensionality P where applicable, p-values from a t-test testing the hypothesis that there is no difference between the results from BayesCluster and from the comparison methods in terms of ARI, and computation time per run.

3.4 Discussion

We have presented the application of BayesCluster to both synthetic and real datasets and shown that the model provides competitive clusterings of real-world data as measured by the adjusted Rand index with respect to known labels.

There were cases in the experiments with continuous data where BayesCluster failed to identify clusters close to the ground truth, for example in the case of the iris dataset, it merged two of the clusters into one. Some of the reasons for that could be that this was a particularly difficult clustering problem or that some of the model assumptions are not appropriate.

In the experiments with discrete datasets, BayesCluster provided competitive results in comparison with k-modes and iClusterPlus. The only dataset on which BayesCluster did not perform well was the Spect heart dataset, which could be because of uninformative features result of the preprocessing performed on the original Spect data.

The experiments with mixed data, in particular with the credit approval data, highlighted the importance of data quality and informative/relevant features to the quality of the clustering results.

In terms of computation time, BayesCluster is slower than all comparison methods, which is mainly due to not making use of the fast C++ libraries available in R, which are used in the other methods. We expect that further work on the code base could substantially close this gap in the run time.

In this Chapter we applied BayesCluster to datasets from a range of different applications - ecology, medicine, politics. The results from the numerical experiments show that BayesCluster can provide competitive clustering results regardless of the context of the data. There were cases where a simpler model such as k-means outperformed BayesCluster, which could be due to the model assumptions and inference scheme we use. There are different ways to counteract the shortcomings of BayesCluster, which will explore in Chapter 4. We will also present extensions to BayesCluster which lead to the identification of more interpretable and well-defined clusters. We will investigate the application of BayesCluster to genomic data in Chapter 5.

Chapter 4

Extensions to BayesCluster

In Chapter 2 we presented the theory behind BayesCluster, which can be used to identify clusters in mixed datasets. In building the model, we made use of the flexibility of Dirichlet process mixture models and of the efficiency of simulated annealing. The numerical experiments in Chapter 3 demonstrated well the advantages of BayesCluster over other clustering methods when applied to synthetic and real data.

However, the methods for clustering and inference we use have inherent drawbacks that can affect the quality of the model output. For example, Dirichlet process mixture models have been shown to often overestimate the number of true clusters [West and Escobar, 1993; Onogi et al., 2011; Miller and Dunson, 2018], and simulated annealing can result in non-optimal solution. In addition, often, when dealing with noisy genomic and clinical data, which will be the case in Chapter 5, we wish to prioritise clearly-separated (and hence biologically distinctive) clusters.

In this chapter, we address these issues and present extensions to BayesCluster based on the ideas of *non-local priors, split-merge* and *cluster-size priors*, which lead to stronger model parsimony and to the identification of more interpretable clusters.

4.1 Non-local priors extension to BayesCluster

Non-local priors [Johnson and Rossell, 2012; Fúquene et al., 2016; Rossell and Telesca, 2017] encourage parsimony by enforcing separation between the clusters under consideration.

4.1.1 Overview of non-local priors

Fúquene et al. [2016] introduce non-local priors in the context of mixture models. They consider data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ arising from density

$$p(\mathbf{x}_i | \boldsymbol{\vartheta}_k, \mathcal{M}_k) = \sum_{j=1}^{k} \pi_j p(\mathbf{x}_i | \theta_j), \tag{4.1}$$

where $\mathcal{M}_k$ is a model with k components and parameters $\boldsymbol{\vartheta}_k$, which include the mixture proportions π_j and the component parameters θ_j for $j = 1, \ldots, k$, and present a statistical framework for selecting the number of components k. If we have multiple models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ and for example, $\mathcal{M}_1$ is nested within $\mathcal{M}_2$, these models are not very well separated. If $\mathbf{x}$ was truly generated from $\mathcal{M}_1$, then $\mathcal{M}_1$ will receive high marginal likelihood. However, the marginal likelihood for $\mathcal{M}_2$ will also be relatively large since $\mathcal{M}_1$ is contained in $\mathcal{M}_2$. If we perform Bayesian model selection via posterior model probabilities, then $\mathcal{M}_1$ will be eventually favoured as the sample size n grows towards infinity since Bayesian model selection automatically incorporates Occam's razor.

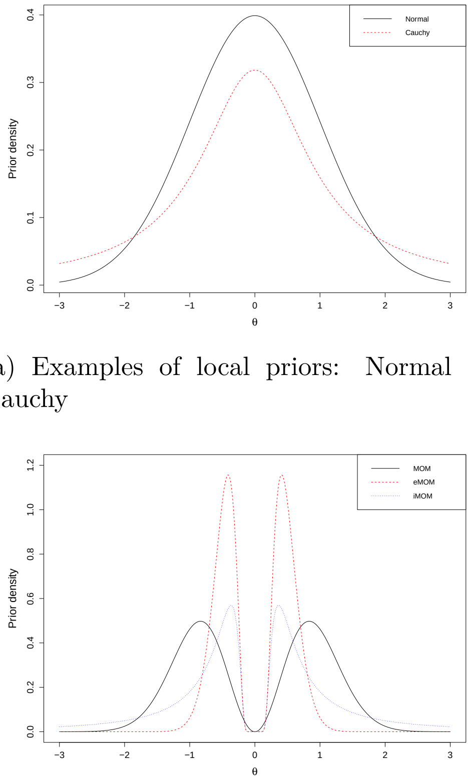
To address the problem of weakly separable models, Fúquene et al. [2016] build upon the idea of repulsive mixtures [Petralia et al., 2012] and avoid the limitations of shrinkage priors such as inference sensitive to value of the concentration parameter α or the number of components k, and lack of posterior model probabilities. They introduce non-local priors, formally defined as follows:

Definition 1. Let $\mathcal{M}_k$ be a mixture with k components as in (4.1). A continuous prior density $p(\boldsymbol{\vartheta}_k | \mathcal{M}_k)$ is a **non-local prior** if and only if

$$\lim_{\boldsymbol{\vartheta}_k \to t} p(\boldsymbol{\vartheta}_k | \mathcal{M}_k) = 0, \tag{4.2}$$

for any $t \in \Theta_k$ such that $p(\mathbf{x} | \mathbf{t}, \mathcal{M}_k) = p(\mathbf{x} | \boldsymbol{\vartheta}_{k'}, \mathcal{M}_{k'})$, for some $\mathcal{M}'_k$ with k' components as in (4.1) and $\boldsymbol{\vartheta}_{k'} \in \Theta_{k'}, k' < k$.

This means that a non-local prior under $\mathcal{M}_k$ assigns vanishing density to any $\boldsymbol{\vartheta}_k$ such that (4.1) is equivalent to a mixture with $k' < k$ components. A **local prior** is defined as any prior $p(\boldsymbol{\vartheta}_k | \mathcal{M}_k)$ not satisfying (4.2), and examples of local priors include the normal and Cauchy distributions. Figure 4.1 presents a comparison between local priors and non-local priors in the context of hypothesis testing.



(a) Examples of local priors: Normal and Cauchy

(b) Examples of non-local priors: moment (MOM), exponential moment (eMOM), inverse moment (iMOM)

Figure 4.1: Comparison between local and non-local priors

We can reformulate **Definition 1.** to simplify checking whether a prior is non-local as follows: $p(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ defines a non-local prior if and only if $\lim p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = 0$ as either

1. $\pi_j \to 0$ for any $j = 1, \ldots, k$;
2. $\boldsymbol{\theta}_i \to \boldsymbol{\theta}_j$ for any $i \neq j$.

We can easily construct a non-local prior from an arbitrary local prior in the following way:

$$p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = d_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}_k)p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k), \tag{4.3}$$

where $d(\boldsymbol{\vartheta}_k)$ is a continuous penalty function converging to 0 under condition 1. or 2., and $p^L(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ is an arbitrary local prior such that $p(\boldsymbol{\vartheta}_k|\mathcal{M}_k)$ is a proper prior.

We can express (4.3) as

$$p(\boldsymbol{\vartheta}_k|\mathcal{M}_k) = d_{\boldsymbol{\theta}}(\boldsymbol{\theta})p^L(\boldsymbol{\theta}|\mathcal{M}_k)\text{Dir}(\boldsymbol{\pi}|\alpha), \tag{4.4}$$

where

$$d_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{C_k}\left(\prod_{1\leq i<j\leq k} d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\right), \tag{4.5}$$

with $C_k = \int \left(\prod_{1\leq i<j\leq k} d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\right) p^L(\boldsymbol{\theta}|\mathcal{M}_k)\mathrm{d}\boldsymbol{\theta}$ being the normalising constant. The form of $d_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ depends on the model under consideration and some of the most commonly used penalties are:

- *moment* (MOM) [Johnson and Rossell, 2010] $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\intercal \mathbf{A}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)/g$;
- *exponential moment* (eMOM) [Rossell et al., 2013] $d(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \exp\{-g/(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\intercal \mathbf{A}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)\}$,

where $\mathbf{A}$ is a symmetric positive matrix and g is a dispersion parameter similar to the parameter used in the repulsive mixtures introduced by Petralia et al. [2012].

We choose to use the MOM prior in BayesCluster as the normalising constant C_k can be computed in closed form in specific model settings and it has been shown empirically that both MOM and eMOM result in strong model separation [Johnson and Rossell, 2012, 2010].

Fúquene et al. [2016] have shown that non-local priors induce extra parsimony via the penalty term $d_{\boldsymbol{\vartheta}}(\boldsymbol{\vartheta}_k)$ and provide posterior consistency. An approximation to the marginal likelihood $p(\mathbf{x}|\mathcal{M}_k)$ can be derived as well but as we do not require the computation of marginal likelihood, we omit the derivation of its approximation here. Proofs of the parsimony and posterior consistency properties are omitted but the reader can refer to Fúquene et al. [2016] if interested.

We adapt the Moment non-local prior for BayesCluster. Following Fúquene et al. [2016], we place a MOM prior on the cluster means $\boldsymbol{\mu}_k$:

$$p(\boldsymbol{\mu}_k) = \frac{1}{C_k}\prod_{1\leq i<j\leq k}\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\intercal \mathbf{A}_\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{g}\mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, g\mathbf{A}_\Sigma) \tag{4.6}$$

We set $\mathbf{A}_\Sigma$ to be the identity matrix. We set up g so that there is small prior probability that any 2 components are poorly separated and give rise to unimodality. Ray and Lindsay [2005] point out that although the number of modes in normal mixtures depends on nontrivial parameter combinations, in the case when $\pi_1 = \pi_2 = 0.5$ and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ the mixture is bimodal if and only if $\kappa = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal \mathbf{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 4$. Hence, we need to set the dispersion parameter g so that $p(\kappa < 4|\mathcal{M}_k) = 0.05$ or 0.1 (we use 0.05 based on the numerical results provided in Fúquene et al. [2016]). As (4.6) implies a Gamma prior on κ $\text{Ga}(\kappa|p/2+1, 1/4g)$ where p is the dimension of the data, setting g amounts to numerically solving an integral. We present in the table below (Table 4.1) the values of g for a range of number of dimensions p:

p	g
1	5.68
2	2.81
3	1.745
4	1.225
5	0.922
6	0.731
7	0.60
8	0.505
9	0.437
10	0.384

Table 4.1: Values of the dispersion parameter g for number of dimensions between 1 and 10.

Using non-local priors in the BayesCluster model does not require the computation of the normalising constant C_k, which is quite expensive, as it is not involved in the allocations of the observations or in the Metropolis Hastings/simulated annealing updates. The only differences from the models presented in Chapter 2, are in computation of the log posterior of the model where we have an extra term for the penalty, and in the calculation of the probability of starting a new cluster. Because of the penalty term, we cannot integrate the base measure but we can still find the probability of starting a new cluster using the ‘no gaps’ algorithm proposed by MacEachern and Müller [1998] and adapted by Neal [2000]. We compute the probability of starting a new cluster by introducing an auxiliary variable c^*, representing the new state, and we sample a cluster mean $\boldsymbol{\mu}^*$ for the new (temporary) cluster from the base measure. After that we compute the conditional probability using:

$$p(c_i = c^* | c_{-i}, \mathbf{Z}) \propto \frac{\alpha}{\alpha + N - 1} \mathcal{N}(\mathbf{z}_i | \boldsymbol{\mu}^*, \mathbf{I}). \tag{4.7}$$

Similarly to the models from Chapter 2, we then assign the i^{th} item to the most probable cluster. We use the same proposal distributions for the other model variables as outlined in Chapter 2.

We illustrate how using non-local priors leads to better cluster separability using some of the real datasets from Chapter 3. We picked the iris dataset, since BayesCluster was not able to identify the true number of clusters in this case and merged two of the clusters; the Wisconsin breast cancer dataset and the glass dataset as BayesCluster did not perform well on these datasets. We set the concentration parameter α to be equal to 3 for all the experiments as recommended by Fúquene et al. [2016].

Iris dataset

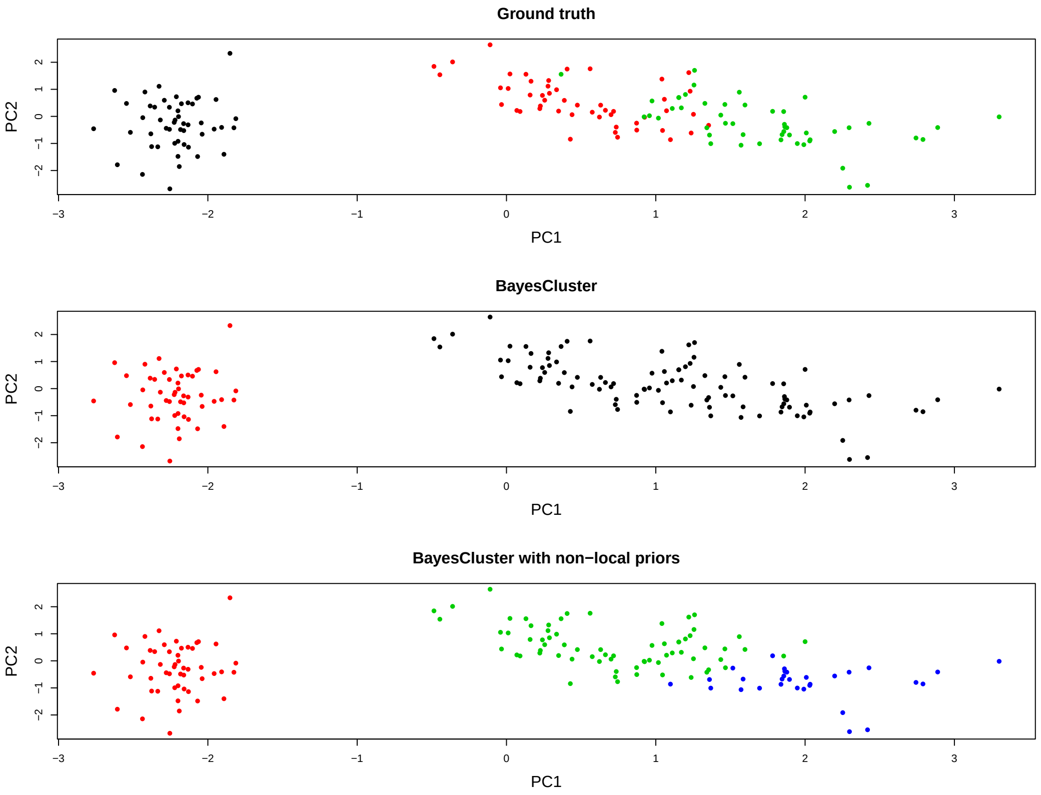
Using the non-local priors for the cluster means, BayesCluster was able to identify the true number of clusters ($K = 3$) in the iris dataset. Although it did not model well the overlap between the versicolor and virginica clusters, the clusters it identified were well separated (see Figure 4.2).

Model	mean ARI (± st.error)	Est. K	Est. P	p-value	Comp. time
BayesCluster	0.567 (0.565,0.569)	2	2	0.06362	29.87 sec
BayesCluster (non-local priors)	0.588 (0.557,0.62)	3	2	-	1.12 min

Table 4.2: Comparison of the results on iris data in terms of mean adjusted Rand index (ARI), ± standard error, estimates of the number of clusters K, estimates of the latent dimensionality P, p-value from a t-test testing the hypothesis that there is no difference between the results from BayesCluster with and without non-local priors in terms of ARI, and computation time per run.

Wisconsin breast cancer dataset

The non-local priors (NLP) extension of BayesCluster achieved better results in regards with adjusted Rand index (Table 4.3, mean ARI =0.762, p-value = 0.03328). However, it did not identify the correct number of cluster $K = 2$ and instead es-


Ground truth

BayesCluster

BayesCluster with non−local priors

Figure 4.2: Comparison between the final partitions of the iris data, identified by BayesCluster and BayesCluster with non-local priors. The first two principal components are plotted and coloured according to the cluster membership.

timated that there were 3 clusters. This could be because it decided to create a separate cluster to model the observations with similar characteristics (see Figure 4.3) or it is finding another subtype of breast cancer.

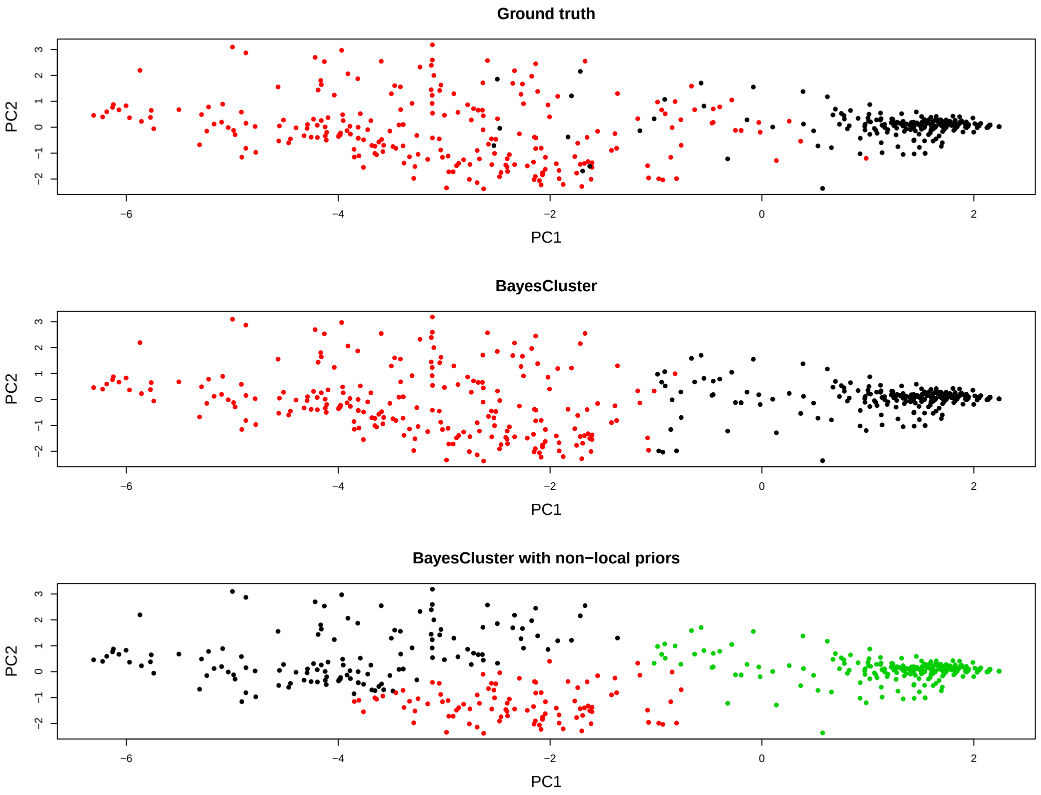


Figure 4.3: Comparison between the final partitions of the Wisconsin breast cancer data, identified by BayesCluster and BayesCluster with non-local priors. The first two principal components are plotted and coloured according to the cluster membership.

Model	mean ARI (± st. error)	Est. K	Est. P	p-value	Comp. time
BayesCluster	0.728 (0.699,0.757)	2	2	0.03328	21.45 min
BayesCluster (non-local priors)	0.762 (0.726,0.798)	3	2		27.95 min

Table 4.3: Comparison of the results on Wisconsin breast cancer data in terms of mean adjusted Rand index (ARI), ± standard error, estimates of the number of clusters K, estimates of the latent dimensionality P, p-value from a t-test testing the hypothesis that there is no difference between the results from BayesCluster with and without non-local priors in terms of ARI, and computation time per run.

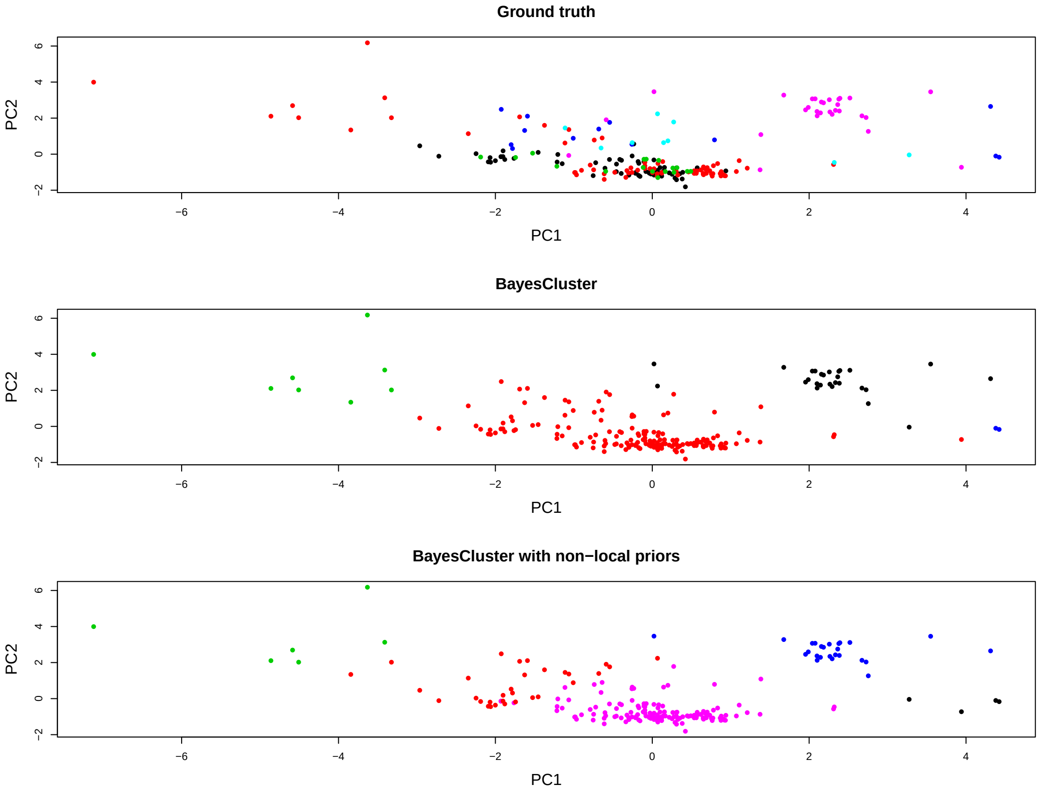


Figure 4.4: Comparison between the final partitions of the glass data, identified by BayesCluster and BayesCluster with non-local priors. The first two principal components are plotted and coloured according to the cluster membership.

Glass dataset

The non-local priors extension of BayesCluster achieved better results on the glass dataset in regards with adjusted Rand index (Table 4.4). Although it did not identify the correct number of clusters ($K = 7$), the NLP version estimated that there were 5 clusters and it got closer to the ground truth K. Both versions of BayesCluster modelled well the lilac cluster (Figure 4.4) but struggled to model the other clusters as a lot of them have similar characteristics (see heatmap on Figure 3.1b).

Model	mean ARI	Est.	Est.	p-value	Comp.
	(± st. error)	K	P		time
BayesCluster	0.182	4	3	0.05806	7.02 min
	(0.165,0.191)				
BayesCluster	0.195	5	2		9.61 min
	(0.179, 0.219)				
(non-local priors)					

Table 4.4: Comparison of the results on the glass data in terms of adjusted Rand index (ARI), mean ARI ± 1 standard error, estimates of the number of clusters K, estimates of the latent dimensionality P, p-value from a t-test testing the hypothesis that there is no difference between the results from BayesCluster with and without non-local priors in terms of ARI, and computation time per run.

4.2 Drawbacks of Dirichlet process mixture models and possible solutions

Although Dirichlet process mixture models have been shown to provide consistent estimates for the density [Ghosh and Ramamoorthi, 2003; Wu and Ghosal, 2010], this does not imply they give consistent estimates for the number of components: a good density estimate might include components with very small weights. This problem has been studied in detail by West and Escobar [1993] and Onogi et al. [2011], who have shown empirically that the posterior inference of the number of clusters tends to put its mass on a range of values greater or equal to the true number of clusters. Miller and Harrison [2013] show with the simple example of fitting a Dirichlet process mixture model to data generated from a single univariate Normal distribution, that the posterior probability of the number of clusters being equal to 1 does not converge to 1 almost surely, but decreases to 0 instead as the amount of data increases. The reason for this inconsistency is that Dirichlet process mixture models strongly prefer having some very small clusters and will introduce extra clusters even when they are not needed. Arratia et al. [2003] showed that as the number of observations $N \to \infty$, the expected number of clusters K is equal to

$$\mathbb{E}[K] = \sum_{i=1}^{N} \frac{\alpha}{N-1+\alpha} \approx \alpha \log N \tag{4.8}$$

and that the expected number K_M of clusters of size M is

$$\lim_{N \to \infty} \mathbb{E}[K_M] = \frac{\alpha}{M}, \tag{4.9}$$

which implies that in expectation, there will be a small number of large clusters, corresponding to the 'rich-get-richer' property of Dirichlet process, and a large number of small clusters.

Often these tiny clusters are dealt with by being removed, which results in an inaccurate model. Some researchers have put a prior on the number of components [McCullagh et al., 2008; Nobile and Fearnside, 2007; Green and Richardson, 2001; Nobile, 1996] but this becomes an unrealistic approach to take when working with real complex, high-dimensional datasets and not knowing the ground truth or having prior knowledge of the number of clusters. This might suggest that using a finite mixture model would be a more appropriate approach. However, Miller and Dunson [2018] show empirically that finite mixture models with unknown clusters exhibit a similar inconsistency. In addition, Cai et al. [2017] demonstrate this inconsistency of finite mixtures theoretically and highlight the importance of understanding under what conditions we can make robust inference and deal effectively with model misspecifications.

We take a different approach to solving this issue and propose an extension which deals with tiny clusters in a more principled way. This is desirable particularly in biological contexts such as those we are interested in because it is hard to interpret small clusters.We achieve this by putting a prior on the size of the clusters which discourages very small clusters.

4.2.1 Prior on the cluster size

The idea of controlling the cluster size with a prior or constraints is not new - it has been applied in both deterministic and model-based clustering over the last decade, for example in the constrained k-means clustering developed by Bradley et al. [2000].

There are different ways of controlling the size of the clusters: for example, Wallach et al. [2010] explore the properties of the uniform process introduced by Qin et al. [2003] and Jensen and Liu [2008], and show that using the uniform process instead of Dirichlet process or Pitman-Yor process as a prior leads to more uniformly-sized clusters. Bradley et al. [2000] transform the k-means clustering algorithm into a linear programming problem with constraints on the minimum cluster size. Banerjee and Ghosh [2006] study a variant of k-means which enforces equal size for all clusters, whereas Zhu et al. [2010] apply balancing constraints to k-means and make use of prior knowledge of the distribution of the data by trying to find a partition that satisfies the constraints. Another approach to controlling the cluster size involves the concept of *microclustering*, introduced by Miller et al. [2015]. In microclustering models the size of the clusters grows sublinearly with respect to the sample size and to ensure that, negative binomial [Miller et al., 2015] and uniform priors [Klami and Jitta, 2016] have been used for the cluster size. Jitta and Klami [2018] generalise this work to the use of wider range of priors on the cluster size. The model they propose for mixture-based clustering replaces the i.i.d. observations with i.i.d. clusters, and the joint density factorises as

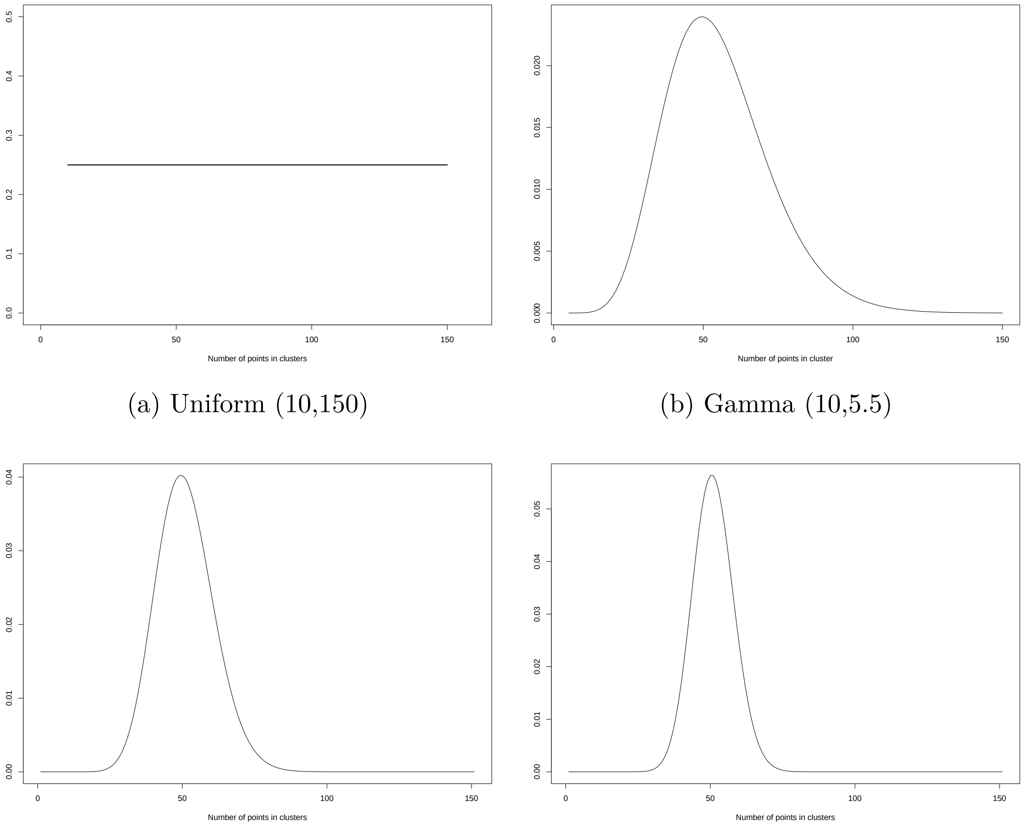
$$p(\mathbf{X}, C|\theta) = \prod_{k=1}^{K} \left(p(s_k) \prod_{n=1}^{N} p(\mathbf{X}_n|\theta_n)^{I[c_n=k]} \right), \tag{4.10}$$

where s_k is the number of samples in the kth cluster, $\mathbf{X}_n$ is the nth observation and c_n is the corresponding indicator variable. This formulation allows control over the cluster size but removes the ability to sample the cluster indicators individually - now they all have to be updated simultaneously.

We adapt the approach of Jitta and Klami [2018] but incorporate the cluster size prior as an implicit constraint, and include a term $\sum_{k=1}^{K} \log p(s_k)$ in the computation of the model log posterior.

We consider the following priors on the cluster size: Poisson, Gamma, negative binomial and uniform. Other possible choices for the prior include the delta distribution, and a mixture of Gaussian/Gamma distributions. We illustrate the effect of the different priors over the size of the clusters using three of the continuous synthetic datasets generated in Chapter 3.

We consider four models, where we have placed uniform ($\mathcal{U}(10, 150)$), Gamma ($\mathrm{Ga}(10, 5.5)$), negative binomial ($\mathrm{NBin}(50, 1/2)$) and Poisson ($\mathrm{Poi}(50)$) priors (see Figure 4.5 for illustration of the priors) on the size of the clusters. We have picked the parameters of the priors so that small clusters with fewer than 10 data points are heavily disfavoured.



(a) Uniform (10,150)

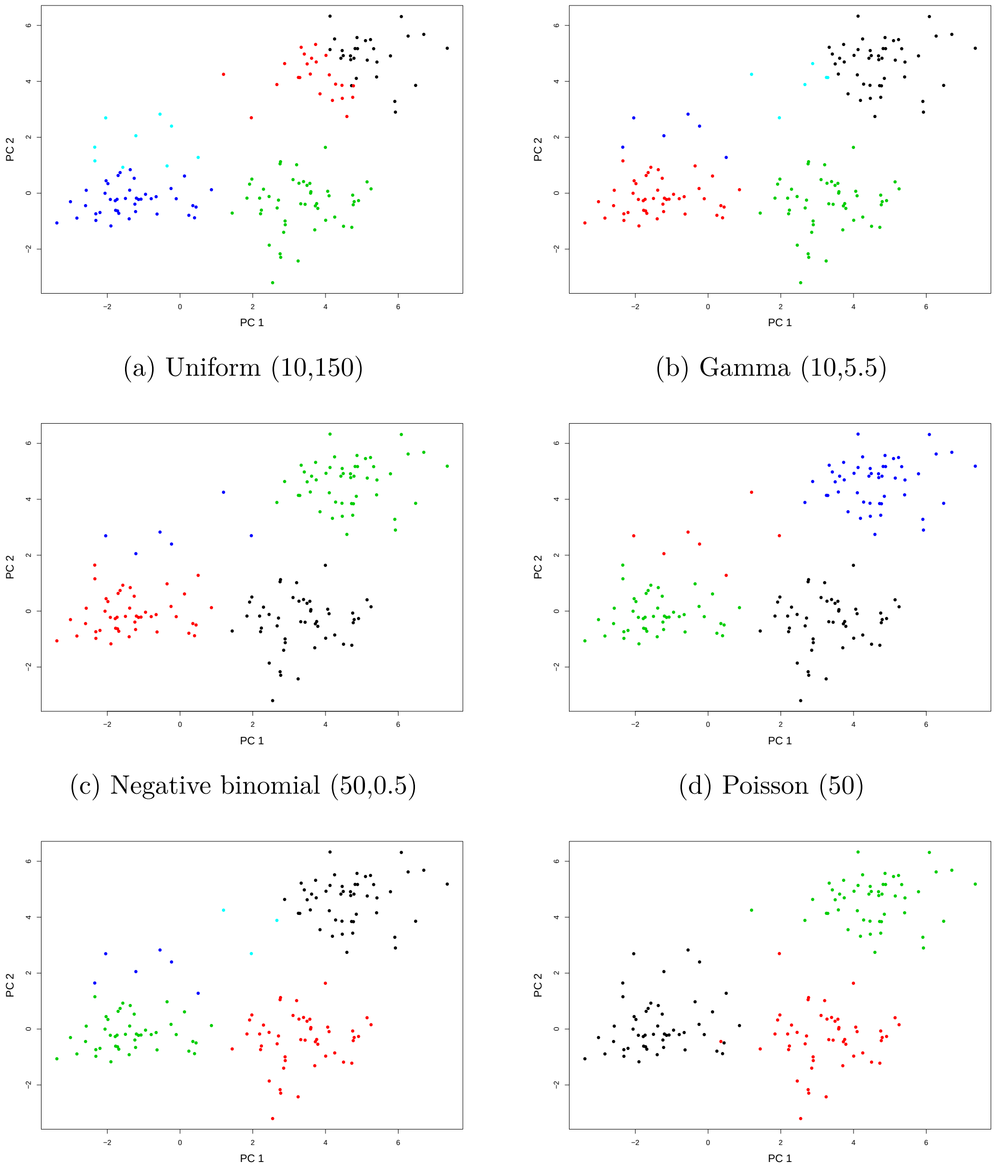
(b) Gamma (10,5.5)

(c) Negative binomial (50,0.5)

(d) Poisson (50)

Figure 4.5: Examples of priors on the cluster size we consider in this chapter.

We present a comparison between the different priors on a synthetic dataset, generated in the same manner as the continuous synthetic datasets in Chapter 3. We performed experiments only with synthetic data since we did not observe very small clusters in the experiments with real data in Chapter 3. Using continuous BayesCluster on the dataset, there were 5 clusters identified, with one of them (the light blue) having fewer than 10 data points in it. By setting a prior on the cluster size, we can see that the resulting partitions do not have any small clusters and include clusters of similar sizes (Figure 4.6).



(a) Uniform (10,150)

(b) Gamma (10,5.5)

(c) Negative binomial (50,0.5)

(d) Poisson (50)

(e) no prior on the cluster size

(f) ground truth

Figure 4.6: Comparison between the cluster partitions identified by BayesCluster when using different priors on the cluster size, the cluster partition obtained BayesCluster without any prior on the cluster size, and the ground truth. The use or the lack of prior on the cluster size is indicated in the figure captions.

We further compared the different cluster size priors in regards with adjusted Rand index and estimated number of clusters to pick a prior to work with later. All models apart from the one using the Negative binomial prior identified 5 clusters.

The models produced similar final partitions and performed similarly in regards with adjusted Rand index, with the model with negative binomial prior having the highest mean ARI of 0.884 (Table 4.5). As the model with negative binomial prior was the closest to the ground truth and modelled well the tiny clusters, it is a more preferable choice to use for prior on the cluster size.

prior	mean ARI (± st. error)	Estimated K	Estimated P	p-value
Uniform	0.836 (0.764,0.907)	5	2,3	0.8612
Gamma	0.863 (0.806,0.921)	4,5	2,4	0.6525
Negative binomial	0.884 (0.843,0.925)	4	2,3	0.3131
Poisson	0.844 (0.785,0.903)	5	2,3,4	

Table 4.5: Comparison of the results obtained using different cluster size priors with BayesCluster on a synthetic dataset, in terms of mean adjusted Rand index (ARI), ± standard error, estimates of the number of clusters K, estimates of the latent dimensionality P. The p-values are from a t-test testing the hypothesis that there is no difference between the results from BayesCluster with Poisson prior on the cluster size and with any of the other 3 cluster size priors in terms of ARI.

4.3 Split-merge MCMC

We investigated the idea of split-merge sampling because we often observed a tail of small clusters in our experiments which might be in part due to simulated annealing getting stuck in local modes. To address this issue, we adopted a method, inspired by split-merge MCMC [Jain and Neal, 2004; Green and Richardson, 2001; Wang and Blei, 2012; Hughes et al., 2012; Jain et al., 2007; Dahl, 2003; Wang and Dunson, 2011]. Split-merge MCMC methods have been motivated by some of the drawbacks of Gibbs sampling, which can become trapped in isolated modes and result in an inappropriate clustering of the data. Celeux et al. [2000] point out that this problem is due to the incremental nature of Gibbs sampling which is unable to simultaneously move a group of observations to a new component. In addition, the incremental updates are unlikely to move a single observation to a new cluster because such move has a low probability and is unlikely to be accepted [Celeux et al., 2000].

There have been developed many different split-merge MCMC methods. For example, Green and Richardson [2001] propose a complex split-merge in the reversible jump networks, which is based on conserving specific moment conditions, and is accepted or rejected by Metropolis-Hastings acceptance probability. This scheme, however, is not particularly practical when working with high-dimensional data as the computation of the probability of accepting/rejecting the proposed move becomes more complex.

Jain and Neal [2004] propose a simpler scheme, which is more suitable for high-dimensional data. They consider two different split-merge moves: simple random split and restricted Gibbs sampling. In the random split, two random points are considered, and if they are in the same cluster k, they are split into two clusters i and j, and all points from the k^{th} cluster are added to either the i^{th} cluster or the j^{th} cluster, which is accepted with a Metropolis-Hastings probability. If the points are in different clusters, then all the points from the two clusters are merged in one and this move is accepted with a Metropolis-Hastings probability. This split-merge move is unlikely to be often accepted as it does not take into account the cluster information. In the restricted Gibbs sampling move, the points after the split are assigned to a new cluster in a deterministic manner and then a restricted Gibbs sampling is performed on the new cluster.

We propose a different approach: we split the clusters containing fewer than 10 data points into singletons, and then consider a merge where we add all points from the tiny clusters to the most likely large cluster, and accept the merge by using a simulated annealing acceptance probability. Note that the detailed balance equations do not need to be satisfied since we apply the split-merge approach in the context of simulated annealing. Modifying the inference scheme with split-merge helps guard against the inference getting stuck in local modes. It does not change the posterior, and thus, does not change the tendency of the Dirichlet process to want a tail of small clusters. Hence, we need to use both split-merge and cluster size prior to counteract the inherent drawbacks of the model and the inference scheme. We summarise the steps involved in the algorithm, as applied in the case of continuous

BayesCluster, below:

Algorithm 4.1: Continuous BayesCluster with split-merge

```
Perform PPCA to initialise the latent variables $\mathbf{Z}$, sample the loadings matrices $\mathbf{W}$ and residual error $\epsilon$ from the corresponding priors ;
Initialise the cluster partition using k-means clustering;
while $t < num_{iterations}$ & not converged do
    Sample a random permutation $\tau$ of $1, \ldots, N$ ;
    for $j \in \tau$ do
        Remove $\mathbf{z}_j$ from its current cluster and update cluster's sufficient statistics ;
        Compute the probabilities of joining an existing cluster and of starting a new cluster using (2.14) and (2.15);
        Set $c_j = \arg\max_{1,\ldots,K,k^*} \log p(c_j = k | c_{-j}, \mathbf{Z}, \boldsymbol{\pi}, \alpha)$ and update the cluster's sufficient statistics ;
    end
    if $num_{points\ in\ (a)\ cluster(s)} < 10$ then
        Compute the probabilities of joining all the points to a cluster with more than 10 data points ;
        Add all points to the cluster with the highest probability and accept/reject with a simulated annealing acceptance probability ;
        Update the cluster's sufficient statistics
    end
    Update model parameters using simulated annealing ;
    Compute the model log posterior ;
    Check for convergence.
end
```

We examined a merge move based on Euclidean distance as well. In this case, we considered merging clusters with fewer than 10 data points with the closest large cluster, where we determined the closest cluster by measuring the distances to cluster means. This, however, was not a very efficient split-merge scheme because the proposed merges were often rejected as different data points from the tiny clusters were often close to two different large clusters and hence, adding all points from the small cluster to one of these large clusters had a low acceptance probability (see Figure 4.7).

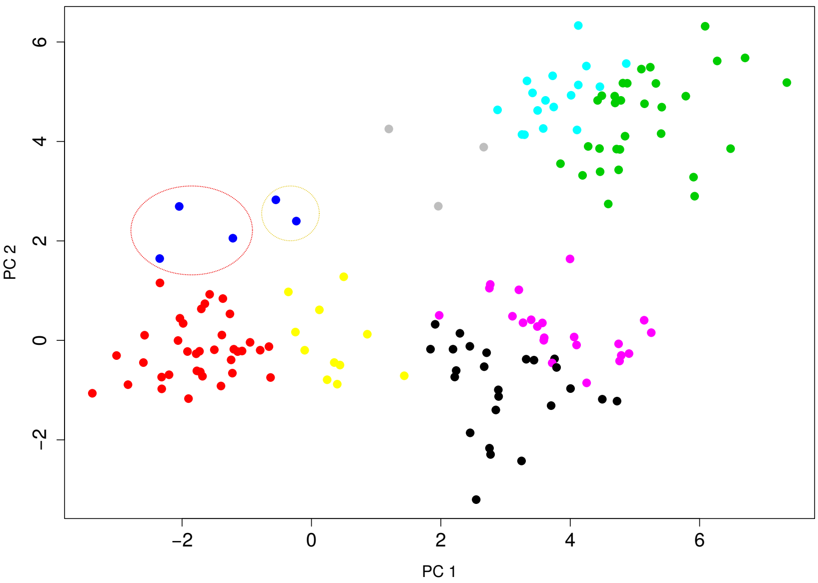


Figure 4.7: An example of why a merge move based on Euclidean distance does not work well in the case of BayesCluster. If we propose a merge of the blue cluster with a larger cluster (in this case the red cluster) based on the distance of the blue points from the centres of the large clusters, it will be rejected since some of the blue points (circled in yellow) would favour a merge with another cluster (the yellow one).

We present below a comparison between the results of using split-merge and of not using split-merge on a synthetic continuous dataset, generated in the same way as outlined in Chapter 3.

Using the adjusted Rand Index to compare the clustering partitions (Table 4.6), we see that we managed to escape local maxima consistently with the proposed-split merge, and obtained better defined clusters (see Figure 4.8 and Figure 4.10), with no very small clusters. Without applying split-merge, there were cases when we ended up stuck in local maxima with a partition including small clusters (see Figure 4.9).

Approach	Mean ARI (± std. error)	Est. K	Est. P	p-value
No split-merge	0.757	2,3	2	0.03066
	(0.612,0.902)	3 - 8	2,3	
Split-merge	0.891			
	(0.795,0.987)			

Table 4.6: Comparison between the accuracy of the clustering with split-merge and no split-merge on a single synthetic dataset in terms of mean adjusted Rand index (± std.error), estimated number of clusters K, estimated number of latent dimensions P. The p-value is from a t-test testing the hypothesis that there is no difference between the results from BayesCluster with and without split-merge in terms of ARI.

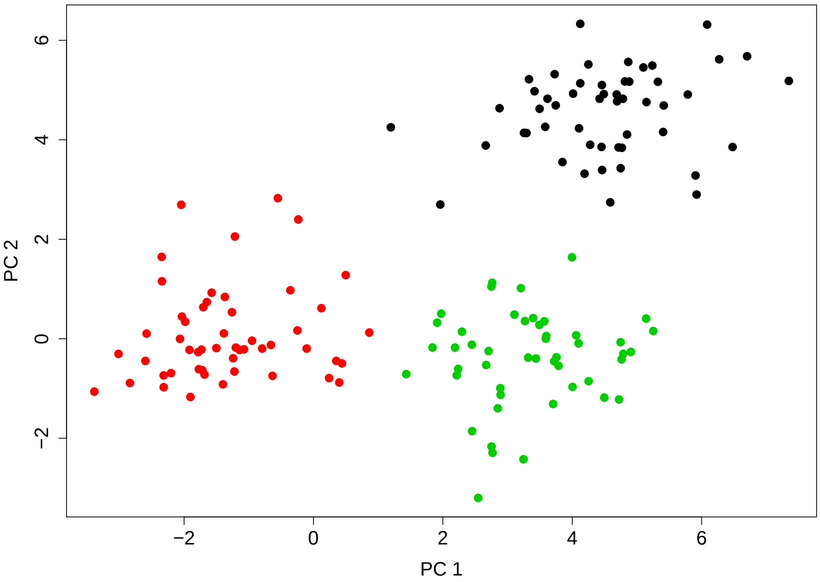


Figure 4.8: An example of a cluster partition when we use split-merge.

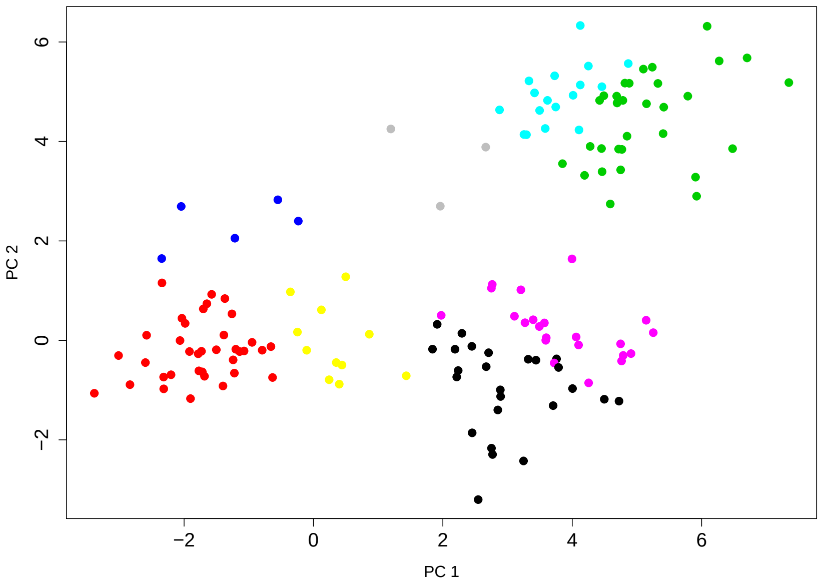


Figure 4.9: An example of a cluster partition when we do not use split-merge and get stuck in a local maximum.

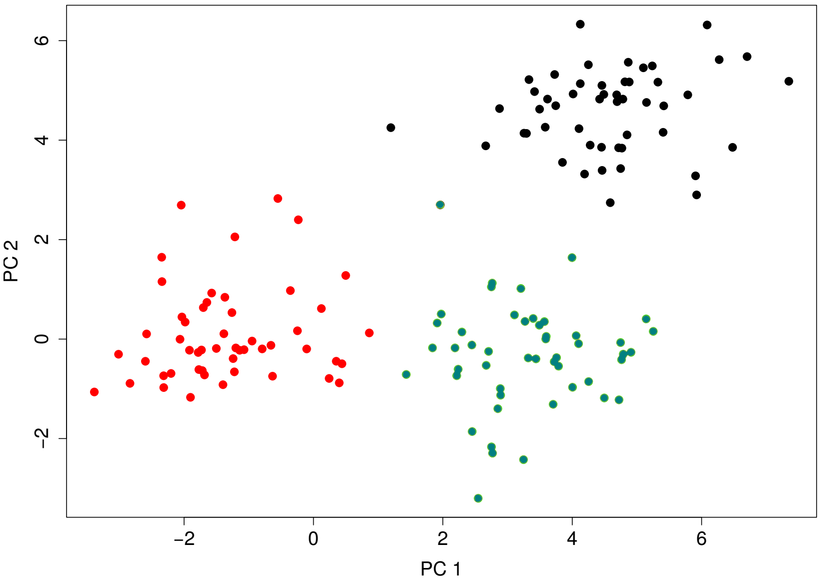


Figure 4.10: The ground truth of the clustering partition for the synthetic dataset under consideration.

4.4 Concentration parameter inference

So far we have fixed the concentration parameter α to a small value in the experiments in Chapters 3 and 4. This is a standard practice when working with Dirichlet process mixture models [Gelman et al., 2014a] as in such way, allocation to only a small number of clusters is favoured. However, it is possible to learn α from the data as both Dunson [2009] and Gelman et al. [2014a] point out that in practice, the data are highly informative about the concentration parameter and a Bayesian approach to learning α is more appropriate. For example, Wang and Dunson [2011] use a prespecified grid with a large range and put a prior on each value. A common approach is to first choose a Gamma hyperprior $\alpha \sim \text{Gamma}(a, b)$, and then use MCMC methods to update α [West, 1992; Escobar and West, 1995; Richardson and Green, 1997].

Considering the different approaches, we decided to place a Gamma prior on α and update the concentration parameter with Metropolis-Hastings or simulated annealing. As the definition of non-local priors requires that $\alpha > 1$, a Gamma$(2, 1)$ prior on α would be a better choice when we use non-local priors and resampling α if we draw a value less than 1, and Gamma$(1, 1)$ prior when we do not use non-local priors (see Figure 4.11 for illustration of the hyperpriors).

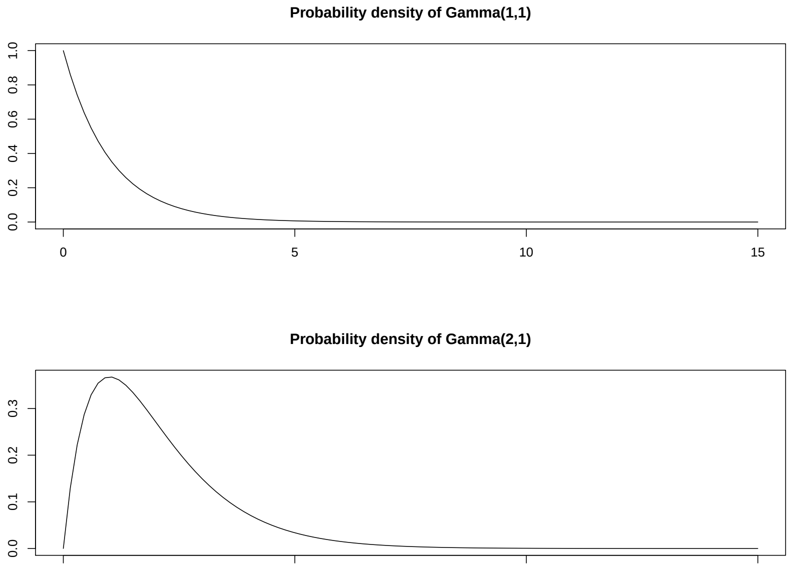


Figure 4.11: Examples of Gamma priors on the concentration parameter α.

To study the effect of updating the concentration parameter α on the final partition, we generated a continuous synthetic dataset where we sampled α from $\text{Gamma}(1, 1)$ obtaining a value of 0.456. We then generated the mixing proportions using the stick-breaking process, introduced in Section 1.3.4 by truncating the stick at $K = 3$ and setting β_3=1 [Ishwaran and James, 2001], and sampled the other model parameters and observations in the same manner as in Chapter 3. The dataset has 100 observations (3 clusters with 44, 38 and 18 observations respectively) and 100 features. The heatmap of the resulting dataset is presented on Figure 4.12, and the latent variables used to generate the dataset are plotted on Figure 4.13, coloured according to their cluster membership.

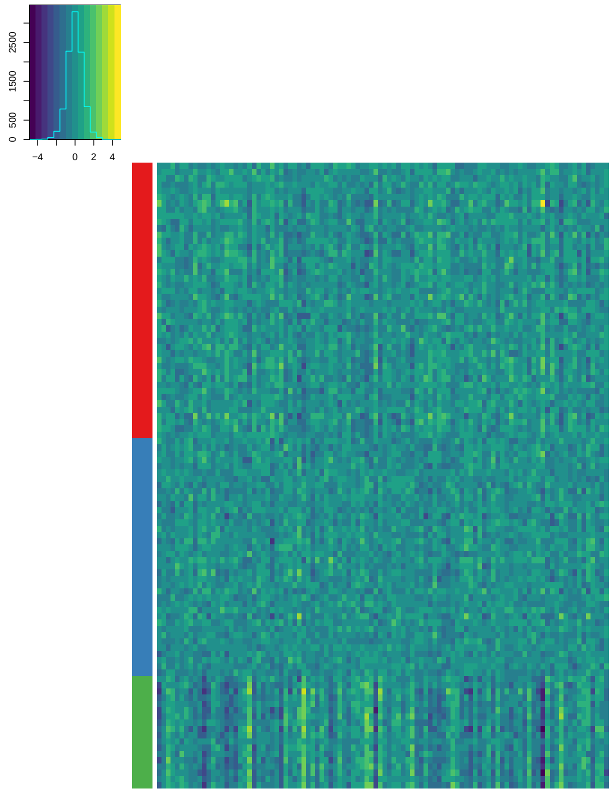


Figure 4.12: The heatmap of the synthetic dataset, generated with $\alpha = 0.456$.

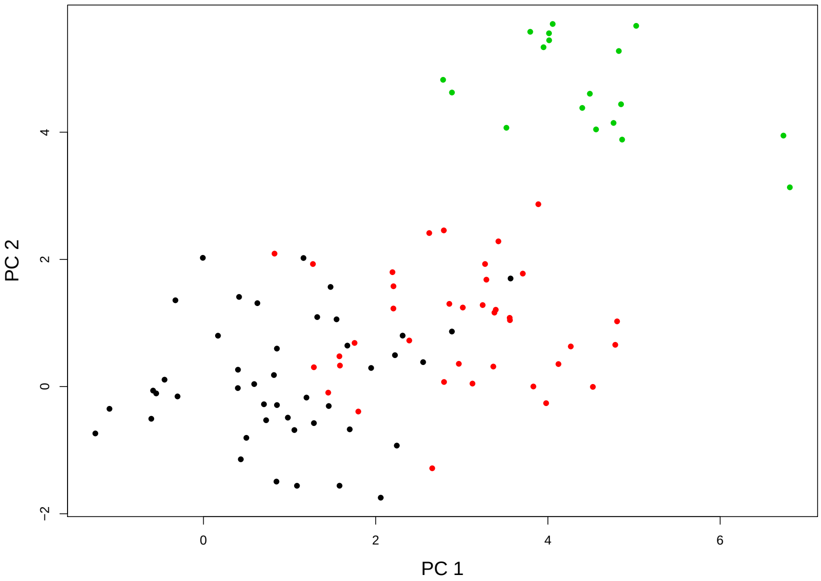


Figure 4.13: The latent variables used for the generation of the synthetic dataset (Figure 4.12) are plotted and coloured accoding to their cluster membership.

We compared the partitions we obtained when we set α to a low value (0.01 in this case) and when we updated it using the procedure outlined above. We repeated the experiment 5 times with the true value of α being 0.456 in each simulation.

If we do not update α, BayesCluster often ends up overestimating the number of clusters and sets them to 5 (see Figure 4.14). If we update α at each iteration, the final partitions have higher adjusted Rand index (mean of 0.517) and the number of clusters is closer to the ground truth ($K = 3$). Although the resulting partitions (Figure 4.16 and Table 4.7) may be seen as better in comparison with the fixed α version ones, the α values learned do not match the generating α well (Figure 4.16).

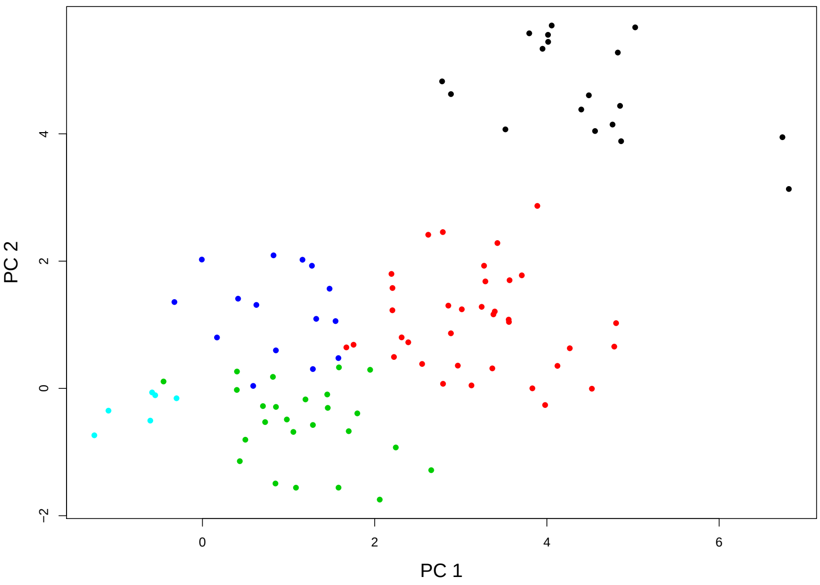


Figure 4.14: Example of a clustering partition when we do not update α.

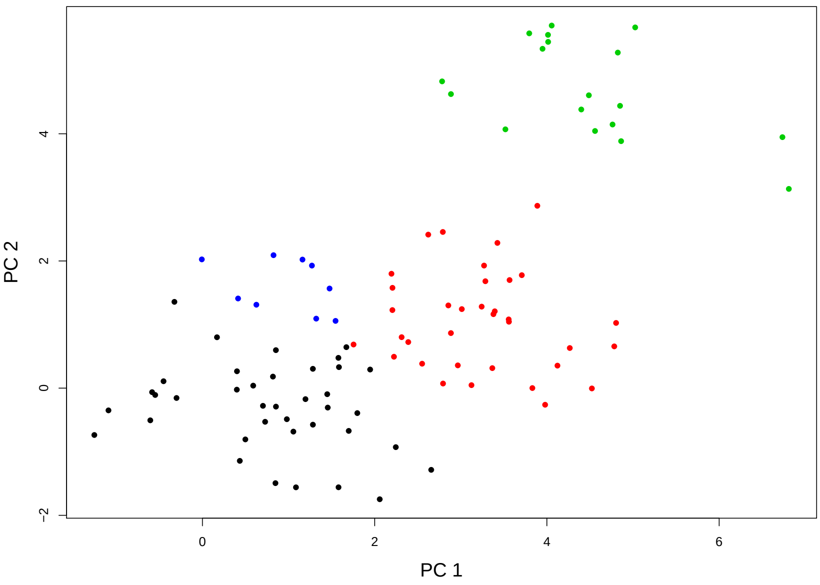


Figure 4.15: Example of a clustering partition when we update α.

Approach	Mean ARI	± std. error ARI	estimated K	estimated P
Fixed α	0.486	(0.432,0.546)	5	2
Update α	0.517	(0.500,0.534)	4	2

Table 4.7: Comparison between the partitions obtained when we update α and when we have it fixed. The p-value from t-test testing the hypothesis that there is no difference between updating and not updating α is 0.292.

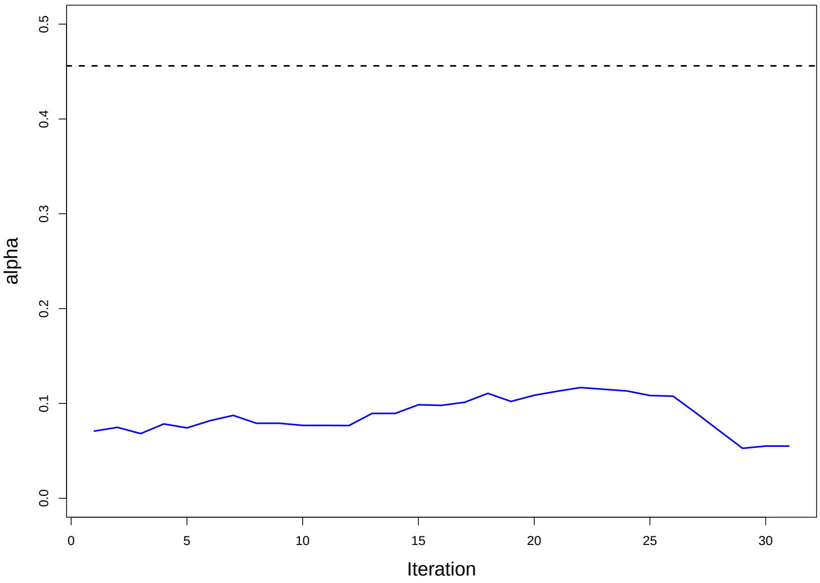


Figure 4.16: Plot of concentration parameter α, learned by BayesCluster. The generating α is plotted in a black dashed line, and the learned values are in the blue.

4.5 Conclusions

In this Chapter we developed extensions to BayesCluster that deal in a principled manner with drawbacks inherent to the Dirichlet process mixture model and MCMC methods, which play an important role in BayesCluster. We have combined the concepts of non-local priors, prior on the cluster size and split-merge MCMC and illustrated the efficacy of the extensions with real and synthetic data examples. Our experiments with the iris, Wisconsin breast cancer and glass datasets demonstrated how by incorporating non-local priors, BayesCluster could identify well defined clusters which are closer to the ground truth in cases where it previously could not. In particular, BayesCluster identified the true number of clusters $K = 3$ in the iris data, where most of the comparison methods found only 2. In addition, the split-merge remedied the issue of the inference getting stuck in local maxima, whereas imposing priors on the size of the clusters was shown experimentally in this Chapter to deal with very small clusters in a principled way.

We investigate further the applicability of BayesCluster to integrate the information from multiple real, noisy genomic datasets to identify more clinically meaningful subtypes in Chapter 5.

Chapter 5

Data Integration using BayesCluster

5.1 Motivation

Cancer is a complex disease, driven by a range of genetic and environmental factors. It is highly heterogeneous in its formation and progression, and response to treatment. This heterogeneity is usually a consequence of genetic, transcriptomic, epigenetic, and/or phenotypic changes in different cancers and even within tumours [Vogelstein et al., 2013], and this creates great challenges to treating patients effectively and to developing novel treatments. The advancement of high-throughput technologies, however, has made it possible to collect more detailed and precise data about large cohorts of patients that could be used to gain better understanding of the genetic makeup of many cancers, and to detect and treat cancer in a timely manner [Levy and Myers, 2016; Lipinski et al., 2016; Gagan and Van Allen, 2015]. These advances also signify a shift in the treatment paradigm - from 'one size fits all' to more personalised, genotype-guided treatments. For example, there are available screening tests for bowel cancer in the UK and Canada [Aubin et al., 2011; Tan and Du, 2012], which enable patients with unaltered KRAS gene to get more appropriate treatment much quicker. There is an oestrogen receptor test for breast cancer as well which can be used to guide the patients' treatment [National Cancer Institute, 2018a].

This personalised approach is expected to lead to accelerated and more precise diagnosis, early disease detection and improved targeted therapies to boost efficacy and to reduce adverse drug reactions [Ginsburg and Phillips, 2018]. Personalised medicine is an exciting opportunity with many benefits not only for the patients but also for the doctors, hospitals and the health system. To make the most of it, however, the right patient needs to be identified first. Molecular data has shown great promise in identifying cancer subtypes that are indicative of response to treatment and overall survival [Guinney et al., 2015; Haque et al., 2012; Collisson et al., 2011]. Many different molecular data types can be informative of the disease progression and response to treatment. For example, gene expression has been used in identifying subtypes prognostic of survival outcome [Tibshirani et al., 2002; Van't Veer et al., 2002; Golub et al., 1999], whereas methylation levels have been shown to provide good biomarkers for different tumours [Cancer Genome Atlas Research Network and others, 2014; Rodríguez-Rodero et al., 2013; Kulis and Esteller, 2010]. These data types often provide distinct but also complementary views of cancers as they interact with each other [Kulis and Esteller, 2010]. This is why the focus of research has shifted towards integrative approach to identifying cancer subtypes [Shen et al., 2009; Mo et al., 2013; Savage et al., 2010; Yuan et al., 2011; Kirk et al., 2012; Lock and Dunson, 2013]. This approach identifies patient groups that share similar molecular characteristics across the different data types and incorporate the interactions between the different molecular data in the final output.

In Chapter 1, we saw that a lot of the current integrative clustering methods suffer from slow parameter inference and are limited in the type of the datasets they can model. These drawbacks motivate the development of an integrative algorithm, which can model different types of data and has efficient inference. We extend BayesCluster to model more than one dataset in this chapter. The use of a mixture model in this extension allows to model easily different types of data and perform faster, more efficient inference.

5.2 Data Integration with BayesCluster

5.2.1 The integrative clustering framework

BayesCluster can be extended to a combined data integration and clustering model and applied to a wide range of different data types. The core idea of integrative clustering is that the model learns a common set of latent features (in our case cancer subtypes) jointly from the multiple data types.

Let us consider a study where we have m datasets from different data sources about the same set of patients, and we want to identify patient groups that share similar molecular signatures.

We assume that the datasets we model share the same set of latent variables $\mathbf{Z}$, which represent the shared structure between the datasets.

We can jointly estimate $\mathbf{Z}$ from the available datasets. The key idea of the integrative framework is to reduce the high-dimensional datasets to a low-dimensional subspace which still captures the major data variations. We then model the lower-dimensional representation of the data, rather than the high-dimensional dataset, and determine the patient subtypes using a Dirichlet Process mixture model [Rasmussen, 1999]:

$$p(\mathbf{Z}) = \sum_{k=1}^{\infty} \pi_k f(\mathbf{Z}|\theta_k), \tag{5.1}$$

where $p(\mathbf{Z})$ is the probability density model for the latent variables, π_k's are the mixing proportions, f is a parametric density and $\boldsymbol{\theta}_k$ are the parameters associated with the k^{th} component.

Using this integrative framework not only identifies the common structure shared between the different data types, but also models appropriately the individual dataset structure, which ultimately leads to the identification of more clinically meaningful subtypes.

We assume that the means of the clusters of latent variables $\boldsymbol{\mu}_k$ have the Moment prior, as introduced in Chapter 4:

$$p(\boldsymbol{\mu}_k) = \frac{1}{C_k} \prod_{1 \leq i < j \leq k} \frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\intercal (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{g} \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{0}, g\mathbf{I}), \tag{5.2}$$

where C_k is the normalising constant and g is the dispersion parameter which drives the separation between the clusters. Using non-local priors in the model offers better separability between the clusters and does not involve computations with high complexity as we already saw in Chapter 4.

5.2.2 Statistical models

For the analyses in this Chapter, we use two different statistical models, which are applied for real-valued and discrete data, respectively.

We model each D-dimensional continuous observation in dataset t (such as gene expression, copy number variation, microRNA) $\mathbf{x}_{it}$ by a Gaussian likelihood with unknown mean and precision

$$p(\mathbf{x}_{it}|\mathbf{z}_i, \mathbf{W}_t, \boldsymbol{\varepsilon}_{it}) = \mathcal{N}(\mathbf{x}_{it}|\mathbf{W}_t\mathbf{z}_i, \sigma_t^2\mathbf{I}), \tag{5.3}$$

where the P-dimensional latent variables $\mathbf{z}_i$ represent the molecular subtypes to be discovered. $\mathbf{W}_t \in \mathbb{R}^{D\times P}$ is the loadings matrix associated with dataset t and that maps the data to a lower dimensional space, and $\boldsymbol{\varepsilon}_{it}$ is the error term, representing the residual variance.

If we have discrete data, for example binary data indicating the presence of mutations, we assume that observations are modelled as the realisation of multinomial distribution whose parameters are achieved through a softmax transformation of the linear projection of the latent factor vector. Therefore the probabilistic model has the following form:

$$p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}^D_{1:R}, \mathbf{w}^D_{01:0R}) = \prod_{r=1}^{R} \mathrm{Cat}(x_{ir}|\mathcal{S}(\mathbf{W}^D_r\mathbf{z}_i + \mathbf{w}^D_{0r})) \tag{5.4}$$

where $\mathbf{W}^D_r$ is the loadings matrix for the r^{th} response variable and $\mathbf{w}^D_{0r}$ is the offset term for the r^{th} response variable.

In the case of mixed type of data (for example, a clinical dataset containing information such as age, gender, follow-ups, TNM staging), we can treat the dataset as the result of integration of two datasets - a continuous one and a discrete one, and the probabilistic model has the following form:

$$p(\mathbf{x}_i|\mathbf{z}_i, \mathbf{W}^C, \boldsymbol{\varepsilon}_i, \mathbf{W}^D, \mathbf{w}^D_0) = \mathcal{N}(\mathbf{x}^C_i|\mathbf{W}^C\mathbf{z}_i, \sigma^2\mathbf{I}) \prod_{r=1}^{R} \mathrm{Cat}(x^D_{ir}|\mathcal{S}(\mathbf{W}^D_r\mathbf{z}_i + \mathbf{w}^D_{0r})), \tag{5.5}$$

where $\mathbf{x}^C_i$ and $\mathbf{x}^D_i$ are the continuous and discrete part of the i^{th} observation respectively, $\mathbf{W}^C$ is the loadings matrix associated with the continuous part, $\mathbf{W}^D$, $\mathbf{w}^D_0$ are the loadings matrices and offsets associated with the discrete part. This case has already been extensively considered in Chapter 2 where we presented the model, and in Chapter 3 where we performed numerical experiments.

If we have m datasets $\mathbf{X}_1, \ldots, \mathbf{X}_m$ from different data sources (continuous and discrete), then the mathematical form of the model which integrates the information from all datasets is as follows:

$$
\begin{aligned}
\mathbf{X}_1 &= \mathbf{W}_1\mathbf{Z} + \boldsymbol{\varepsilon}_1 \\
&\vdots \\
\mathbf{X}_k &= \mathbf{W}_k\mathbf{Z} + \boldsymbol{\varepsilon}_k \\
p(\mathbf{X}_{k+1}|\mathbf{W}_{k+1,1:R},{}^D\,\mathbf{w}_{k+1,r:k+1,R}) &= \textstyle\prod_{i=1}^N \prod_{r=1}^R \mathrm{Cat}(X_{i,k+1}|\mathcal{S}(\mathbf{W}^D_{k+1,r}\mathbf{z}_i + \mathbf{w}^D_{k+1,r})) \\
&\vdots \\
p(\mathbf{X}_m|\mathbf{W}^D_{m,1:R},\mathbf{w}^D_{m,r:m,R}) &= \textstyle\prod_{i=1}^N \prod_{r=1}^R \mathrm{Cat}(X_{i,m}|\mathcal{S}(\mathbf{W}^D_{m,r}\mathbf{z}_i + \mathbf{w}^D_{m,r}))
\end{aligned}
\tag{5.6}
$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_k$ are the continuous datasets which have been normalised (have mean zero and variance 1), $\mathbf{X}_{k+1}, \ldots, \mathbf{X}_m$ are the discrete datasets, $\mathbf{W}_1, \mathbf{W}_2, \ldots \mathbf{W}_k$, $\mathbf{W}^D_{k+1,1}, \ldots, \mathbf{W}^D_{k+1,R}, \ldots, \mathbf{W}^D_{m,1}, \ldots, \mathbf{W}^D_{m,R}$ are the loading matrices which map the corresponding data onto a lower dimensional space, $\mathbf{w}^D_{k+1,1}, \ldots, \mathbf{w}^D_{k+1,R}, \mathbf{w}^D_{m,1}, \ldots,$ $\mathbf{w}^D_{m,R}$ are the offset terms and $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_k$ are the remaining variances unique to each data type after accounting for correlation between data types.

5.2.3 Inference

If we are integrating $\mathbf{X}_1, \ldots, \mathbf{X}_m$, in order to infer the model parameters, we need to derive the full joint distribution. Using the full joint distribution, we can easily find the expressions for the conditional distributions of the model parameters. However, modelling multiple datasets from different data sources means that we can rarely derive closed-form expressions for the conditional distributions; for example, here we can derive a closed-form expression for the conditional distribution of the cluster means only. Hence, we can not use Gibbs sampling for the inference.

As in Chapters 2 and 3, we use simulated annealing [Kirkpatrick et al., 1983], which allows to explore efficiently a high-dimensional sample space, which would take significantly longer time if we were to use an MCMC algorithm.

In the case when we integrate only continuous datasets, we infer the latent variables $\mathbf{Z}$, cluster means $\boldsymbol{\mu}$, the noise variables $\boldsymbol{\varepsilon}_i$ using the proposal distributions (2.1.1), and use the approximation $\mathbf{X}_i \approx \mathbf{W}_i\mathbf{Z}$ to learn the corresponding loadings matrices.

When we integrate continuous and discrete datasets, we infer the latent variables $\mathbf{Z}$, cluster means $\boldsymbol{\mu}$, the noise variables $\boldsymbol{\varepsilon}_i$, the loadings matrices for the discrete observations $\mathbf{W}^D_1 \ldots, \mathbf{W}^D_R$ and the offset terms $\mathbf{w}^D_{01}, \ldots, \mathbf{w}^D_{0R}$ using the proposal distributions (2.1.1) and (2.23), and use the approximation $\mathbf{X}_i \approx \mathbf{W}_i\mathbf{Z}$ to learn the loadings matrices for the continuous variables.

In the case of integration of discrete datasets only, we infer all model variables using simulated annealing.

Initialisation

When we integrate continuous datasets, we apply probabilistic principal component analysis (PPCA) [Tipping and Bishop, 1999] to each dataset, and use the output loadings to initialise $\mathbf{W}_i$. We pick the latent variables $\mathbf{Z}$ to be the PPCA scores from one of the datasets (chosen at random). We initialise the error terms using random draws from an inverse Gamma prior $\text{IG}(1,1)$. We apply k-means clustering to the latent variables to initialise the cluster partition.

When we model discrete data, we initialise the loadings $\mathbf{W}_1^D, \ldots, \mathbf{W}_R^D$ and the offsets $\mathbf{w}_{01}^D, \ldots, \mathbf{w}_{0R}^D$ by sampling from their Normal priors, and the latent variables $\mathbf{Z}$ by using the output from logistic PCA [Landgraf and Lee, 2015] applied to the discrete data.

Since the initialisation of the model parameters depends on the dataset we choose for estimating $\mathbf{Z}$, we run the data integration method for all possible initialisation scenarios, i.e. we use each of the datasets in turn to initialise the latent variables $\mathbf{Z}$ and run the algorithm for a range of number of principal components (from 2 to 10). In addition, simulated annealing is not guaranteed to find the optimal solution [Kirkpatrick et al., 1983] and the estimates will depend on the initial parameters. Therefore, for each initialisation scenario, we run BayesCluster from multiple different starting points and select the one that corresponds to the highest log posterior.

We place a negative binomial prior on the cluster size (NBin(100,0.5) or NBin(20,0.5) in the pancreatic cancer study) and Gamma(2,1) prior on the concentration parameter α.

We run BayesCluster for 1000 iterations and assess convergence using the model log posterior and the stopping condition $|f_t - f_{t-1}| < 0.0001$, where f_t is the log posterior at iteration t.

We can summarise the steps involved in applying BayesCluster to integrate the information from multiple datasets $\mathbf{X}_1, \ldots, \mathbf{X}_m$ in the following algorithm:

Algorithm 5.1: BayesCluster for the integration of information from multiple datasets $\mathbf{X}_1, \ldots, \mathbf{X}_m$

```
for $i \in 1, \ldots, m$ do
    Perform PPCA/logistic PCA on $\mathbf{X}_i$ to initialise the latent variables $\mathbf{Z}$,
      sample the loadings matrices $\mathbf{W}^C$, $\mathbf{W}^D$, error terms $\boldsymbol{\varepsilon}_m$ and offsets $\mathbf{w}_0^D$
      from the corresponding priors (depending on the type of the data) ;
    Initialise the cluster partition by using k-means clustering on the latent
      variables $\mathbf{Z}$;
    while $t < num_{iterations}$ & not converged do
        Sample a random permutation $\tau$ of $1, \ldots, N$ ;
        for $j \in \tau$ do
            Remove $\mathbf{z}_j$ from its current cluster and update cluster's sufficient
              statistics ;
            Compute the probabilities of joining an existing cluster and of
              starting a new cluster ;
            Set $c_j = \arg\max_{1,\ldots,K,k^*} \log p(c_j = k | c_{-j}, \mathbf{Z}, \boldsymbol{\pi}, \alpha)$ and update
              cluster's sufficient statistics ;
        end
        Update model parameters using simulated annealing ;
        Compute the model log posterior ;
        Check for convergence.
    end
end
Use BIC to select the final model.
```

Illustration of the workflow involved in using BayesCluster for data integration is provided in Figure 5.1

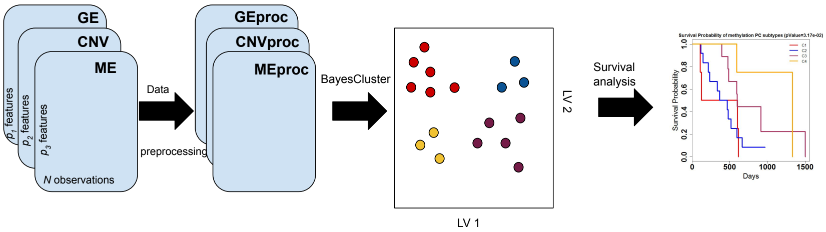


Figure 5.1: The workflow of BayesCluster. We first reduce the dimensionality of the datasets to obtain the latent variables $\mathbf{Z}$, i.e. the latent subtypes. We then model the patient subtypes using Dirichlet process mixture model and obtain a cluster partition. After that we use the clinical and omics data to specify the cancer subtypes and investigate whether they differ in overall survival. Using the follow-up data, we perform survival analysis and using the clinical data, to examine the differences between the subtypes with different survival prognosis.

5.2.4 Model selection

Since the initialisation of the model parameters depends on the dataset we choose for that, we run the data integration method for all possible initialisation scenarios for a range of number of latent dimensions ($P = 2, \dots, 10$) and we use the Bayesian information criterion to select P.

5.2.5 Connections with other models

In its construction, integrative BayesCluster is similar to iCluster, and its extensions iClusterPlus and iClusterBayes. Both BayesCluster and iCluster model continuous data using the latent variable model $\mathbf{X} = \mathbf{W}\mathbf{Z} + \boldsymbol{\varepsilon}$. However, iCluster collates the datasets $\mathbf{X}_1, \dots, \mathbf{X}_m$, which are integrated, into one data matrix $\mathbf{X}$, and the corresponding loadings $\mathbf{W}_1, \dots, \mathbf{W}_m$ into one loadings matrix $\mathbf{W}$, and then clusters the latent variables $\mathbf{Z}$ using k-means clustering.

iClusterPlus and iClusterBayes treat each of the datasets we want to integrate separately and use latent variable models differently from the ones used in BayesCluster (see Chapter 1 for more details on the specific models) to model the different data types. Similarly to BayesCluster, iClusterBayes uses Metropolis Hastings to infer the model parameter because the schemes used in iClusterPlus and iCluster are not efficient and require search over a large parameter space.

The generalised mixture of factor analysers, presented in [Khan et al., 2010], models mixed type of data in the following way:

$$p(\mathbf{x}_i|\mathbf{z}_i, c_i = k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_i^C|\mathbf{W}_k^C\mathbf{z}_i + \boldsymbol{\mu}_k^C, \boldsymbol{\Sigma}_k^C) \prod_{d=1}^{D} \mathcal{M}(\mathbf{x}_{id}^D|\mathcal{S}(\mathbf{W}_{dk}^D\mathbf{z}_i + \boldsymbol{\mu}_{dk}^D)), \quad (5.7)$$

which is similar to the approach of BayesCluster to model mixed data. However, this generalised mixture of factor analysers clusters the observations rather than the latent variables.

The integrative version of BayesCluster is built upon the BayesCluster model specification for mixed data. The main difference between the two models is the inclusion of non-local priors, prior on the cluster size and a split-merge move in the integrative model (as described in Chapter 4) in order to obtain more biologically meaningful clusters.

5.3 Data

5.3.1 Data types

In this chapter, we work mainly with the following genomic data types: gene expression, copy number variation, methylation and microRNA data.

Gene expression is expression of messenger RNA (mRNA) [1] from a given gene. The transcription of genes can be switched on/off depending on the needs and circumstances of the cell and this process is regulated by **DNA methylation** together with other mechanisms. In cancer cells, this regulation is often affected which leads to uncontrollable cancer cell proliferation [Delgado and León, 2006].

DNA methylation is an epigenetic modification of the genome that is involved in regulating many cellular processes such as transcription, carcinogenesis, X-chromosome inactivation, and genomic imprinting [Robertson, 2005]. Properly established and maintained DNA methylation patterns are essential for the normal functioning of people. There is evidence that aberrant DNA methylation is associated with multiple human diseases [Brown and Strathdee, 2002].

[1] carries the genetic information copied from DNA in 3-letter genetic code [Lodish et al., 2008]

Copy number is the number of copies of particular region of the genome occurring in that genome. Redon et al. [2006] define **copy number variation** as a DNA segment of one kilobase or larger that is present at a variable copy number in comparison with the reference genome. Some copy number variations have no effect on the phenotype whereas others have been linked to disease susceptibility [Gamazon and Stranger, 2015], for example in Mendelian disorders [Blair et al., 2013; Al-Thihli et al., 2008] and in common, complex diseases such as diabetes and cardiovascular diseases [Mitchell, 2012]. Studying the gene expression levels together with copy number variations can improve our understanding of their effect on the disease [Lonsdale et al., 2013; Henrichsen et al., 2009].

MicroRNA (miRNA) are small single-stranded, non-coding RNA molecules and can silence the expression of a particular target gene within the cell [MacFarlane and R Murphy, 2010]. They bind to target messenger RNA molecules and suppress translation of the mRNA into protein. miRNAs play an important role in the regulation of numerous metabolic and cellular pathways, including those controlling cell proliferation, differentiation and survival [Zhao et al., 2005; Monticelli et al., 2005; Garzon et al., 2006b]. Dysregulation of miRNA expression profiles has been observed in many different tumours [Garzon et al., 2006a; Volinia et al., 2006].

These processes interact with each other and are involved in the normal cell functioning. Disruption in any of them will affect the other processes. Hence, an integrative clustering algorithm which uses all data types would be able to capture these interactions and identify more clinically meaningful cancer subtypes in comparison with a clustering algorithm which uses an individual data source.

5.3.2 TCGA data

We downloaded genomic and clinical data for breast cancer, pancreatic cancer, glioblastoma and colorectal cancer from the Synapse homepage of the project `https://www.synapse.org/` (accession numbers: syn1910185, syn1910259, syn1910197, syn1910201 and syn1910239, respectively). We matched samples across all data types for each type of cancer, and removed any duplicate sample for the same patient by making a blind selection of the first sample, based on barcode ordering.

Breast cancer

We downloaded breast cancer data [Cancer Genome Atlas Network and others, 2012a], including gene expression and methylation data, as well as clinical data. After matching samples across all data types, we were left with 313 samples for which we have complete genomic data.

We used the publicly available level 3 gene expression data on the UNC AgilentG4502A_07 platform and level 3 methylation data on HumanMethylation450 platform. We selected the genes to work with based on their variability within each of the datasets. We set the threshold for gene expression to 2.1 and the threshold for methylation to 0.3. The threshold values were selected so that the number of genes fulfilling the criterion is as close to 100 as possible. This approach left us with 122 genes in the gene expression dataset and 115 in the methylation dataset.

We included clinical data about the patients, which contains information about the tumour stage, the patient treatments, age, ethnicity. We use this information to further specify the patient subtypes.

Pancreatic cancer

We downloaded pancreatic cancer data, including gene expression, copy number variation and methylation data, as well as clinical data. After matching samples across all data types, we are left with 34 samples for which we have complete genomic data.

We use the publicly available gene expression data on the Illumina HiSeq platform. We use the publicly available level 3 methylation data on HumanMethylation450 platform. We selected the genes to work with based on their variability within each of the datasets. We set the threshold for gene expression to 2.1 and the threshold for methylation to 0.26. The threshold values were selected so that the number of genes fulfilling the criterion is as close to 100 as possible. This approach left us with 141 genes in the gene expression dataset and 143 in the methylation dataset.

We included clinical data about the patients, which contains information about the tumour stage, the patient treatments, age, smoking status, presence of diabetes. We use this information to further specify the patient subtypes.

Colorectal cancer

We downloaded colorectal cancer data [Cancer Genome Atlas Network and others, 2012b], including gene expression, copy number variation and methylation data, as well as clinical data. After matching samples across all data types, we were left with 147 samples for which we had complete data.

We used the publicly available level 3 gene expression data on the UNC AgilentG4502A_07 platform, publicly available level 2 copy number data and publicly available level 3 methylation data on HumanMethylation27 platform, and set all missing values to 0. We used level 2 copy number data as it gives us access to all probes unlike level 3 data which are segmented into regions probes. For each of the datasets we selected the most highly variable genes, which resulted in the selection of $108, 145, 103$ genes from the gene expression, copy number variation and methylation datasets. We set the threshold for gene expression to 1.8, the threshold for copy number variation to 0.6 and the threshold for methylation to 0.3. The threshold values were selected so that the number of genes fulfilling the criterion is as close to 100 as possible.

We included clinical data about the patients, which contains information about the tumour stage, the patient treatments, age, presence of colon polyps, mutations.

Glioblastoma

We downloaded glioblastoma data [McLendon et al., 2008], including gene expression, copy number variation, microRNA, methylation data, as well as clinical data. After matching samples across all data types, we were left with 211 samples for which we have complete data.

We used the publicly available level 3 gene expression data on the UNC AgilentG4502A_07 platform, publicly available level 2 copy number data, and the publicly available level 3 microRNA data, generated by UNC on the H-mirna 8x15K platform. We set all missing values to 0 as we assumed zero-centred and normalised data. We use the publicly available level 3 methylation data on HumanMethylation27 platform. The data were in the form of beta values, which measure the ratio of methylation signal to methylation + background signal. After selecting the genes based on their variability, with threshold set to 0.29, we binarised the data ($\beta > 0.85$) as the data were noisy and removed any features with fewer than 10 hits. This left us with 106 features.

We selected the genes to work with based on their variability within each of the datasets. We set the threshold for gene expression to 1.95, the threshold for copy number variation to 0.8 and the threshold for miRNA to 0.6. The threshold values were selected so that the number of genes fulfilling the criterion is as close to 100 as possible. This approach left us with 122 genes in the gene expression dataset, 115 in the copy number variation dataset, and 125 in the miRNA dataset.

We included clinical data about the patients, which contains information about the tumour stage, the patient treatments, age, ethnicity.

We summarise the characteristics of the datasets for each type of cancer in the table below:

Cancer	nSamples	Clinical	GE	ME	CNV	microRNA
breast cancer	213	23	122	115	-	-
pancretic cancer	34	23	141	142	-	-
glioblastoma	212	23	122	106	115	125
colorectal	214	23	108	145	103	-

Table 5.1: Summary of the number of patients and number of features per data type for each of the types of cancer.

5.4 Methods

5.4.1 Inference, initialisation and model selection

We follow the inference and initialisation procedures outlined in Section 5.2 - we infer all model parameters by using simulated annealing. We use BIC to select the final model.

5.4.2 Comparison methods

We compare the performance of BayesCluster with another integrative clustering method, iClusterPlus [Mo et al., 2013]. As BayesCluster can be used on single datasets, we select methods to compare its performance with. In the case of clustering real-valued data, we choose k-means and Gaussian mixture model, and in the case of discrete data, we pick k-modes.

5.5 Experiments with synthetic data

5.5.1 Experiment 1

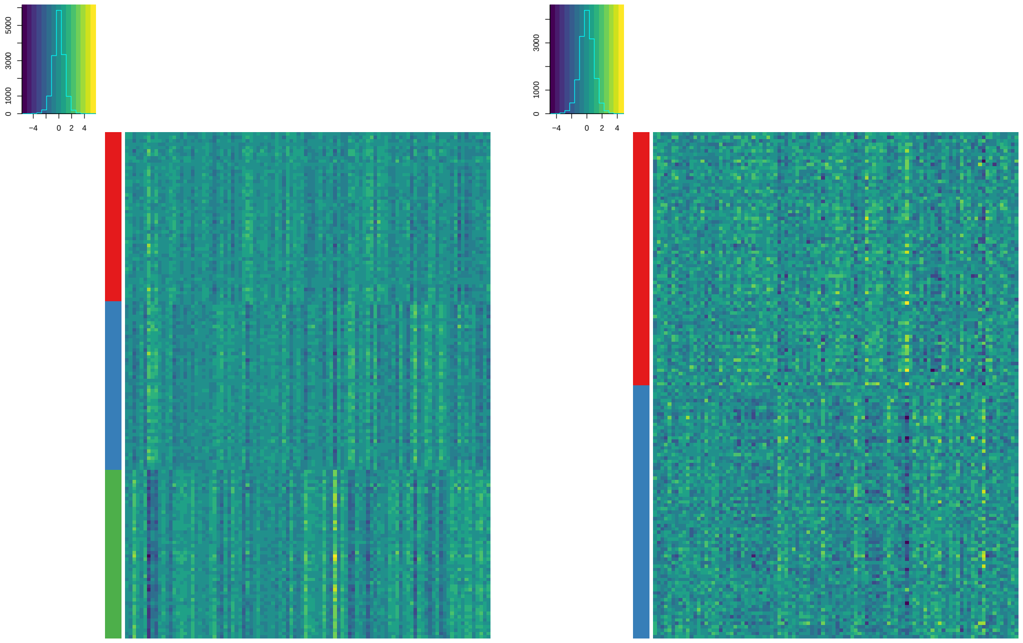
One of the main model assumptions of BayesCluster in the integrative case is that the datasets we model have the same underlying structure. This facilitates the inference scheme and the incorporation of information from different data types. However, it is not unreasonable to expect that the datasets we want to integrate have different clustering structure and that this could have an effect on the final partition. To investigate the impact of this model assumption, we performed an experiment where we used BayesCluster to cluster 10 pairs of synthetic datasets, where one of the datasets has 3 clusters, each with 50 observations and 100 features, and the other has 2 clusters, each with 75 observations and 100 features. We used the generative model of continuous BayesCluster, presented in Section 3.1 of Chapter 3, to generate the datasets and we chose the first two principal components to capture the most of the variation for each dataset. We sampled the cluster means $\boldsymbol{\mu}_{11}, \boldsymbol{\mu}_{12}, \boldsymbol{\mu}_{13}$ of the first dataset and $\boldsymbol{\mu}_{21}, \boldsymbol{\mu}_{22}$ of the second dataset as follows:

$$\begin{aligned} \boldsymbol{\mu}_{11}, \boldsymbol{\mu}_{21} &\sim \mathcal{N}((0,0), \mathbf{I}) & (5.8)\\ \boldsymbol{\mu}_{12}, \boldsymbol{\mu}_{22} &\sim \mathcal{N}((2,2), \mathbf{I}) & (5.9)\\ \boldsymbol{\mu}_{13} &\sim \mathcal{N}((4,4), \mathbf{I}). & (5.10) \end{aligned}$$

We then generated the latent variables $\mathbf{Z}_1$ of the first dataset by sampling from the following Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_{11}, \mathbf{I})$, $\mathcal{N}(\boldsymbol{\mu}_{12}, \mathbf{I})$ and $\mathcal{N}(\boldsymbol{\mu}_{13}, \mathbf{I})$, and $\mathbf{Z}_2$ of the second dataset by sampling from the following Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_{21}, \mathbf{I})$ and $\mathcal{N}(\boldsymbol{\mu}_{22}, \mathbf{I})$. After that, we generated the loadings matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ of the first and second dataset, respectively, by sampling from the priors on the rows of the loadings matrices $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the error terms $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ by sampling from their priors $\mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I})$ and $\mathcal{N}(\mathbf{0}, \sigma_2^2\mathbf{I})$, where $\sigma_1 \sim \mathrm{IG}(1,1)$ and $\sigma_2 \sim \mathrm{IG}(1,1)$. We finally generated the pair of datasets $\mathbf{X}_1$ and $\mathbf{X}_2$ using

$$\begin{aligned} p(\mathbf{x}_{1i}|\mathbf{z}_{1i}, \mathbf{W}_1, \boldsymbol{\varepsilon}_1) &= \mathcal{N}(\mathbf{x}_{1i}|\mathbf{W}_1\mathbf{z}_{1i}, \sigma_1^2\mathbf{I}) & (5.11)\\ p(\mathbf{x}_{2i}|\mathbf{z}_{2i}, \mathbf{W}_2, \boldsymbol{\varepsilon}_2) &= \mathcal{N}(\mathbf{x}_{2i}|\mathbf{W}_2\mathbf{z}_{2i}, \sigma_2^2\mathbf{I}). & (5.12) \end{aligned}$$

The heatmaps on Figure 5.2 present one pair of the synthetic datasets that were generated in this manner.



(a) dataset with 3 clusters

(b) dataset with 2 clusters

Figure 5.2: Heatmaps of two datasets with different underlying structure generated using continuous BayesCluster

We ran BayesCluster with 5 random initialisations for 1000 iterations, with the threshold for convergence set to 1e-4. We used an exponential cooling schedule for the simulated annealing scheme with starting temperature $T_0 = 100$ and cooling rate $C = 0.95$, with proposal distributions as outlined in Section 2.1.1 of Chapter 2. We initialised the model parameters as described in Chapter 2 and set α to 1. We compare the final partition to the clustering structures of the individual datasets in terms of adjusted Rand index.

Table 5.2 demonstrates the effect of modelling two datasets with different underlying structure on the performance of BayesCluster. The model overestimates the number of clusters in each of the experiments and its accuracy as compared with any of the two clustering structures is very low (mean ARIs of 0.233 and 0.182). However, these results are overly pessimistic as there is not a single ground truth to compare against, so we expect the drop in the performance of BayesCluster not to be that dramatic in reality.

Comparison dataset	Mean ARI ($\pm$ std. error)	Est. K (prop.)	Est. P (prop.)
Dataset 1	0.233 (0.062,0.405)	5-7 (0.2,0.4, 0.4)	2-4 (0.3,0.4, 0.3)
Dataset 2	0.182 (0.109,0.255)	5-7 (0.2,0.4,0.4)	2-4 (0.3,0.4, 0.3)

Table 5.2: Comparison between the accuracy of BayesCluster when integrating 2 informative datasets and when integrating an informative and a noisy datasets, in terms of mean adjusted Rand index (± std.error), estimated number of clusters K, estimated number of latent dimensions P.

5.5.2 Experiment 2

Not every dataset used in the integrative scenario will contain relevant information to obtain the final output. To study the effect of the inclusion of a non-informative dataset, we consider an experiment where we integrate two synthetic datasets where one of them has a well-defined structure and the other one is made of noise. We generated 10 pairs of datasets, with one informative and one noisy dataset each. We used the generative model of continuous BayesCluster, presented in Section 3.1 of Chapter 3, to generate the datasets and we chose the first two principal components to capture the most of the variation for each dataset. Both datasets have 3 clusters, each with 50 observations and 100 features. We sampled the cluster means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$ as follows:

$$\boldsymbol{\mu}_1 \sim \mathcal{N}((0,0), \mathbf{I}) \tag{5.13}$$

$$\boldsymbol{\mu}_2 \sim \mathcal{N}((2,2), \mathbf{I}) \tag{5.14}$$

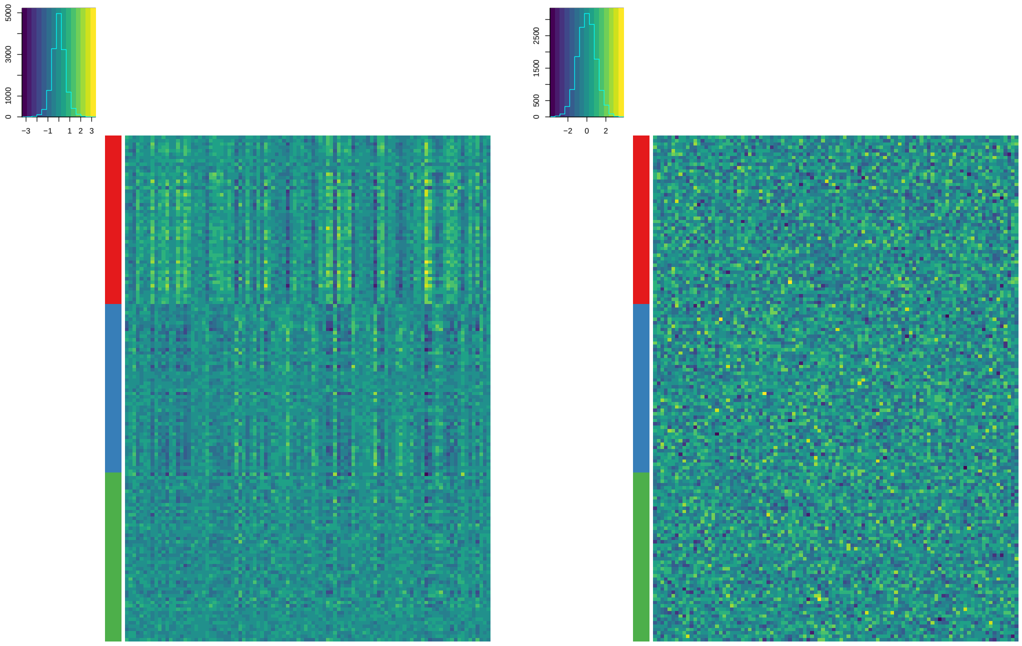
$$\boldsymbol{\mu}_3 \sim \mathcal{N}((4,4), \mathbf{I}). \tag{5.15}$$

We then generated the latent variables $\mathbf{Z}$ by sampling from the following Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I})$, $\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{I})$ and $\mathcal{N}(\boldsymbol{\mu}_3, \mathbf{I})$. After that, we generated the loadings matrix $\mathbf{W}$ by sampling from the prior on the rows of the loadings matrix $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the error terms $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ by sampling from their prior $\mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I})$, where $\sigma_1 \sim \mathrm{IG}(1,1)$. We set σ_2 to 100. We finally generated the pair of datasets $\mathbf{X}_1$ and $\mathbf{X}_2$ using

$$p(\mathbf{x}_{1i}|\mathbf{z}_i, \mathbf{W}, \boldsymbol{\varepsilon}_1) = \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{z}_i, \sigma_1^2\mathbf{I}) \tag{5.16}$$

$$p(\mathbf{x}_{2i}|\mathbf{z}_i, \mathbf{W}, \boldsymbol{\varepsilon}_2) = \mathcal{N}(\mathbf{x}_{2i}|\mathbf{W}\mathbf{z}_i, \sigma_2^2\mathbf{I}). \tag{5.17}$$

The heatmaps on Figure 5.3 present one pair of the synthetic datasets that were generated in this manner.



(a) informative dataset

(b) noisy dataset

Figure 5.3: Heatmaps of an informative and a noisy dataset generated using continuous BayesCluster.

We ran BayesCluster with 5 random initialisations for 1000 iterations, with the threshold for convergence set to 1e-4. We used an exponential cooling schedule for the simulated annealing scheme with starting temperature $T_0 = 100$ and cooling rate $C = 0.95$, with proposal distributions as outlined in Section 2.1.1 of Chapter 2. We set α to 1. We use adjusted Rand index to assess whether including a noisy dataset in the data integration task affects the final output and compare against the performance of BayesCluster on integrating pairs of informative datasets, which we presented in Chapter 3.

Table 5.3 illustrates well the drop in the performance of BayesCluster when we include a noisy dataset in the integrative scenario: the mean ARI falls from 0.52 to 0.365, and the number of datasets is consistently overestimated. In addition, BayesCluster does not estimate the number of latent dimensions P as accurately as in the informative datasets case. Although the difference in the results is not statistically significant with a p-value of 0.06717, we would recommend a careful variable

Datasets	Mean ARI (± std. error)	Est. K (prop.)	Est. P (prop.)	p-value
2 informative datasets	0.52 (0.343,0.657)	3,4 (0.6,0.4)	2,6 (0.7,0.3)	0.06717
An informative and a noisy dataset	0.365 (0.229,0.502)	5-7 (0.3,0.3,0.4)	2,3 (0.5,0.5)	

Table 5.3: Comparison between the accuracy of BayesCluster when integrating 2 informative datasets and when integrating an informative and a noisy datasets, in terms of mean adjusted Rand index (± std.error), estimated number of clusters K, estimated number of latent dimensions P. The p-value is from a t-test testing the hypothesis that there is no difference between the results from BayesCluster in the two cases.

selection in order to ensure that none of datasets we model is made predominantly of noise.

5.6 Results

We performed survival analysis on the clusters identified by the different methods, with the right-censored event being death. We considered only clusters with at least 5 patients in each case study and tested the hypothesis that there was no difference in the overall survival between the subtypes. We plotted the Kaplan-Meier curves and included the unadjusted p-value from the performed log-rank test.

We used the clinical data to further specify the identified cancer subtypes and to investigate the differences between them.

We used the following R packages:

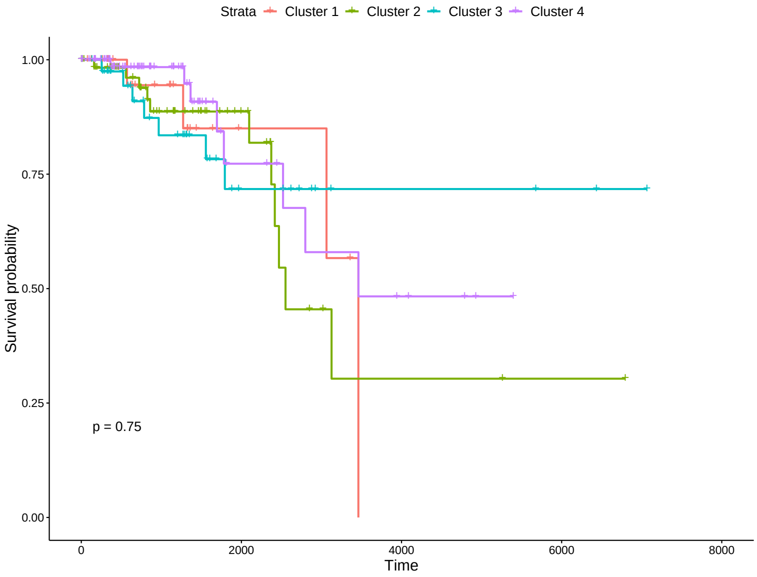
- **survival** [Therneau and Grambsch, 2013] to create the survival objects;
- **survminer** [Kassambara and Kosinski, 2018] to perform the survival analysis and plot the Kaplan-Meier curves;
- **gplots** [Warnes et al., 2016] to plot the heatmaps of the data sources;
- **viridis** [Garnier, 2018] to use a nicer colour palette for the heatmaps.

5.6.1 Breast cancer

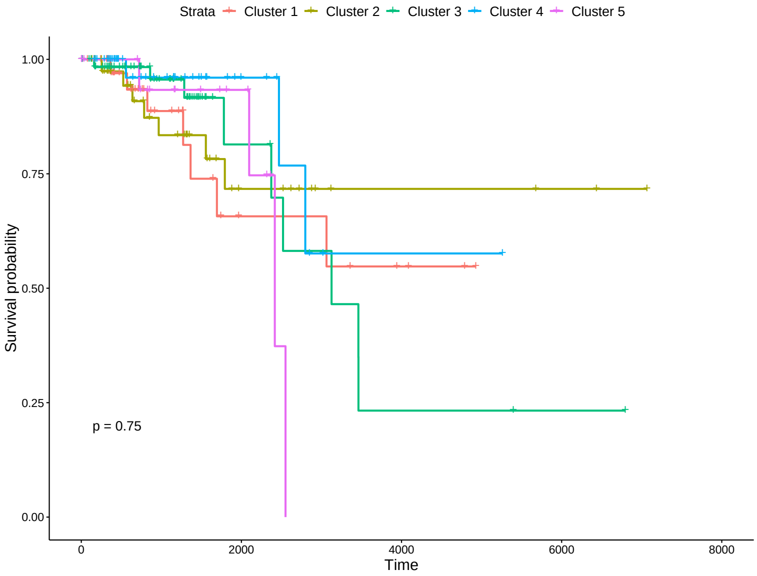
We analysed gene expression and methylation data for the 216 patients with breast cancer. We compared the results we obtained from using the individual datasets and from the integration of the data from the two sources. This can help us answer questions such as which data type drives the difference in the patient survival and whether the integration of the datasets leads to more precise subtype specification.

BayesCluster identifies 4 breast cancer subtypes using the information from all datasets. Although we cannot reject the null hypothesis that the subtypes have the same survival outcome as the unadjusted p-value of the log-rank test is 0.75 (see Figure 5.4a), it will be worth investigating further whether they are prognostic for other right-censored outcomes such as new tumour event, tumour regression or recurrence. The clinical data shows that one of the BayesCluster subtypes (Cluster 3) consists predominantly (23 out of 44, with 13 additional patients with equivocal or indeterminate Her2 status that could potentially have triple negative breast cancer) of patients with triple negative breast cancer and who do not have receptors for oestrogen, progesterone and Her2 protein (see Appendix H for more details and comparison between the patient subtypes).

We also looked at patient clusters identified by BayesCluster on the gene expression and methylation datasets individually. We used log-rank test to test that there was no difference between the subtypes in their survival prognosis. BayesCluster finds 6 and 8 patient clusters, respectively, with the survival outcome prognosis not being statistically significant in both cases (unadjusted log-rank p-values of 0.82 and 0.6, respectively, see Figures F.1 and F.2 in Appendix F).



(a) BayesCluster



(b) iClusterPlus

Figure 5.4: Breast cancer subtypes identified using integration of gene expression and methylation by BayesCluster and iClusterPlus

We compare these results with the output from iClusterPlus (in the case of integrating the gene expression and methylation data) and from k-means/k-modes and Gaussian mixture model applied to the individual datasets.

iClusterPlus identifies 5 breast cancer subtypes using the gene expression and methylation data (Figure 5.4b). Although the identified subtypes do not have statistically different survival prognosis outcome (with unadjusted p-value of 0.75), it manages to identify a subtype where the majority of the patients have triple negative subtype (25 out of 43 patients in Cluster 2 have triple negative breast cancer, with 13 additional patients with equivocal or indeterminate Her2 status that could potentially have triple negative breast cancer; see Appendix H).

Clustering the gene expression dataset using k-means and a Gaussian mixture model identifies 2 and 3 clusters, respectively, with similar survival prognosis (unadjusted log-rank p-values of 0.8 and 0.97). The results from the application of k-means to the methylation data show that methylation patient clusters do not have different survival prognosis (unadjusted log-rank p-value of 0.88), whereas the Gaussian mixture model finds only one cluster. This shows that we are able to discover more specific characteristics about the patient subgroups by using data integration and that one dataset cannot capture the complexity of breast cancer.

We present a summary of the comparisons between the models in Table 5.4. We test the null hypothesis that there is no difference between the subtypes in their survival prognosis.

model	gene expression	methylation	all
k-means	0.8	0.88	-
GMM	0.97	-	-
iClusterPlus	0.16	0.91	0.75
BayesCluster	0.82	0.6	0.75

Table 5.4: Unadjusted p-values for Kaplan-Meier survival curves (breast cancer data). We test the null hypothesis that there is no difference between the subtypes in their survival prognosis using a log-rank test.

5.6.2 Pancreatic cancer

We analyse gene expression and methylation data for the 34 patients with pancreatic cancer, and consider the subtypes identified using each individual data type only and both data types. For each case, we plot the Kaplan-Meier curves for the patient groups, identified by BayesCluster, after the removal of any patients with no follow up. We test the null hypothesis that there is no difference between the subtypes in their survival prognosis, and the resulting unadjusted log-rank p-values are 0.09

and 0.027 for the subtypes identified using only methylation and both gene expression and methylation data, respectively. In the case of the gene expression data, BayesCluster identifies only one cluster and we can see from the heatmap of the gene expression data, the measurements across the patients are similar (Figure 5.5). This implies that using only gene expression data cannot capture the complexity of pancreatic cancer, and that an integrative approach would be more appropriate.

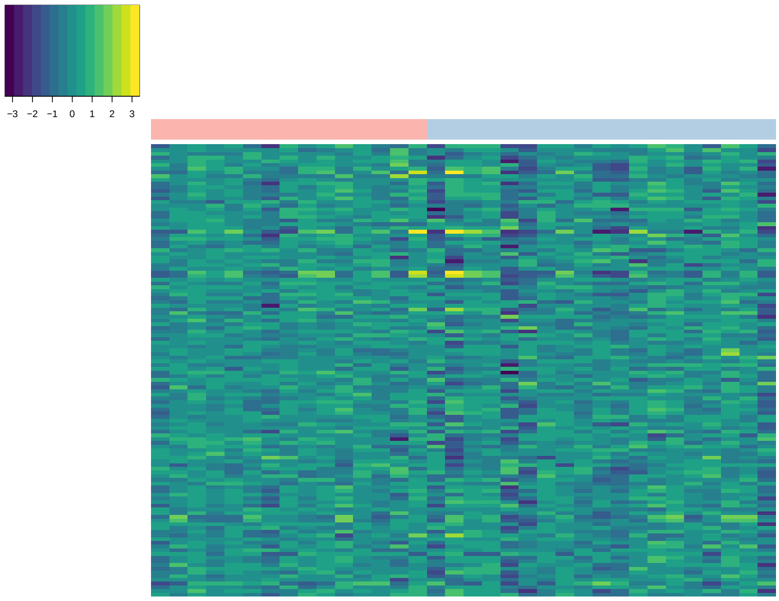


Figure 5.5: Heatmap of the pancreatic cancer gene expression data. The patients are on the x-axis and are ordered according to their membership to integrative cluster, and the normalised gene expression measurements on the y-axis.

We have identified 2 pancreatic cancer subtypes using both gene expression and methylation data (unadjusted log-rank p-value of 0.027, Figure 5.6). The patients from both clusters have similar characteristics in regards with median age at diagnosis, tumour stage, presence of diabetes and smoking history (see Appendix H for comparison between the two clusters in terms of tumour stage, age at diagnosis and other clinical characteristics). However, a third of the patients from cluster 2 (7 out of 19), which has median survival of less than 5 months, have family history of cancer, which has been associated with increased risk of pancreatic cancer [Jacobs et al., 2010].

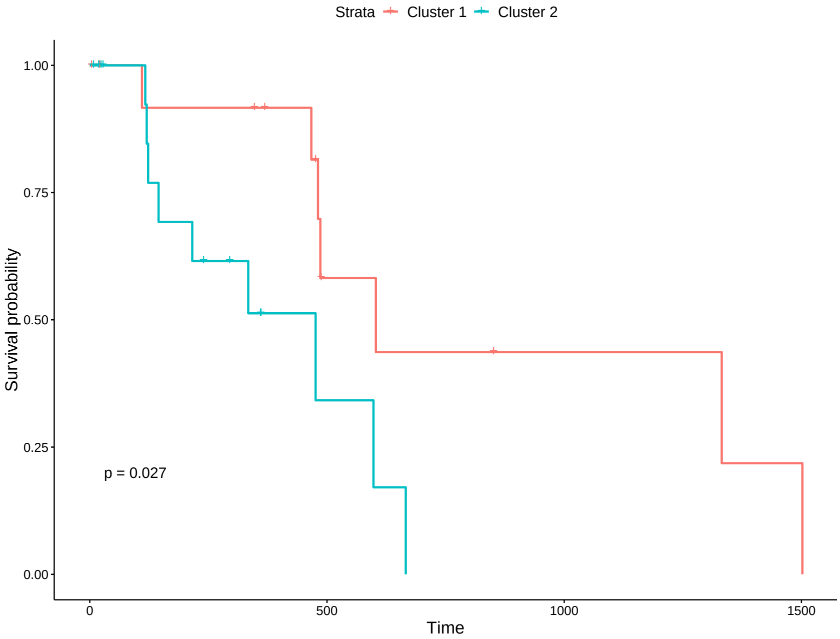


Figure 5.6: Comparison of the survival of the pancreatic cancer subtypes identified by BayesCluster using gene expression and methylation data.

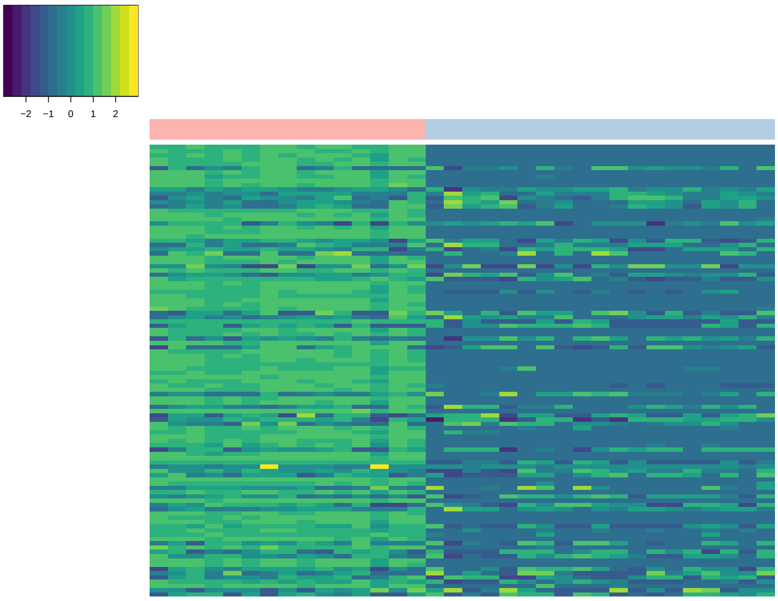


Figure 5.7: Heatmap of the methylation pancreatic cancer data. The patients, grouped using the integrative version of BayesCluster, are on the x-axis, while the genes, clustered using hierarchical clustering with average linkage, are on the y-axis.

We compare theses results with the output from iClusterPlus, k-means and Gaussian mixture model.

iClusterPlus identifies 4 pancreatic cancer subtypes using the gene expression and methylation data (Figure 5.8), for which we cannot reject the null hypothesis of no difference between the survival of the different subtypes (unadjusted log-rank p-value of 0.09).

The Gaussian mixture model finds only one gene expression subtype, similarly to BayesCluster, and identifies the same three methylation subtypes as iClusterPlus (see Figures F.3 and F.4 in Appendix F for a comparison between the different subtypes).

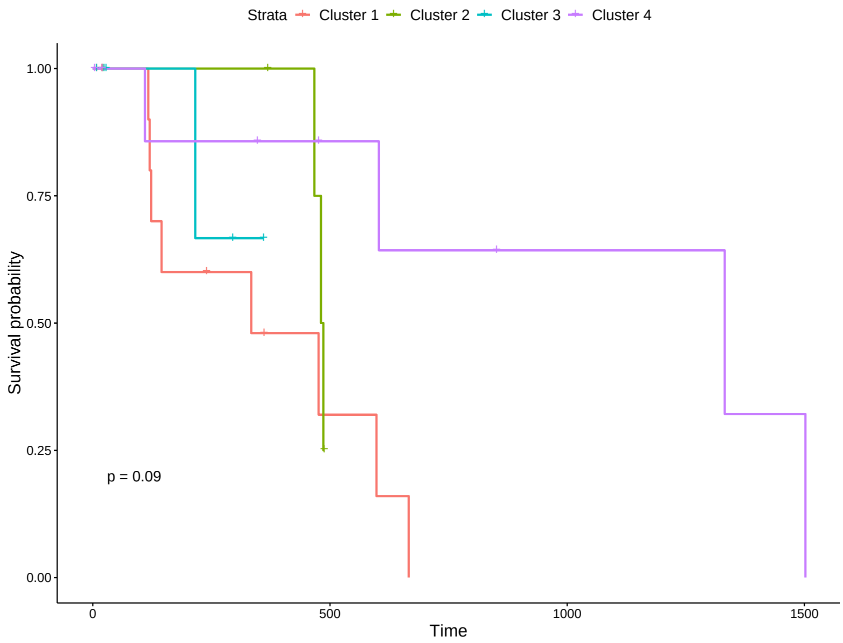


Figure 5.8: Comparison of the survival of the pancreatic cancer subtypes identified by iClusterPlus using gene expression and methylation data.

model	gene expression	methylation	all
k-means	0.178	0.054	-
GMM	-	0.082*	-
iClusterPlus	0.36	0.246	0.27
BayesCluster	-	0.18	0.054

Table 5.5: Bonferroni-adjusted p-values for Kaplan-Meier survival curves (pancreatic cancer data). We test the null hypothesis that there is no difference between the subtypes in their survival prognosis using a log-rank test. The p-value for the Gaussian mixture model is unadjusted as there is only cluster found using the gene expression data.

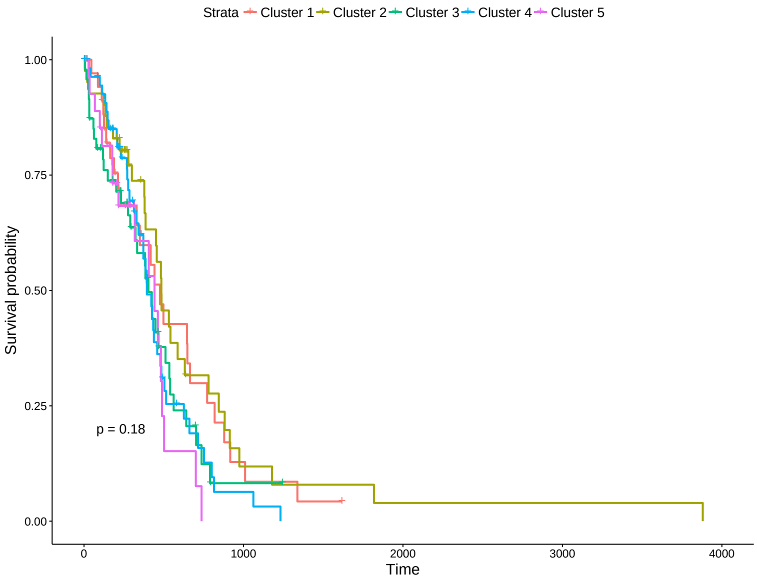
5.6.3 Glioblastoma

We analyse gene expression, copy number variation, microRNA and methylation data for the 211 patients with glioblastoma. We consider five different ways for deriving the disease subtypes: from each individual data type and from the integration of all datasets.

We perform survival analysis, with the right-censored event being death. We have considered only clusters with at least 5 patients.

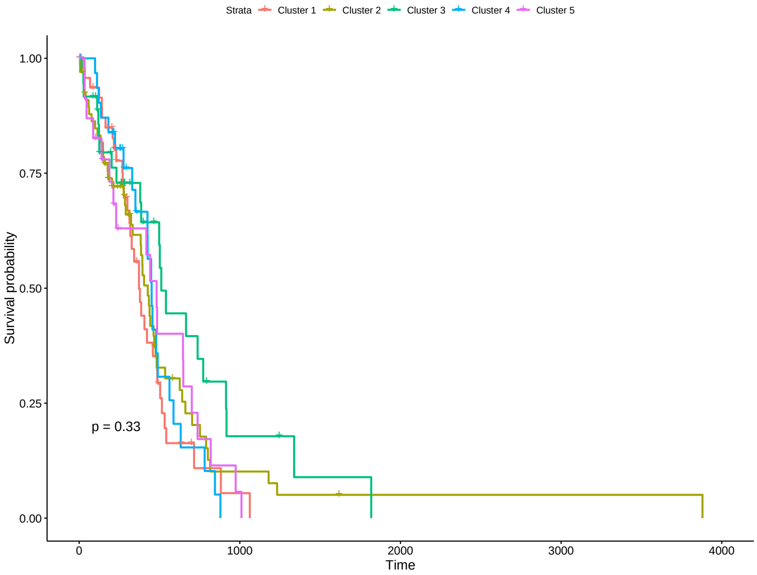
For each case, we plot Kaplan Meier curves for the patient groups, identified by BayesCluster, after the removal of any patients with no follow up. We test the null hypothesis that there is no difference between the subtypes in their survival prognosis, and the resulting unadjusted log-rank p-values are 0.18, 0.22, 0.77, 0.031, 0.33 for the subtypes identified using only gene expression, only copy number variation, only microRNA, only methylation and all data types, respectively.

We have identified 5 glioblastoma subtypes using the information from all 4 datasets (see Figure 5.9a). Patients from cluster 3, who have the best survival prognosis and median survival of over year and a half, have the lowest median age in comparison with the other patient groups (59 vs 60.5 (cluster 1), 60 (cluster 2), 59.5 (cluster 4), 60 (cluster 5)). A more detailed comparison of the clinical characteristics of the clusters can be found in Appendix H.



(b) iClusterPlus

Figure 5.9: Glioblastoma subtypes identified using integration of gene expression, copy number variation, miRNA and methylation by BayesCluster and iClusterPlus



(a) BayesCluster

Figure G.3 in Appendix G shows that integrative cluster 5 (in pink), which includes 25 patients, has distinctive copy number variation patterns (loss of copies, in particularly in genes part of the *TTTY* family) and methylation (high levels).

We have also identified a large subtype of 177 patients based on methylation (Figure 5.10) for which there is an extremely poor survival outcome, with a third of the patients dying within 6 months of diagnosis. This group of patients has low levels of methylation in the majority of genes chosen for the analysis(see the red cluster on Figure 5.11). A larger study with more patients is required to investigate this further and confirm the low methylation levels as biomarker.

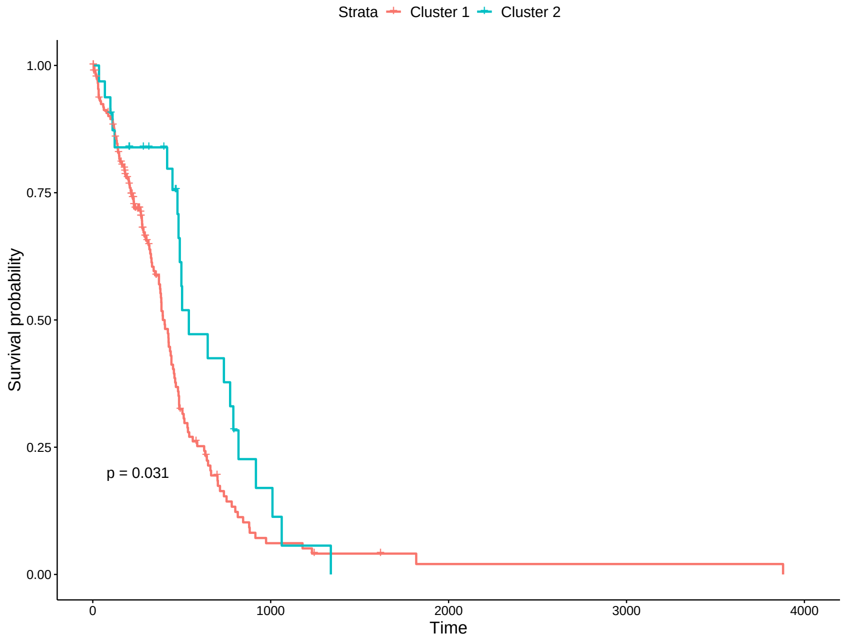


Figure 5.10: Comparison of the survival of the glioblastoma subtypes identified by BayesCluster using methylation data.

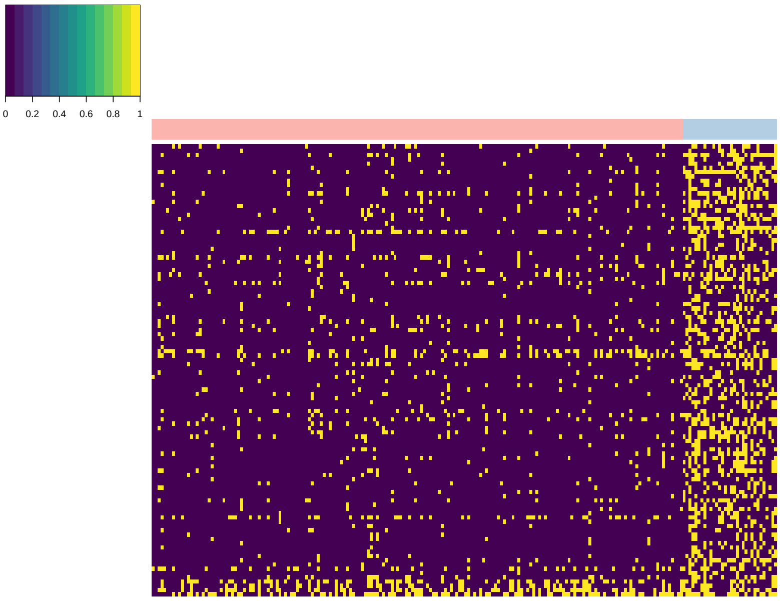


Figure 5.11: Heatmap of glioblastoma methylation data. The patients are on the x-axis, sorted by their membership to one of the two methylation clusters, whereas the genes are on the y-axis, sorted using hierarchical clustering with average linkage.

iClusterPlus identifies 5 subtypes of glioblastoma patients using the information from all 4 datasets (Figure 5.9b, unadjusted log-rank p-value of 0.18). The five groups do not have statistically different survival outcome, and there is no patient subtype that has a noticeably better survival prognosis than the rest.

We also look at the patient clusters, identified by k-means/k-modes and Gaussian mixture models (GMM) for each of the glioblastoma datasets. In each of the cases, k-means/k-modes and GMM are not able to capture the difference between the patient subtypes. This suggests that combining the information from different glioblastoma data sources could identify more clinically meaningful subtypes than using a single data source.

We present a summary of the comparisons of the models in Table 5.6. We have adjusted the p-values using Bonferroni correction in the cases where we have performed multiple hypothesis testing.

model	GE	CNV	microRNA	ME	all
k-means/k-modes	0.6	0.06	1.00	0.59*	-
GMM	0.16*	-	-	-	-
iClusterPlus	0.95	0.225	1.00	0.036	0.9
BayesCluster	0.90	1.00	1.00	0.155	1.000

Table 5.6: Bonferroni-corrected p-values for Kaplan-Meier survival curves (glioblastoma data). We test the null hypothesis that there is no difference between the subtypes in their survival prognosis using a log-rank test. As GMM resulted in one large cluster and singletons in the case of copy number variation and microRNA, we excluded these results from the analysis. The p-value for the k-modes model was not adjusted as well as the model was applied only to methylation data.

5.6.4 Colorectal cancer

We analyse gene expression, copy number variation and methylation data for the 147 patients with colorectal cancer. We compare the results we obtained from using the individual datasets and from the integration of the data from all sources.

The survival analysis of the integrative clusters found by BayesCluster reveals 2 clusters with poor survival (Clusters 1 and 5), 2 clusters with good outcomes (Clusters 3 and 4) and 1 intermediate cluster (Figure 5.12).

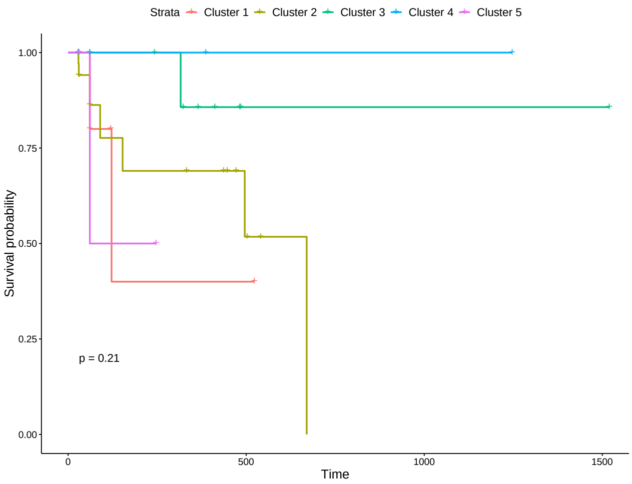


Figure 5.12: Comparison of the survival of the colorectal cancer subtypes identified by BayesCluster using gene expression, copy number variation and methylation data.

Integrative cluster 4, which has the best survival prognosis, includes patients with very low relative levels of methylation (the blue cluster on Figure G.4c), which occurs less often in cancer, and normal copy number variation measurements, apart from a few gains (Figure G.4b). In addition, the median age at diagnosis of patients from clusters 3 and 4 (63 and 68, respectively) are lower in comparison with the other three clusters with median ages of 69.5, 74 and 77 (Table H.4).

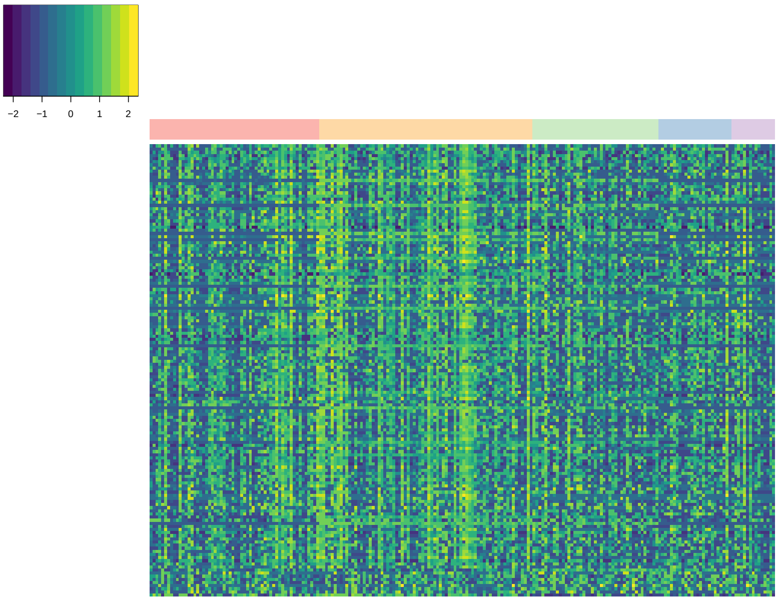


Figure 5.13: Heatmap of the colorectal cancer methylation data. The patients are on the x-axis sorted by their membership to one of the 5 integrative clusters, whereas the genes are on the y-axis, sorted using hierarchical clustering with average linkage.

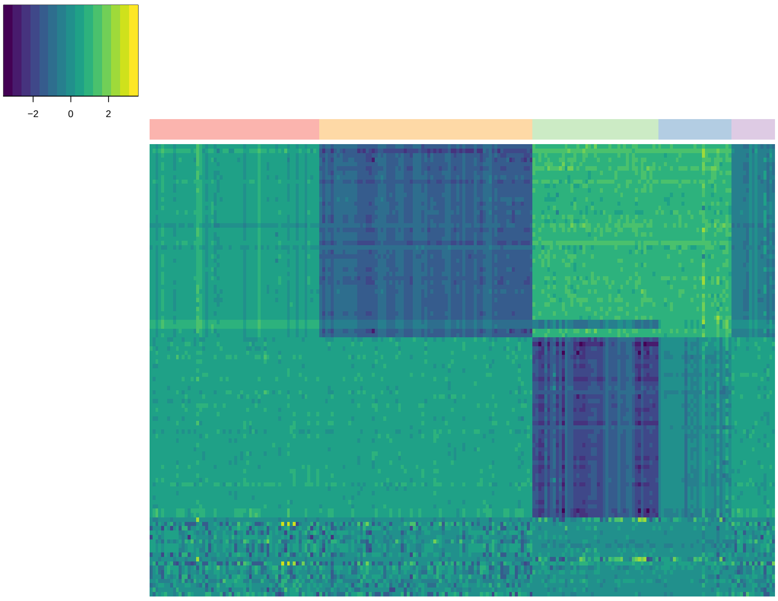


Figure 5.14: Heatmap of the colorectal cancer copy number variation data. The patients are on the x-axis sorted by their membership to one of the 5 integrative clusters, whereas the genes are on the y-axis, sorted using hierarchical clustering with average linkage.

Clustering only the copy number variation data using BayesCluster results in 5 clusters, whereas clustering only the methylation data produces 6 patient clusters as well (see Figures F.10 and F.11 in Appendix F). The lack of statistically different survival prognosis in both partitions implies that neither copy number variation data, nor methylation data on their own can capture the mechanisms underlying colorectal cancer.

We compare these results with the output from iClusterPlus (in the case of integrating gene expression, copy number variation and methylation data) and from k-means and Gaussian mixture model applied to the individual datasets.

iClusterPlus identifies 3 colorectal cancer subtypes using the information from all data sources (Figure 5.15) the identified patient subtypes do not differ in the survival prognosis, with cluster 3 having the best survival outcome.

model	gene expression	copy number variation	methylation	all
k-means	0.99	0.75	0.48	-
GMM	1.00	0.36	0.57	-
iClusterPlus	1.00	1.00	0.92	1.00
BayesCluster	0.44	0.72	1.00	0.84

Table 5.7: Bonferroni-corrected p-values for Kaplan-Meier survival curves (colorectal cancer data). The unadjusted p-values can be found on figures of the corresponding Kaplan-Meier curves. We test the null hypothesis that there is no difference between the subtypes in their survival prognosis using a log-rank test.

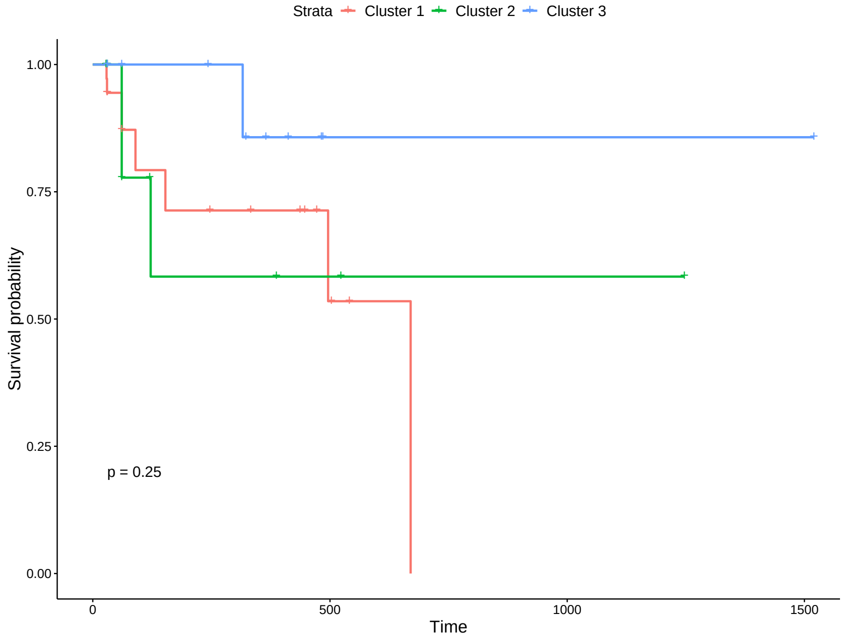


Figure 5.15: Comparison of the survival of the colorectal cancer subtypes identified by iClusterPlus using gene expression, copy number variation and methylation data.

Table 5.7 shows that using k-means clustering or Gaussian mixture model fails to capture any difference between the patient subtypes. This might be because the mechanisms driving the gene expression, methylation levels and copy number changes are too complex to be properly captured, or because the data used in the analysis were not informative.

5.6.5 Comparison of computational speed

We compared the performances of BayesCluster and iClusterPlus in terms of computational speed. We summarise the results for each of the case studies in Table 5.8 below:

Dataset(s)	nItems	nFeatures	BayesCluster	iClusterPlus
BRCA (GE&ME)	213		2hr 16min	19.25min
BRCA (GE)	213	122	55.68min	7.47min
BRCA (ME)	213	115	1hr 45min	1.93min
PAAD (GE&ME)	34		1.24 min	1.32 min
PAAD (GE)	34	141	8.16min	22 s
PAAD (ME)	34	142	6.63min	22 s
CRC (GE&CNV&ME)	214		1hr 22min	16min
CRC (GE)	214	108	32.33min	1.21min
CRC (CNV)	214	103	32.33min	1.16min
CRC (ME)	214	145	3.84min	1.11min
GBM (GE&CNV&ME&MiRNA)	212		2hr 40min	3hr 40min
GBM (GE)	212	122	50.33min	1.74min
GBM (CNV)	212	115	1hr 10min	1.45min
GBM (ME)	212	106	37.25min	1.55min
GBM (MiRNA)	212	125	34.40min	1.69min

Table 5.8: Comparison of the computational speed of BayesCluster and iClusterPlus. We used the following abbreviations for ease: BRCA (breast cancer), PAAD (pancreatic ductal adenocarcinoma), CRC (colorectal cancer), GBM (glioblastoma); GE (gene expression), ME (methylation), CNV (copy number variation), MiRNA (microRNA)

BayesCluster and iClusterPlus have similar speed performance in the pancreatic cancer data integration study. Although iClusterPlus is faster in most of the case studies, especially in the integrative cases, due to its implementation in C++, it takes over 3 hours to find the optimal integrative clustering partition in the glioblastoma study. The reason for this is that iClusterPlus has to search through 307 possible values of the penalty parameter λ for every selected number of principal components (from 1 to 9) and to find a cluster partition for each combination of λ and latent dimensionality in order to select the optimal λ. Using parallel computation to do that does not lead to speed up in the computation time for the integration of 4 datasets but the speed-up is noticeable in the case when we integrate 2 or 3 datasets. With more datasets to integrate, iClusterPlus will need to perform a search of even bigger parameter space, which will lead to a much longer computation time. One of the reasons for this is the not straightforward statistical inference, used in iClusterPlus.

5.7 Discussion and future work

We have presented an application of BayesCluster in integrative context where we aimed to discover cancer subtypes indicative of overall survival. We have applied the integrative version of BayesCluster to four different types of cancer: breast cancer, glioblastoma, pancreatic cancer and colorectal cancer. Using BayesCluster, we were able to identify subtypes which are prognostic of survival outcome in pancreatic cancer, and which we were not able to identify using iClusterPlus, k-means and Gaussian mixture model.

However, there were cases where BayesCluster could not identify cancer subtypes with different survival prognosis, which could be due to selecting uninformative features and to learning only the cluster means. We intend to explore the feature selection and interpretability in more detail by incorporating estimation of the posterior probability of each omics feature, which can be used as a criterion for feature selection as suggested by Mo et al. [2017]. We could model more appropriately the cluster variability by detecting signals in both cluster-specific means and covariances, in a manner similar to [Taschler et al., 2019].

In addition, the analysis undertaken in this chapter highlighted the great level of disagreement between the subtypes identified with the different methods. In some of the case studies, such as the pancreatic cancer one, that could be due to the low power of the study. A natural next step for each of these studies would be to validate the results by considering a different patient cohort, for example part of the ICGC project, which would allow the confirmation of low methylation levels and of EGFR mutations as biomarkers for aggressive subtypes of glioblastoma and of colorectal cancer, respectively.

Chapter 6

Studying the impact of clinical factors on 90-day and long-term survival following surgery for pancreatic cancer

This chapter explores the impact of different clinical factors on the short- and long-term survival of pancreatic cancer patients following pancreatoduodenectomy (PD), also known as Whipple procedure or pancreatic resection. Pancreatoduodenectomy is a complex surgical operation performed to remove tumours of the head of the pancreas, and is the only potentially curative procedure currently in clinical use to remove malignant pancreatic tumours [Clancy, 2015]. Centralisation, which was implemented in the UK in 2001 [Department of Health, 2011], aims to improve the outcomes of cancer surgery by centralising the surgical procedures to hospitals with higher annual volume. This could lead to improvement in the access to and quality of care, better outcomes, less invasive procedures and shorter recovery times. We present a study of the 90-day mortality following pancreatoduodenectomy in England, which we performed with data from the Hospital Episode Statistics (HES) database. To the author's best knowledge, there are no previous studies investigating the impact of centralisation on the 90-day survival of pancreatic cancer patients in the UK; hence, it is of interest to study the potential positive/negative impact of it on survival. We also look at the impact of different clinical factors on the longer term survival, in particular 2-year survival.

A note on autorship: the thesis' author performed the multivariate statistical analysis for the 90-day mortality study. This study [Liu et al., 2018] is a result of a collaboration with a great team from the University Hospitals Birmingham, the Informatics Unit, and researchers from the University of Warwick. The data were extracted by Felicity Evison, the univariate statistical analysis was conducted by Zhangdaihong Liu. Dr Richard Savage, Dr Keith Roberts and Felicity Evison provided guidance and assistance during the project. The study on the long-term survival was performed entirely by the thesis' author.

6.1 Motivation

Pancreatic cancer is the 10^{th} most common cancer in the UK [Office for National Statistics, 2017; IDS, 2018; Welsh Cancer Intelligence and Surveillance Unit, 2018]. Around 8800 people in UK are diagnosed with the disease every year [Cancer Research UK, 2018]. It most frequently occurs from the ducts within the pancreas (ductal adenocarcinoma) when abnormal cells in the pancreas grow out of control, forming a mass of tissue (tumour). Pancreatic cancer is classified as either exocrine [1] tumour (accounting for 95% of the pancreatic cancer cases) or endocrine [2] tumour based on the location of the tumour. They are diagnosed and treated differently, and they exhibit different symptoms. The most common type of cancer is pancreatic ductal adenocarcinoma, and it is predicted to become the second leading cause of cancer mortality by 2030 [Rahib et al., 2014].

It is difficult to diagnose pancreatic cancer as it usually does not give rise to any symptoms or signs in the early stages. There is no programme for pancreatic cancer anywhere in the world to screen the general population as there is no suitable test that has been developed to do this. Tobacco smoking is the only established environmental risk factor for pancreatic cancer, and patients with diabetes are also at increased risk of getting pancreatic cancer [Lowenfels and Maisonneuve, 2006]. The risk of pancreatic cancer increases with age - with over 8 in 10 cases of pancreatic cancer occurring in people over 60 [Cancer Research UK, 2018]. Despite recent medical advances, the survival rate of pancreatic cancer patients has not shown statistically significant improvement since 1971 (Figure 6.1).

[1] producing digestive enzymes

[2] producing hormone (endocrine)

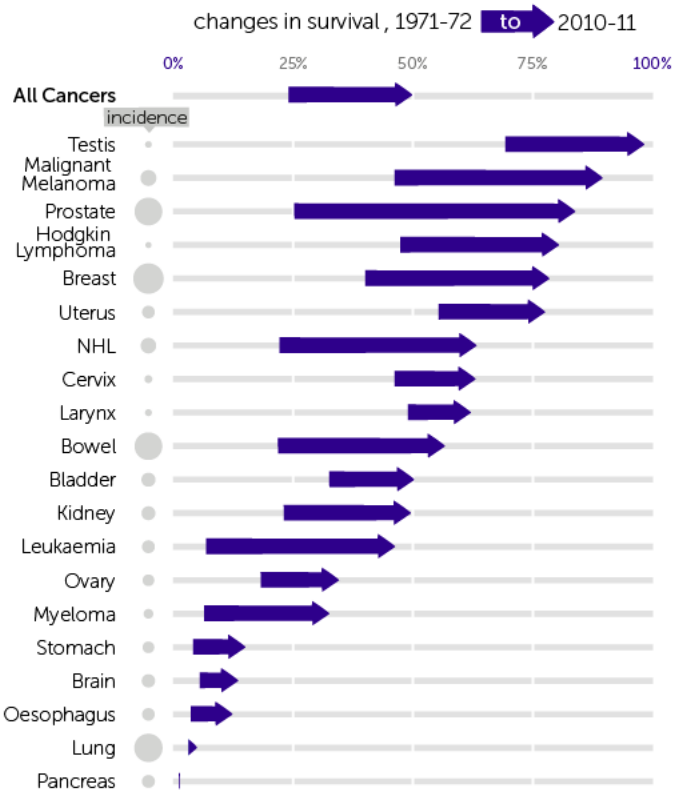


Figure 6.1: Comparison of age-standardised ten-year net survival trends of the most common cancers (adults, aged 15-99) in England and Wales over the period 1971-2011. There is no signifcant improvement in the survival prognosis for pancreatic cancer patients unlike the noticeable improvement for most of the other types of cancer. Credit: Cancer Research UK

6.2 Towards a risk score model

Despite recent advances in surgery procedures and clinical care, the perioperative [3] mortality rate associated with pancreatic resection remains very high [Büchler et al., 2007; McPhee et al., 2007]. There has been a lot of effort focused on the development of a score model [Lowenfels and Maisonneuve, 2005; Hassan et al., 2007; Raimondi et al., 2009; Yadav and Lowenfels, 2013; Maisonneuve and Lowenfels, 2014] to predict the risk of in-hospital mortality following pancreatoduodenectomy. A risk score model has the potential to lead to improvements in patient care as patients who have statistically higher risk of mortality could be provided with more support, and

[3] occurring at or around the time of an operation

with the help of their surgeon and clinicians, reduce the risk by adjusting modifiable risk factors such as their diet, alcohol intake. Here we summarise some of the risk score models, developed in the USA, the Netherlands and Japan.

Hill et al. [2010] use the **Nationwide Inpatient Sample** (NIS) to develop a model for preoperative evaluation of patients in the USA. The predictor variables, chosen based on clinical usefulness and biological plausability, include age, sex, Charlson comorbidity score [4], type of pancreatectomy performed (what proportion of the pancreas is removed) and hospital volume. The statistical analyses, performed in this study, identify all predictive variables as statistically significant factors affecting patient mortality, with patient age over 80 years and having a pancreatic resection at a low-volume centre being the factors with the largest effect on the survival. It should be noted that although this is a nationwide study, it includes only 20% of the US hospitals and hence, it may not include certain centres of excellence in pancreatic resection.

Are et al. [2009] use the NIS database as well to develop a nomogram, that can be used in the preoperative setting to counsel patients about the perioperative mortality associated with pancreatectomy. The nomograms are graphical models that use models such as Cox proportional hazards model to estimate the probability of an outcome such as cancer recurrence or death, for a given individual [Evesham, 2010] (see Figure 6.2 for an example of nomogram). The data used in the study include information about the patient's age, sex, admission type, hospital size and type, pancreatectomy type. The information about patients admitted between 2000 and 2004 was used to create a predictive model and the data from year 2005 was used to validate the model. The results showed excellent agreement between the observed and the predicted probabilities.

[4] predicts the 10-year mortality for a patient who may have existing co-morbid conditions. Higher score indicates more present comorbidities.

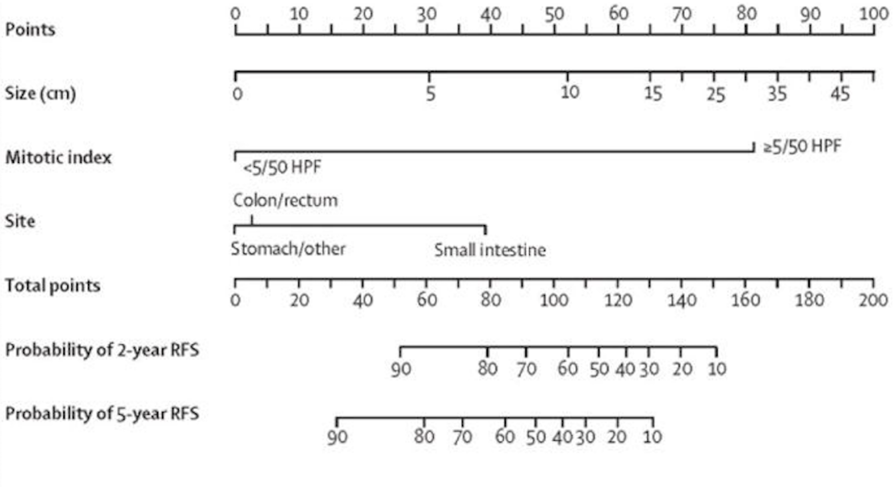


Figure 6.2: An example of nomogram, used to estimate recurrence-free survival in resected primary gastrointestinal stromal tumor. This is done in the following way: we first draw an upward vertical line to the 'Points' bar based on different features of the tumour to calculate points. Based on the sum, after that we draw a downward vertical line from the 'Total Points' line to calculate the recurrence-free survival. Credit: [Balachandran et al., 2015]

Venkat et al. [2011] develop another risk model to predict the 30-day and the 90-day mortality after a PD using data about patients admitted to the **John Hopkins Hospital** from 1st January 1998 to 30th June 2009. They include covariates such as age, Charlson index, albumin level [5], sex, tumour size, creatinine level [6], histologic diagnosis, type of surgery. The analysis shows that age, sex, tumour size, type of surgery and preoperative serum albumin level are predictors for the 30-day mortality rate, whereas age, sex, tumour size, Charlson score, type of surgery, preoperative serum albumin level are predictors for the 90-day mortality rate. The Hosmer-Lemeshow test, which is a statistical test for goodness of fit for logistic regression [Hosmer and Lemesbow, 1980], used to assess whether or not the observed event rates match the expected event rates, confirms that there are no statistical differences between observed and expected 30-day and 90-day mortality rate. Hence, the score models can be used to identify certain risk groups that may be considered for stratification or exclusion from clinical trials.

[5] the most abundant protein in human blood plasma. It is produced in the liver and is responsible for the transportation of thyroid hormones, fatty acids and many drugs. It maintains the oncotic pressure, which is generated by proteins such as albumin. Decreased oncotic pressure leads to decreased effective circulating fluid volume [Koeppen and Stanton, 2012]

[6] important indicator of renal health because it is an easily measured byproduct of muscle metabolism that is excreted unchanged by the kidneys

ACS-NSQUIP is another database that contains data about pre-operative risk factors, post-operative morbidity and mortality to assess the surgical quality at more than 200 US hospitals. Parikh et al. [2010] use this data to develop a pancreatectomy risk calculator to predict the post-operative adverse outcomes. The variables included in the model for mortality are age group, systemic sepsis, functional health status, ASA classification [7], history of congestive heart failure, dyspnoea [8], previous/concurrent chemotherapy, esophageal varices [9] and type of surgery. The variables part of the predictive model for morbidity are age group, gender, BMI classification, systemic sepsis, functional status, ASA classification, surgical extent, coronary heart disease, history of severe chronic obstructive pulmonary disease (COPD) [10], smoking status, dyspnoea, bleeding disorders and weight loss greater than 10%. The results from the fitted forward stepwise logistic regression models show that age over 74 years, male gender, BMI over 40, pre-operative sepsis, dependent functional status, ASA class more than II, history of coronary heart disease, dyspnoea, a bleeding disorder and the contemplated procedure are risk factors for pancreatic cancer.

Vollmer et al. [2012] evaluate whether any of the risk assessment tools presented above can sufficiently predict and account for actual clinical events that are often identified by root-cause analysis. A **root-cause analysis** is a retrospective method employed to understand adverse events. It allows for a more objective review of the events which lead to an endpoint. In this study, high-volume pancreatic surgical specialists from 14 academic/affiliate or private institutions and 4 countries had to provide data on preoperative demographics, disease process, medical comorbidities, operative details, and the course of postoperative care for all mortalities in their practice during the study period. They were asked to comment on the cause of the death and whether it was predictable. The study shows that none of the risk score models is superior to the others and that in many cases (about a quarter of the analysed in the paper), the death cause cannot be determined and hence prevented.

We summarise the main aspects of the presented studies in the table below:

[7] classes range from ASA I - normal healthy patient, to ASA VI - a declared brain-dead patient whose organs are being removed for donor purposes [American Society of Anesthesiologists]

[8] difficult breathing

[9] abnormally enlarged veins in the tube connecting the throat and the stomach (esophagus). It occurs most often in people with serious liver conditions

[10] a group of lung conditions (emphysema and chronic bronchitis) that cause breathing difficulties; it is common amongst middle-aged and older people who smoke.

Study	Data	Study period	Risk factors
Hill et al. [2010]	NIS	1998 - 2006	age over 80 and having a pancreatic resection in a low-volume centre
Are et al. [2009]	NIS	2000 - 2005	age, sex, admission type, hospital size and type, pancreatectomy type
Venkat et al. [2011]	John Hopkins Hospital	1998 - 2009	age, sex, tumour size, type of surgery, preoperative serum albumin level
Parikh et al. [2010]	ACS-NSQUIP	2005 - 2008	age over 74, BMI over 40, male gender, pre-operative sepsis, ASA class more than II, history of coronary heart disease, dysponea, bleeding disorder, contemplated procedure

Table 6.1: Summary of studies developing a risk score model for pancreatic cancer

6.3 Studying the impact of centralisation in other countries

Gooiker et al. [2011] aim to evaluate whether the centralisation of pancreatic surgeries in **the Western part of the Netherlands** has improved clinical outcomes and changed referral patterns. The data used in this study include information from Leiden University Medical Centre and Reinier de Graaf Hospital, and provide patient demographics information, pathological notes, TNM staging [11], data on surgical and additional treatments, comorbidities, detailed postoperative complications, length of stay and margin status of all patients who underwent pancreatic surgery between 2006 and 2008.

The researchers compared the outcomes for 3 time periods: 1996 - 2000 (when no

[11] T describes the size of the tumour and whether it has spread to nearby tissue, N - the nearby lymph nodes that are involved, and M - whether and how far the tumour has metastasised [National Cancer Institute, 2018b]

quality control was applied to the surgeries), 2001 - 2005 (when quality standards were implemented) and 2006 - 2008 (when surgeries were centralised to 2 hospitals), in order to assess whether the implemented changes had impact on the 30-day mortality, 90-day survival, 1-year survival and 2-year survival. The performed analysis found that greater risk of death was associated with higher age, a tumour located in the pancreas, stage III and IV pancreatic adenocarcinoma, and diagnosis in the early time periods. The results also show that after centralisation, the survival of the patients improved and a higher proportion of patients received surgery. In addition, the centralisation did not lead to increased waiting times and longer length of stay. Some of the reasons for the improved survival might be the better selection for surgery, improvements in the diagnosis, surgical technique or post-operative care, the better facilities in high-volume centres and the more experienced surgical team.

Although the study used reliable and complete clinical, population-based data, only data of the malignant diagnoses was collected, and no information on the structural changes in the management of the pancreatic cancer was gathered. Since there was no data on comorbid diseases until 2006, no risk adjustments could be made.

de Wilde et al. [2012] present another study performed on data from **the Netherlands**. The aims of this study were to discover whether the concentration of pancreatic cancer surgery led to higher survival and resection rates, and to evaluate the association between hospital volume and survival. The population-based Netherlands Cancer Registry (NCR) was the source of data for the study, and it contained information about patient characteristics (age, sex), tumour characteristics (TNM stage), treatment (resection, adjuvant treatment [12], hospital of treatment), hospital of diagnosis and hospital of treatment. The information covered the period from 1st January 2000 to 31st December 2009.

The comparison between the periods 2000-2004 (before centralisation was introduced) and 2005-2009 (after centralisation was introduced) showed an increase in the resection rate but no difference in the overall survival between the two periods. In addition, there was no statistically significant difference between high volume and medium/low-volume hospitals in terms of postoperative mortality. However, the difference in 1- and 2-year survival rates after resection in high-volume hospitals was statistically significant.

Though this study was based on reliable and complete clinical, population-based data which could be adjusted for confounding factors, it did not include information

[12] additional cancer treatment given after the primary treatment to lower the risk that the cancer will recur.

on comorbidity, which might have a serious impact on the survival outcome, and the type and date of surgery before 2005.

Gooiker et al. [2014] performed another **Dutch** study with data from NCR, aiming to determine the impact of hospital volume on hospital mortality, length of stay and total costs after PD. The clinical data included information such as the patient's age and sex, diagnoses and comorbidities, administered drugs, length of stay and total costs. Patients who had a PD between July and December in each year from 2007 to 2010 were identified. Hospital volume was defined as the number of PDs performed annually at each hospital and was categorised into quantiles (very low, low, medium, high and very high). The primary endpoint was in-hospital mortality, defined as death at any time before hospital discharge. Secondary endpoints were post-operative length of stay and total costs during the hospital stay.

The results of the performed statistical analyses show that patients in very high-volume group were younger than those in the very low-volume group. In addition, higher hospital volume was found to be associated with shorter length of stay and lower total costs. However, surgeon volume, which could be more informative than hospital volume, was not included in the analysis. Moreover, the adoption of the reporting system by community hospitals is voluntary so the database may not be representative of all hospitals.

LaPar et al. [2012] aim to reassess the volume-outcome relationship of the hospital procedures which use volume as a quality measure, which is adopted by the Agency for Healthcare Research and Quality (AHRQ) in **USA** as a quality indicator for several high-risk surgical procedures. The data for the study were extracted from the 2008 Nationwide Inpatient Sample (NIS) database, which contains information about in-hospital mortality, patient's age, gender, comorbidity. The models used were adjusted for differences in patient's age, sex, elective admission status and comorbid disease. Hospitals were included in the models as random effects, allowing the relationship between volume and in-hospital death to be different across hospitals.

Although hospital volume was not associated with mortality, many patient-related factors were strongly associated with mortality. The factors with the strongest associations with mortality included patient's age and comorbidities, elective status, female sex. The major finding of this study is that there was no threshold value for hospital procedure volume at which mortality risk was significantly increased. However, the study did not include analysis of the relationship between individual surgeon volume and outcome, and other important clinical endpoints including long-term survival, inpatient resource utilization or hospital readmission were not addressed.

Ho and Heslin [2003] performed another **US** study, using patient data for California and Florida for the period $1988 - 1998$. The aim was to investigate the relative impact of procedure volume versus years of hospital experience on inpatient death rates after PD. The patient characteristics chosen for the analysis included age $(< 60, 60 - 69, 70 - 79, 80+)$, gender, and comorbidities, and the patients who were treated in 1998 were used to derive the model predictions.

The analysis found that the higher volume hospitals tended to operate on younger patients. In addition, the number of years of experience that a hospital had in performing PD was also associated with a lower probability of inpatient mortality. The results in this study indicate that both increased procedure volume and increased experience were associated with lower mortality rates for patients undergoing the Whipple procedure. High volume rather than experience was associated with marked reductions in inpatient mortality that were statistically significant. As the reduced mortality could be a consequence of more younger patients being operated on in the higher volume hospitals, it is important to include the patient's age in the model to account for it.

6.4 Ninety day mortality following pancreatoduodenectomy in England

In this section we present the first study to the author's knowledge that studies the impact of centre volume on the ninety-day mortality following pancreatoduodenectomy in England. We present the statistical models and data used in the study, the performed statistical analyses and some of the most important results. The reader can refer to the paper [Liu et al., 2018] for a more detailed list of the results.

6.4.1 Statistical models

We briefly outline the models we used in studying the long-term survival patterns following pancreatoduodenectomy.

One of the survival models we use is **Cox proportional hazards model** (Cox PH) [Cox, 1992]. This is a linear survival model, modelling the instantaneous rate at which some event, such as death or tumour progression, occurs at time T given that the event has not yet occurred at time $t < T$. The function modelling this is known as the *hazard function*, taking the form

$$\lambda(t|\mathbf{z}) = \mathrm{lim}_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t, \mathbf{z})}{\delta t}, \tag{6.1}$$

where the numerator is the probability that, given that the event has not occurred before time t, the event will not occur before time $t + \delta t$. This is a general definition applicable to all survival models, where $\mathbf{z}$ is a vector of covariates applying to some individual. In the case of the Cox proportional hazards model, the hazard function takes the particular form

$$\lambda(t|\mathbf{z}) = \lambda_0(t) \exp(\mathbf{z}^\intercal \beta), \tag{6.2}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, which describes how the risk of the event per time unit changes over time at base-line levels of covariates.

In order to use the Cox PH model, we need to make sure that certain assumptions are satisfied. First, we need to ensure that the design of the study is set up so that the mechanisms giving rise to the censoring of the individuals are not related to the probability of the event occurring. The proportional hazards assumption has to be satisfied as well. The survival curves for two strata must have hazard functions that are proportional over time for the assumption to be valid. Due to the proportional hazards assumption we make, the baseline hazard function will turn out to cancel from the analysis and its form will not affect the results. For example, for any two sets of covariates $\mathbf{z}_0$ and $\mathbf{z}_1$, we get that

$$\begin{align} \frac{\lambda(t|\mathbf{z}_1)}{\lambda(t|\mathbf{z}_0)} =& \frac{\lambda_0(t) \exp(\mathbf{z}_1^\intercal \beta)}{\lambda_0(t) \exp(\mathbf{z}_0^\intercal \beta)} \tag{6.3} \\ =& \exp((\mathbf{z}_1^\intercal - \mathbf{z}_0^\intercal)\beta). \tag{6.4} \end{align}$$

We also consider fitting regression models to predict the survival outcomes. We used **stepAIC** [Venables and Ripley, 2013] to select the best fitted regression model. This procedure uses the Akaike Information Criterion (AIC) [Akaike, 1974], which we introduced in Section 1.2.2 in Chapter 1 to decide which model fits the data the best. It does not evaluate the AIC for all models but uses instead a search method that compares models by removing covariates from the model sequentially until there is no improvement in the AIC score.

We applied as well more sophisticated linear models such as the **generalised lin-**

ear model with elastic net penalty to model the survival outcomes, where the following optimisation problem is solved:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}}\Big[\frac{1}{2N}\sum_{i=1}^{N}(y_i+\beta_0-x_i^T\beta)^2+\lambda P_\alpha(\beta)\Big], \tag{6.5}$$

where $P_\alpha(\beta)=\sum_{j=1}^{p}\Big[\frac{1}{2}(1-\alpha)\beta_j^2+\alpha|\beta_j|\Big]$ is the elastic net penalty, $\mathbf{x}$ is the predictor variable and y is the response variable. The elastic net penalty is a compromise between the ridge regression and lasso regression penalties. If $\alpha=0$, then (6.5) is equivalent to ridge regression, which shrinks the coefficients of correlated predictors towards each other, allowing them to borrow strength from each other. If $\alpha=1$, then (6.5) corresponds to lasso regression, which is indifferent to very correlated predictors and will tend to pick one and ignore the rest. As α increases from 0 to 1, for a given λ, the sparsity to a solution of the elastic net problem increases monotonically from 0 to the sparsity of the lasso solution.

Random Survival Forests

Ishwaran et al. [2008] extended random forests to the setting of right-censored survival data by introducing **random survival forest**. The model does not rely on restrictive assumptions such as proportional hazards and does not use parameters. Random survival forests follow the principle outlined by Breiman [2001], which requires that all aspects of growing a random forest take into account the outcome. The splitting criterion used in growing a tree must explicitly involve survival time and censoring information.

A good split for a node maximises survival difference between the resulting two branches. The best split for a node is found by searching over all possible x variables and split values c, and choosing that x^* and c^* that maximise survival difference. By maximising survival difference, the tree pushes dissimilar cases apart. Eventually, as the number of nodes increase, and dissimilar cases become separated, each node in the tree becomes homogeneous and is populated by cases with similar survival.

We implemented the model using R package randomForestSRC (`https://github.com/kogalur/randomForestSRC`).

Performance measures

In order to assess how well the models above describe the data, we need appropriate metrics to measure their performance.

We use **concordance index** (C-index) [Harrell et al., 1996; Pencina and D'Agostino, 2004] to assess the accuracy of a survival model. C-index is defined as the proportion of patients in which predictions and outcomes are concordant, i.e. the number of pairs of patients with predicted survival times correctly ordered among all survival times that can actually be ordered. Hence, a C-index of 1 means perfect rank-ordered prediction accuracy, whereas a C-index of 0.50 is as good as a random predictor. The concordance index was calculated in R using the concordance.index function, part of the package 'survcomp' [Schröder et al., 2011].

We use **receiver operating characteristic curve** (ROC curve) [Zweig and Campbell, 1993; Mason and Graham, 2002; Fawcett, 2006; Powers, 2011] as well. It is a graphical plot that illustrates the performance of a binary classifier as its discrimination threshold is varied. There are four possible outcomes from a binary classifier:

- *true positive* - if the outcome from a prediction is positive and the actual value is **p**;
- *false positive* - if the outcome from a prediction is positive and the actual value is **n**;
- *true negative* - if both the prediction and the actual value are **n**;
- *false negative* - if the prediction is **n** and the actual value is **p**.

The ROC curve is constructed by plotting the True Positive Rate (also called *sensitivity*) versus the False Positive Rate (equal to 1-*specificity*). The true positive rate is defined as

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \tag{6.6}$$

where the false positive ratio is defined as

$$\frac{\text{false positive}}{\text{false positive} + \text{true positive}} \tag{6.7}$$

An example of some realisations of receiver operating characteristic together with their interpretations is provided on Figure 6.3.

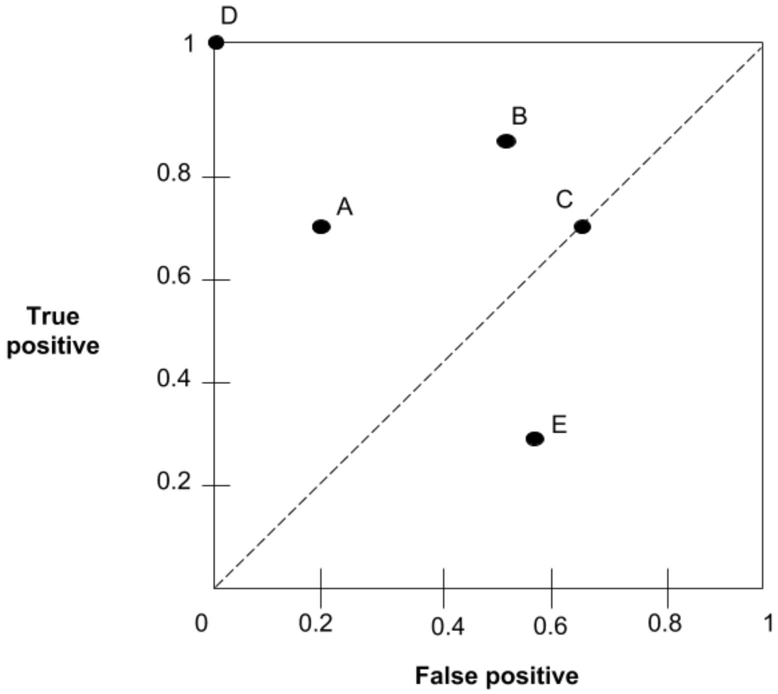


Figure 6.3: An example for the realisations of receiver operating characteristic for different settings. Point D corresponds to perfect classification, point C - to random guessing, point E - to worse than random guessing. Predictions associated with point A are more conservative in comparison with those associated with point B.

The ROC curves in this thesis were plotted using the roc function from the ‘pROC’ package [Robin et al., 2018] in R.

6.4.2 Data

In this study we used data from the Hospital Episode Statistics (HES) database in England. The HES database contains details about all admissions, accident and emergency attendances and outpatient appointments at NHS hospitals in England [NHS Digital, 2018]. Mortality data were provided by the Office for National Statistics and captured in and out of hospital deaths. The data contained information about patients aged 18 or over who underwent PD between April 2001 and March 2016. We removed patients with incomplete information as well any patients with length of stay of four days or less since they were likely to be the result of miscoding.

Following the analyses in the studies presented in Sections 6.2 and 6.3, we chose to include information about the patient's gender, age (categorised based on quartiles [18-59, 60-65, 77-72 and 73-90]), ethnicity (white, Asian, black, Chinese and other, mixed, and unknown), Charlson comorbidity index (categorised into three groups [0, 1-4 and 5+, with 0 corresponding to the healthiest group, and 5+ - to the group with the most existing comorbidities], the index of multiple deprivation (IMD) (categorised into 5 groups, with 1 being the most deprived and 5 being the least deprived), the year of treatment (the fifteen-year period was divided into five 3-year periods) and centre volume.

We followed the definitions used in the Dutch study performed by de Wilde et al. [2012] and grouped the centres initially in the following categories: < 5, $5 - 10$, $11 - 20$ and > 20 PD per year. As there was no plateau of 90-day mortality by centre volume (see Figure 6.5), the centre volumes were based upon quartiles of PD performed per year (very low - ≤ 3, low - 4 to 15, medium - 16 to 35 and high volume > 35 PD per year). Very high volume centres (> 60 per year) were then defined as the top decile of centres for volume.

The outcome of interest in this study was defined to be death from any cause within 90 days of the date of PD.

6.4.3 Statistical analysis

Univariate analysis using log-rank test was performed to determine the variables to be included in the multivariate model. A Cox proportional hazards model [Cox, 1992] was then fitted. The model was fitted using backward stepwise selection.

6.4.4 Results

The goals of this study were to determine whether there were any differences between the groups operated in the different volume centres; whether there were significant differences between the characteristics of the alive and dead patients at 90 days, and if mortality rates changed over the 15-year period. We also investigated whether the mortality rate and centre volume were correlated, and a potential difference in the survival of the patients operated at the high (> 35 PDs per year) and at the very high volume (> 60 PDs per year) centres. We tried to determine as well the factors related to 90-day mortality.

We present a summary of the main results in the next few subsections.

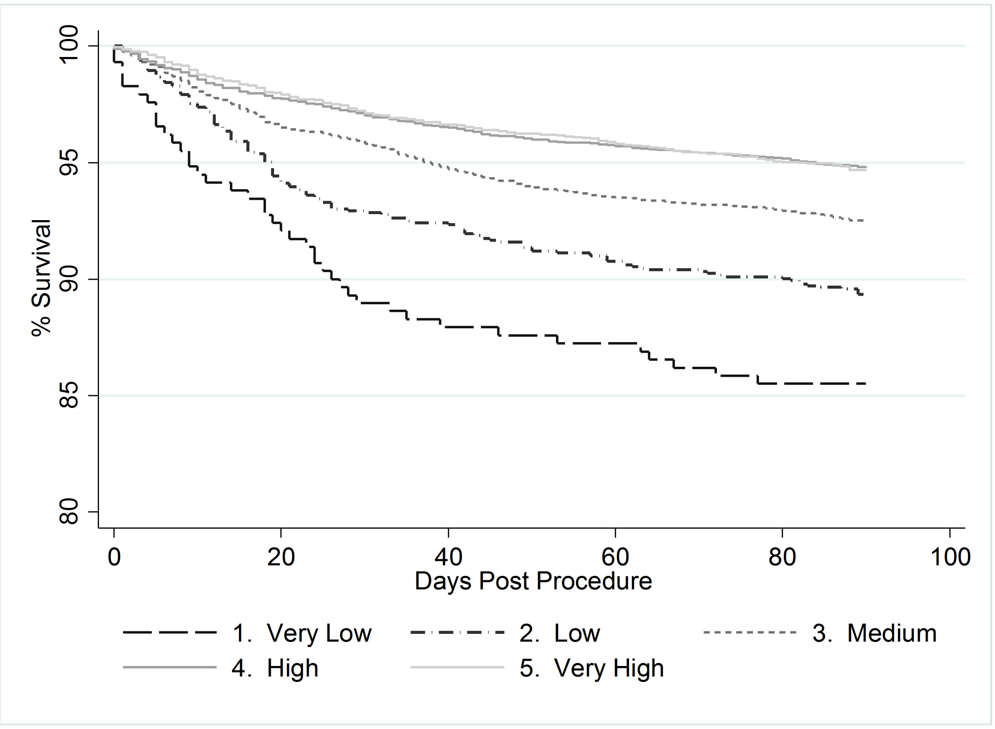


Figure 6.4: Survival analysis of the whole cohort grouped by centre volume (very low ($\leq$ 3), low (4-10), medium (16-35), high (36-60) and very high ($>$ 60) volume centres of PDs per year

Mortality following PD

At the last follow-up 8456 (56.6%) patients had died and 970 (6.5%) patients had died within 90 days of PD. The 30-day and in-hospital mortality rates were 3.7% (551) and 4.7% (700) respectively. The characteristics of the patients grouped by survival status at 90 days following PD are presented in Table 6.2 and Figure 6.4. Age, Charlson score, diagnosis, ethnicity and centre volume all varied significantly on univariate analyses with regards to 90-day mortality.

		Alive at 90 days ($N = 13965$)	Dead at 90 days ($N = 970$)	p-value
Gender	Male	7806 (93.2%)	567 (6.8%)	$p = 0.0121$
	Female	6159 (93.9%)	403 (6.1%)	
Age group	18-59	4559 (96.3%)	177 (3.7%)	$p < 0.001$
	60-65	2765 (94.1%)	172 (5.9%)	
	66-72	3565 (93.3%)	256 (6.7%)	
	73-90	3076 (89.4%)	365 (10.6%)	
Deprivation	1	2419 (93.3 %)	174 (6.7%)	$p = 0.057$
	2	2596 (93.7%)	175 (6.3%)	
	3	2794 (93.2%)	203 (6.8%)	
	4	3173 (93.2%)	230 (6.8%)	
	5	2976 (94.1%)	187 (5.9%)	
Charlson score	0	7878 (94.6%)	448 (5.4%)	$p < 0.001$
	1-4	2071 (94.2%)	127 (5.8%)	
	5+	4016 (91.0%)	395 (9.0%)	
Ethnicity	White	11416 (93.9%)	743 (6.1%)	$p < 0.001$
	Asian	308 (93.9%)	20 (6.1%)	
	Black	176 (95.7%)	2 (4.3%)	
	Other	161 (94.7%)	9 (5.3%)	
	Unknown	1859 (90.8%)	188 (9.2%)	
Volume (PD p.a.)	very low (≤ 3)	248 (85.5%)	42 (14.5%)	$p < 0.001$
	low (4-15)	1200 (89.4%)	143 (10.6 %)	
	medium (16-35)	3723 (92.5 %)	300 (7.5%)	
	high (36-60)	4103 (94.8 %)	225 (5.2%)	
	very high ($>$ 60)	4691 (94.7%)	260 (5.3 %)	
Diagnosis	pancreas cancer	6123 (93.1%)	451 (6.9 %)	$p = 0.048$
	ampullary cancer	2156 (94.8%)	118 (5.2%)	
	cholangio-carcinoma	1432 (94.0%)	92 (6.0%)	
	duodenal cancer	679 (92.3%)	57 (7.7%)	
	other malignant	1710 (93.8%)	113 (6.2%)	
	benign	1865 (93.1%)	139 (6.9%)	

Table 6.2: Summary of the cohort, tested variables as part of univariate analysis and the differences between alive and dead patients at 90 days, following PD

Table 6.2 shows that 90-day mortality has reduced over time. The highest mortality was seen in the first time period (2001-4, 10.0%) with mortality falling sequentially until the most recent period (2013-16, 4.1%).

Time period	very low (≤ 3)	low ($4-15$)	medium ($16-35$)	high ($36-60$)	very high (> 60)
2001/04	26 of 141	73 of 722	74 of 732	17 of 309	12 of 126
2004/07	4 of 76	55 of 453	83 of 1168	29 of 491	37 of 445
2007/10	8 of 37	13 of 126	72 of 992	65 of 832	84 of 1375
2010/13	* of 24	* of 33	43 of 654	62 of 1411	65 of 1279
2013/16	* of 12	0 of 9	28 of 477	52 of 1285	62 of 1726

Table 6.3: 90-day mortality following PD in relation to centre volume and time period. * indicates that the number of patients is so small ($n < 5$) that there is potential for patient identification and thus data is not presented in line with the accepted principles of data reporting from these databases.

Centre volume

The 90-day mortality rates in the highest volume centres were significantly lower than the rates in the lowest volume centres (p-value = 0.001, Table 6.2). In addition, the mortality rates have lowered following the introduction of centralisation as we can see from Table 6.3, which summarises the deaths following PD in the different volume centres over the period 2001-2016, which has been divided in 5 sub-periods.

The highest 90-day mortality rate is observed in the very low volume centres (14.5%). It is worth noting that the mortality rates are similar for the high and very high volume centres over the whole study period (5.2% and 5.3% respectively), which might be due to the skewed data for the very high volume centres between 2001 and 2007. The lowest 90-day mortality rate was observed during 2013 and 2016.

Figure 6.5 illustrates the relationship between centre volume and 90-day mortality. During the early period, before the centralisation, the higher volume centres were associated with lower mortality rates in comparison with the lower volume centres. Interestingly, the highest volume centres in the early period appear to be associated with higher rates of mortality than neighbouring lower volume centres. In the later period ($2009-2016$), the higher volume centres are similarly associated with lower mortality rates; in particular, we observe that the very high volume centres are associated with a further reduction in the 90-day mortality rates.

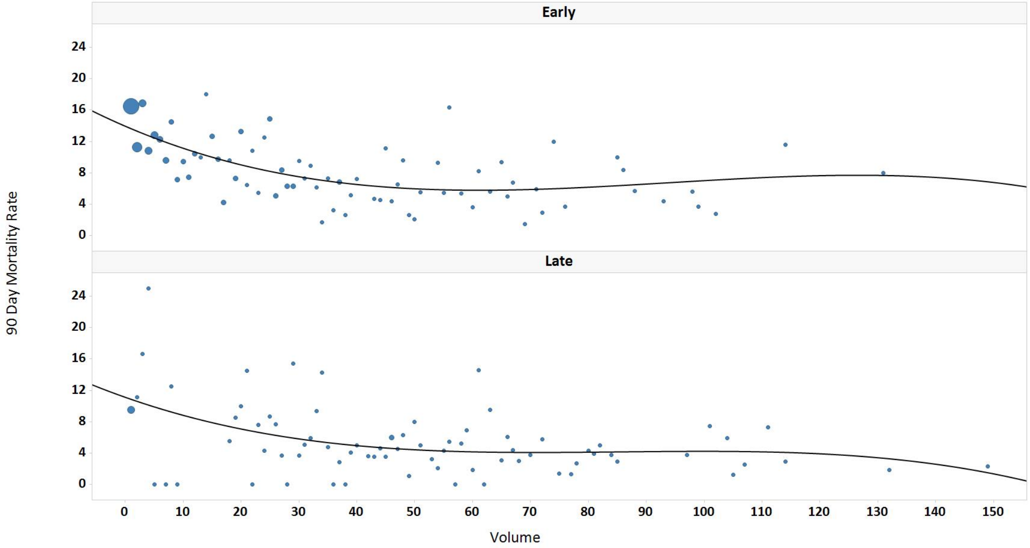


Figure 6.5: The relationship between centre volume and 90-day mortality between early/before centralisation (2001-8) and late/after centralisation (2009-2016) periods. The increase in the centre volume is associated with a decrease in the 90-day mortality. We observe plateaus during both time periods.

Factors related to 90-day mortality

A Cox proportional hazards model was fitted to assess the relationship between the variables we selected and the 90-day mortality (Table 6.4).

The different age groups have statistically significant differences in survival (see Table 6.4) with the oldest group ($73 - 90$) having the worst survival.

Although the index of multiple deprivation was not significant in the univariate analysis, there are significant differences in the survival of patients from different social groups. The patients from the least deprived social group had better 90-day survival relative to the most deprived group (p-value $= 0.022$, Table 6.4).

Patients undergoing resection for ampullary carcinoma [13] have better survival outcome compared to those undergoing resection for pancreatic cancer (p-value $=$

[13] carcinoma that forms in the ampulla of Vater. The ampulla of Vater is a small opening that enters into the first portion of the small intestine, and is the spot where the pancreatic and bile ducts release their secretions into the intestines

		HR	95% CI	p-value
Gender	Male	1		
	Female	0.90	0.79, 1.02	0.095
Age group	18-59	1		
	60-65	1.66	1.34, 2.05	< 0.001
	66-72	1.95	1.60, 2.37	< 0.001
	73-90	3.30	2.74, 3.97	< 0.001
Deprivation	1	1		
	2	0.89	0.72, 1.10	0.299
	3	0.91	0.74, 1.12	0.361
	4	0.91	0.75, 1.12	0.381
	5	0.78	0.64, 0.97	0.022
Charlson score	0	1		
	1-4	1.12	0.92, 1.37	0.251
	5+	1.79	1.56, 2.06	< 0.001
Centre volume	very low	1		
	low	0.70	0.50, 0.99	0.046
	medium	0.58	0.41, 0.80	0.001
	high	0.45	0.32, 0.63	< 0.001
	very high	0.44	0.31, 0.63	< 0.001
Period	2001/04	1		
	2004/07	0.83	0.68, 1.02	0.074
	2007/10	0.80	0.65, 0.99	0.037
	2010/13	0.55	0.44, 0.70	< 0.001
	2013/16	0.45	0.35, 0.58	< 0.001
Diagnosis	pancreas cancer	1		
	ampullary cancer	0.73	0.60, 0.89	0.002
	cholangio-carcinoma	1.11	0.90, 1.37	0.331
	duodenal cancer	1.15	0.87, 1.52	0.316
	other malignant	0.96	0.77, 1.21	0.753
	benign	1.36	1.12, 1.66	0.002

Table 6.4: Multivariate analysis using Cox PH model of factors related to 90-day mortality. The variables were one-hot encoded.

0.002), whereas the survival of the patients undergoing resection for benign disease have worse survival outcome (p-value = 0.002).

6.5 Long-term survival following pancreatoduodenectomy in England

6.5.1 Data

In this study we used data from patients who underwent PD between April 2001 and March 2014 from the Hospital Episode Statistics (HES) database in England, together with mortality data from the Office for National Statistics. We preprocessed the data in similar way to the 90-day survival study and we were interested in studying the factors related to 2-year survival.

6.5.2 2-year survival

We begin the analysis by looking into the difference in the survival outcome of different patient groups. We compared the survival of male and female patients using Cox proportional hazards model and tested the null hypothesis that they come from the same distribution. The p-value of 0.237 suggests that both female and male patients have similar 2-year survival patterns.

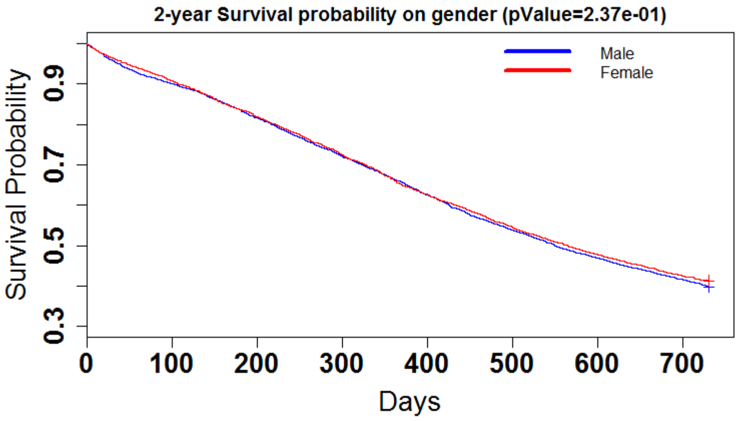


Figure 6.6: Kaplan-Meier curves comparing the 2-year survival of male and female pancreatic cancer patients.The survival patterns are not statistically different

We also looked into how the patient's age affects their long-term survival. We used the same age subgroups as in the short-term survival analysis: between 18 and 59, between 59 and 66, between 66 and 72, and over 72 years old. As we can see from Figure 6.7, the 4 age groups differ in their long term survival (p-value of $7.48e-19$), with the youngest group (18-59) having the best survival and the oldest group (above 72) having the worst survival.

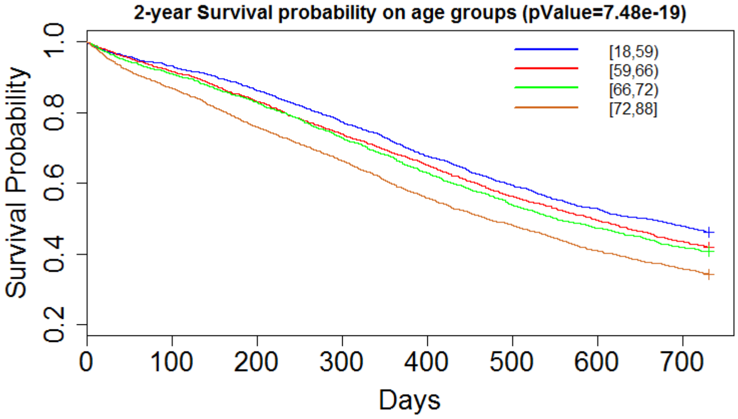


Figure 6.7: Kaplan-Meier curves comparing the 2-year survival of the different age bands.The survival patterns are statistically different

Next we checked how the presence of comorbidities affected the long-term survival. Similarly to the 90-day survival analysis, we divided the patients into 4 groups based on their Charlson score: a group with the fewest present comorbidities (0-5), a group with low number of present comorbidities (5-10), a group with medium to high number of comorbidites (10-20), and a group with very high number of comorbidities (over 20). Figure 6.8 shows that the group with the highest number of existing comorbidities had the worst survival outcome: approximately half of the patients die within one year after pancreatic cancer resection. The group with fewest comorbidities had the best 2-year survival.

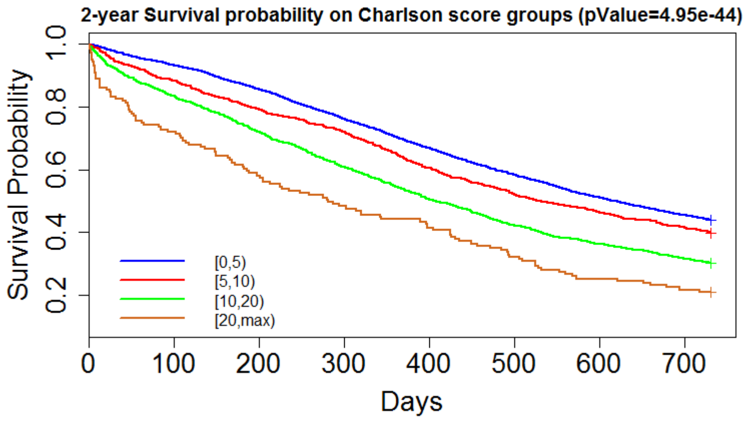


Figure 6.8: Kaplan-Meier curves comparing the 2-year survival of pancreatic cancer patients from different Charlson score groups.The survival patterns are statistically different with the patients with the highest Charlson score having the lowest survival rate

Index of multiple deprivation (IMD) could be an important factor affecting the long-term patient survival as it incorporates information about the patient access to health institutions, living environment conditions and existing barriers to houses and services [Department for Communities and Local Government, 2018]. We divided the patients in 5 groups, corresponding to IMD $= \{1, 2, 3, 4, 5\}$, with IMD$=1$ being the most deprived group and IMD$=5$ being the least deprived group. Figure 6.9 demonstrates that the areas where the patients lived had an affect on their survival - the patients from the most deprived regions had the worst survival which could be due to the lack of easy access to hospitals.

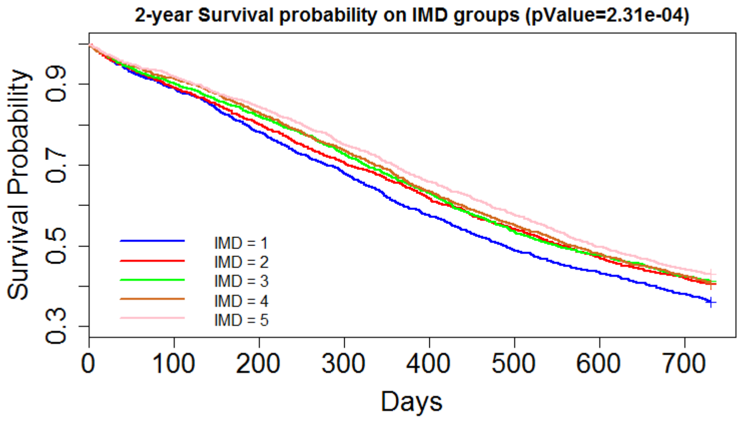


Figure 6.9: Kaplan-Meier curves comparing the 2-year survival of patients from different IMD groups.The survival patterns are statistically different with the patients from the lowest IMD group (the most deprived patients) having the lowest survival rate

We finish this analysis by investigating whether there was any difference between the survival of patients operated in low volume centres and the survival of patients operated in the other centres combined. Figure 6.10 illustrates that the survival of the two groups differ in the first year but in the second year after surgery, this difference becomes less pronounced.

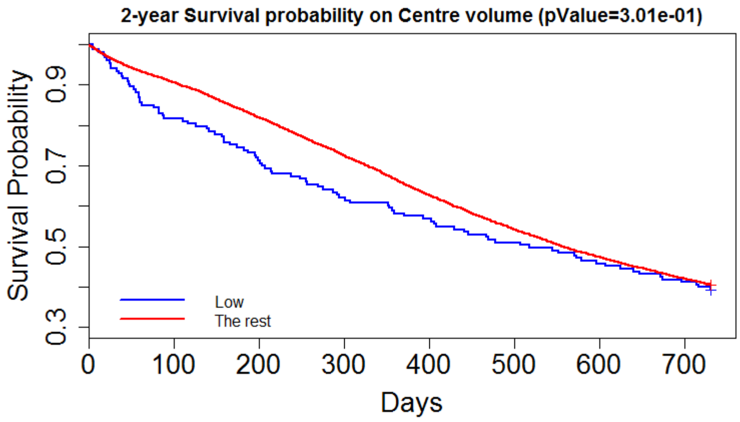


Figure 6.10: Kaplan-Meier curves comparing the 2-year survival of patients operated in low-volume centres and non-low-volume centres.The survival patterns are not statistically different

We present a summary of the results from the univariate analysis in Table 6.5:

Variable		**p-value**
Gender	Male	0.232
	Female	
Age group	[18,59)	< 0.001
	[59,66)	
	[66,72)	
	[72,88]	
Charlson score	[0,5)	< 0.001
	[5,10)	
	[10,20)	
	[20, max)	
IMD	1	< 0.001
	2	
	3	
	4	
	5	
Centre volume	Low	0.301
	the rest	

Table 6.5: Univariate analysis of factors related to 2-year mortality.

Model comparison

We also looked into using different models to study the factors affecting long-term survival. We compared the predictions from Cox proportional hazards model with the results from stepAIC and glmnet using concordance index (Figure 6.11). We included gender, ethnicity, IMD and centre volume as categorical variables, and age at diagnosis and Charlson score as continuous variables. We performed 5-fold cross-validation. The Cox model found the age at diagnosis, IMD, Charlson score and centre volume to be significant, whereas the stepAIC model found the age at diagnosis, IMD, Charlson score and ethnicity to be significant. The models output similar results as we can see from Figure 6.11, and although their predictions were better than random guesses, the accuracy of the survival models was low.

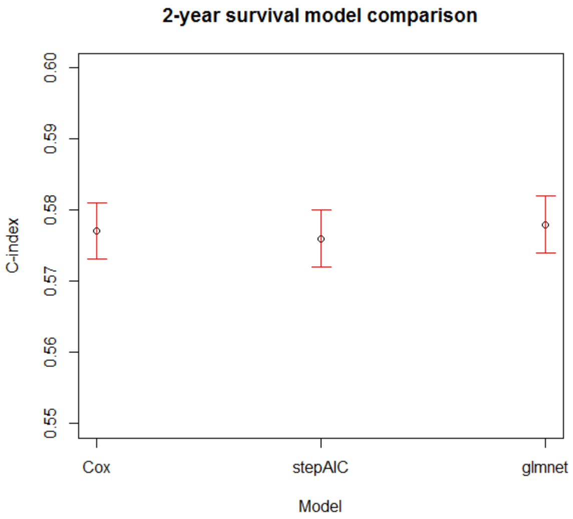


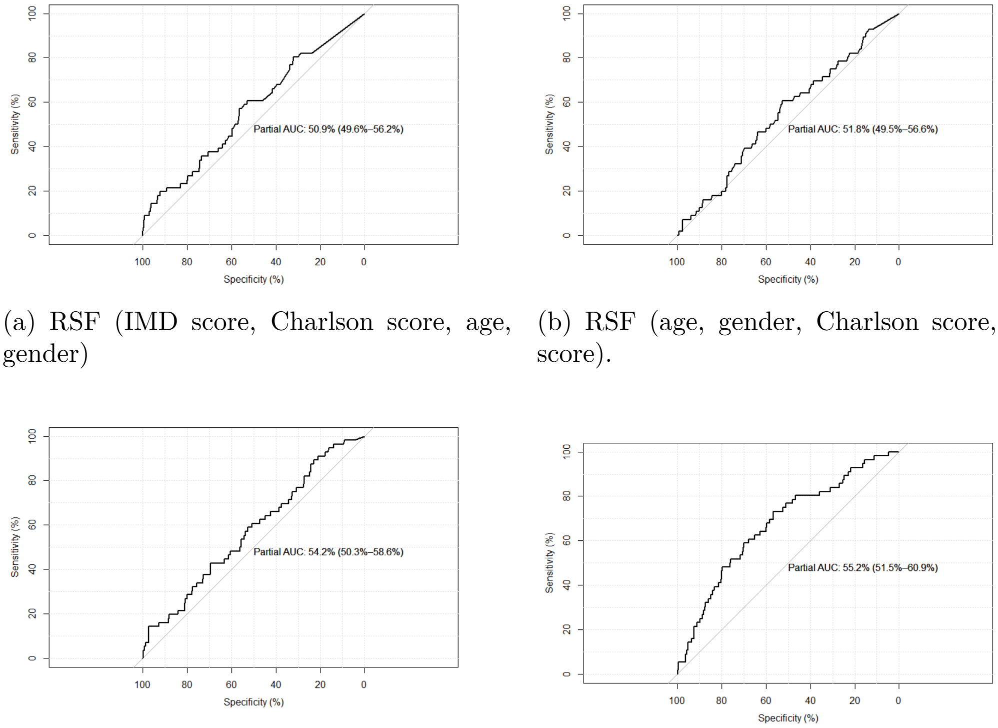
Figure 6.11: Comparison of the different models using concordance index. The mean C-index ± one standard deviation is plotted for Cox PH model, stepAIC model and generalised linear model with elastic net penalty. Cox PH model found age at diagnosis, IMD, Charlson score and centre volume to be significant; stepAIC found age at diagnosis, IMD, Charlson score and ethnicity to be significant.

6.5.3 Predictive models for 90-day survival

We were interested as well whether we could predict the 90-day survival of patients operated in 2013 and 2014 with models trained on the data about patients operated between 2007 and 2012. We chose to consider the patients after 2007 because centralisatiton of pancreatic resections was introduced then, and using the patients operated before then might confound the results.

We compared the predictions from a regularised generalised linear model and from a random survival forest. We included different covariates in the random survival forest classifier as it does not include feature selection. We used AUC as a measure of how accurate the model predictions were.

The results show that the predictions from every model were not much better than just a random guess (see Figures 6.12a, 6.12b,6.12c, 6.12d). We tried including only a few of the covariates for random survival forest, which have been identified as significant in the previous analyses (see the model with including age, Charlson score, IMD score) but this did not lead to an improvement in the accuracy. The model with the best predictions is the generalised linear model with lasso and elastic net penalty (Figure 6.12d).



(a) RSF (IMD score, Charlson score, age, gender)

(b) RSF (age, gender, Charlson score, IMD score).

(c) RSF (age, Charlson score, centre volume, IMD score)

(d) generalised linear model

Model	**AUC**	**95 % CI**
RSF (age, gender, Charlson score, centre volume, IMD)	50.9 %	49.6 % - 56.2 %
RSF (age, gender, Charlson score, IMD)	51.8 %	49.5 % - 56.6 %
RSF (age, Chalson score, centre volume, IMD)	54.2 %	50.3 % - 58.6 %
GLM	55.2 %	51.5 % - 60.9 %

Table 6.6: Summary of the developed predictive models and the corresponding AUCs and confidence intervals (RSF = random survival forest classifier, GLM = generalised linear model)

6.6 Conclusions

In this chapter, we studied the impact of different clinical factors on the short and longer term survival of pancreatic cancer patients following pancreatoduodenectomy in England after the introduction of centralisation in 2001. The multivariate analysis showed a steady reduction in the 90-day mortality associated with increasing annual centre volume. In addition, the data demonstrated similar 90-day mortality rates between centres performing 36-60 PD procedures per year and those undertaking > 60 operations. This might indicate that a threshold for the 'ideal' centre volume has been demonstrated which has not been shown in previous studies, in which increasing surgical volume has been associated with a reduction in the post-operative mortality without reaching a plateau. The analysis showed as well that age, index of multiple deprivation and diagnosis type were significant factors for the short-term survival. We also looked into whether data from previous time periods could be used to help predict the outcomes in future periods. Our analysis using data about patients operated between 2007 and 2012 showed that we could not predict accurately the 90-day survival of patients operated in 2013 and 2014.

In the analysis of the long-term survival data, we compared different survival models: Cox proportional hazards model, stepAIC and generalised linear model with elastic net penalty; which identified age, Charlson score and index of multiple deprivation to be significant factors. However, none of the models we used could predict the outcomes very accurately - the mean concordance indices of all of them were between 0.57 and 0.58. The analyses and the model predictions would benefit from the inclusion of more detailed information about the cause of death and the surgical procedures.

Chapter 7

Conclusions

Cancer research is currently undergoing a data revolution. Multi-omics and clinical data of high dimensionality, resolution and accuracy have been rapidly accumulated across multiple cancer projects as part of The Cancer Genome Atlas and the International Cancer Genome Consortium, and have shown potential to offer valuable insights into the complex processes underlying cancer. The data are promising to change how we diagnose, treat and prevent cancer but in order to be able to achieve this, we need to create efficient methods and tools that can help us model these complex, high-dimensional, heterogeneous data appropriately. In this thesis, we focused on the development of efficient Bayesian clustering methods to identify cancer subtypes that are indicative of overall survival and/or response to treatment. Learning more about the differences between the patient groups can lead to earlier diagnosis, better and more personalised treatment, and help identify biomarkers in very aggressive cancers.

As we saw in Chapter 1, a lot of the currently used clustering methods are limited in the type of data they can model, require data transformations, often have computationally expensive inference schemes or require setting the number of clusters manually. This motivated the development of a Bayesian clustering method, called BayesCluster. The model combines the advantages of latent variable models, which provide an efficient lower-dimensional representation of the data, and Bayesian nonparametric models, which offer flexibility.

We highlighted the advantages of BayesCluster over other clustering methods throughout Chapter 2, and some of these include the ability to model mixed type data, a Bayesian inference scheme and learning the number of clusters from the data.

The experiments in Chapter 3 illustrated the applicability of BayesCluster to synthetic and real datasets from different areas: economics, politics, medicine. The tests with synthetic data allowed us to validate and test the model in different scenarios. Although it provided competitive results on the real datasets in comparison with other commonly used methods such as k-means, k-modes, iClusterPlus, there were cases in which BayesCluster allocated only a small part of the observations to the correct cluster. There are many reasons for this: inappropriate model assumptions, uninformative data, or some of the inherent limitations of the Dirichlet process mixture model, which have been shown to often overestimate the number of true clusters [West and Escobar, 1993; Onogi et al., 2011; Miller and Dunson, 2018]. These issues motivated a detailed study of the properties of Dirichlet process mixture model in Chapter 4. We looked into different ways to counteract the tendency of Dirichlet process mixture models to overestimate the true number of clusters and found that by putting a prior on the cluster size which disfavours very small clusters, we were able to get a consistent and accurate estimate of the true number of clusters. We proposed two further extensions of BayesCluster: one based on the idea of split-merge, which helps the inference algorithm escape local maxima, and another based on non-local priors, which has been found to be particularly helpful in the cases of model misspecification and to lead to more interpretable clusters [Fuquene et al., 2016].

In Chapter 5 we applied BayesCluster in the context of data integration of molecular data. We outlined a simple integrative framework which uses the information from multiple data sources to derive a single clustering partition, and demonstrated its ability to easily implementable inference with four case studies using data from TCGA. Using BayesCluster, we were able to discover subtypes which were prognostic of the overall survival in two aggressive types of cancer: pancreatic cancer and glioblastoma, and which we were not able to identify using iClusterPlus or simpler models. In addition, the clinical data helped us investigate the clinical characteristics of the different subtypes, and explain the different survival outcomes.

There are several different ways that we can extend BayesCluster in order to make modelling more complex data possible and get more interpretable results. In this thesis, we considered only a linear mapping from the latent space to the observed space and a simple variable selection method. However, we saw in Chapter 5 that there were cases, for example the breast cancer case study, where BayesCluster could not identify cancer subtypes that were associated with different survival prognosis. One of the avenues we could explore involves incorporating the variable selection in the model by adopting for example, a Bayesian approach and computing the posterior distribution of selecting a particular gene/mutation for further analysis. [Law et al., 2004] and [Tadesse et al., 2005] adopt this approach by including a binary variable φ_j such that $\varphi_j = 1$ if the j^{th} variable is relevant and $\varphi_j = 0$ otherwise. Law et al. [2004] define the quantity $\rho_j = p(\varphi_j = 1)$ as the saliency of the j^{th} feature or the importance of the variable in characterising the cluster structure of the data. They then place a Dirichlet prior on ρ to infer it and identify the relevant features. Tadesse et al. [2005] assume a prior on φ of the form:

$$p(\varphi|\eta) = \prod_{j=1}^{J} \eta^{\varphi_j}(1-\eta)^{1-\varphi_j}, \tag{7.1}$$

where J is the number of features and the hyperparameter η is the proportion of variables expected to discriminate between the clusters. The best clustering variables are then considered to be those with the largest marginal posterior $p(\varphi_j = 1|X) > t$ with a prespecified t. Another option would be to detect signals not only in cluster-specific means but also in cluster specific covariances, in a manner similar to [Taschler et al., 2019]. In this way, we should be able to model more appropriately the cluster variability.

In Chapter 5 we focused on the application of BayesCluster to genomics data. However, the sequencing technologies used in cancer research generate a wider range of data such as spatial proteomics data [1] and ChIP-seq data [2], which could help us understand better the differences between the cancer subtypes. In order to be able to incorporate these datasets in the data integration framework, we need appropriate statistical models for these data types. For example, we can adapt a Hidden Markov model for the ChIP-seq data [Spyrou et al., 2009], and the Bayesian mixture model proposed by Crook et al. [2018] for the spatial proteomics data.

Cancer is a complex disease, driven not only by changes in the genome, but also by environmental factors. Hence, it is important to study the effect of factors such as patient's age, gender, presence of comorbidities, hospital size, access to health services, on the short- and long-term survival following cancer-related surgeries as often these factors are easily modifiable and can improve the survival outcome. In Chapter 6, we presented a pilot study using Hospital Episode Statistics data about pancreatic cancer patients from England who underwent pancreatoduodenectomy, which aimed to investigate the effect of the centralisation on the patient survival.

[1] it is used to study the location of proteins on large scale

[2] it is used to analyse the interactions of the proteins with the DNA.

Our analysis indicated that higher volume surgery centres were associated with lower 90-day mortality rates. In addition, the multivariate analysis showed that age, index of multiple deprivation and diagnosis type were significant risk factors for the short-term survival, whereas age, Charlson score and index of multiple deprivation were important factors for the long-term survival. This study shows promising results that the patient survival for aggressive diseases such as pancreatic cancer could be improved by modifying factors such as centre referral and access to healthcare.

In this thesis, we have seen how the analysis of large complex molecular and clinical datasets coupled with methodology advances can help us understand better what drives the different patient survival. The rapid accumulation of high-dimensional and heterogeneous data requires the creation and use of models that could enable us to extract useful patterns and identify novel patient subtypes to provide personalised treatment and better monitoring. Here we proposed Bayesian methods that take into account the interactions between the different data sources to determine cancer subtypes and thus, offer valuable insights into the biological dynamics of cancer. The increasing availability of precise, detailed molecular and clinical data together with the development of new statistical methods show a great promise of making a more personalised cancer treatment the new standard of care.

Appendix A

Efficient computation of sufficient statistics

A.1 Sufficient statistics for multivariate Normal distribution

We use definitions and propositions from Bernardo and Smith [2001] to derive results for efficient computations of the sufficient statistics for multivariate Normal distribution. We have omitted the proofs of the propositions.

Definition. Given random vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ with specified sets of possible values $X_1, \ldots, X_m$ respectively, a random vector $\mathbf{t}_m : X_1 \times \ldots \times X_m \to \mathbb{R}^{k(m)}$ with $k(m) \leq m$, is called a $k(m)$-dimensional **statistic**.

Some examples of statistic are the sample mean; the sample size, sum and sum of squares.

Proposition *The sequence* $\mathbf{t}_1, \mathbf{t}_2, \ldots,$ *is parametric sufficient for infinitely exchangeable* $\mathbf{x}_1, \mathbf{x}_2, \ldots,$ *if and only if, for any* $m \geq 1$*, the density* $\mathbb{P}(\mathbf{x}_1, \ldots, \mathbf{x}_m | \theta, \mathbf{t}_m)$ *is independent of* θ*.*

Using results from [Bernardo and Smith, 2001], it can be shown that the sufficient statistics of the multivariate Normal distribution are the mean and the covariance matrix.

As the collapsed Gibbs sampling (Chapter 2) requires the update of the sufficient statistics after the removal/addition of an observation to a cluster, we have derived calculations to speed up the step. We provide more details in the following subsections.

A.1.1 Removing an observation from a cluster

Let assume that the cluster consists of N data points $\mathbf{x}_1, \dots, \mathbf{x}_N$, the mean of this cluster is denoted by $\bar{\mathbf{x}}$ and the covariance is $\mathbf{S}$, and the mean of the cluster after the point removal is $\bar{\mathbf{x}}_\star$ and the corresponding covariance is $\mathbf{S}_\star$. We assume for simplicity that the data point we remove is $\mathbf{x}_N$.

Updating the cluster mean

In the case of updating the cluster mean, we have that

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{A.1}$$

$$\bar{\mathbf{x}}_\star = \frac{1}{N-1}\sum_{i=1}^{N-1}\mathbf{x}_i. \tag{A.2}$$

Hence,

$$\begin{aligned}\bar{\mathbf{x}}_\star &= \frac{1}{N-1}(\sum_{i=1}^{N-1}\mathbf{x}_i + \mathbf{x}_N - \mathbf{x}_N) &\text{(A.3)}\\ &= \frac{1}{N-1}(\sum_{i=1}^{N}\mathbf{x}_i - \mathbf{x}_N). &\text{(A.4)}\end{aligned}$$

Therefore, the mean of the new cluster is $\bar{\mathbf{x}}_\star = \frac{1}{N-1}(\sum_{i=1}^{N}\mathbf{x}_i - \mathbf{x}_N)$.

Updating the cluster covariance

In the case of updating the cluster covariance, we have that

$$\mathbf{S}_\star = \frac{1}{N-2}\sum_{i=1}^{N-1}(\mathbf{x}_i - \bar{\mathbf{x}}_\star)^\intercal(\mathbf{x}_i - \bar{\mathbf{x}}_\star) \tag{A.5}$$

$$= \frac{1}{N-2}(\sum_{i=1}^{N-1}\mathbf{x}_i^\intercal\mathbf{x}_i - (N-1)\bar{\mathbf{x}}_\star^\intercal\bar{\mathbf{x}}_\star) \tag{A.6}$$

$$= \frac{1}{N-2}((\sum_{i=1}^{N}\mathbf{x}_i^\intercal\mathbf{x}_i - \mathbf{x}_N^\intercal\mathbf{x}_N) - \frac{1}{N-1}\sum_{i=1}^{N-1}\mathbf{x}_i\sum_{i=1}^{N-1}\mathbf{x}_i). \tag{A.7}$$

A.1.2 Adding an observation to a cluster

Let assume that the cluster consists of N data points $\mathbf{x}_1, \dots, \mathbf{x}_N$, the mean of this cluster is denoted by $\bar{\mathbf{x}}$ and the covariance is $\mathbf{S}$, and the mean of the cluster after the addition of point is $\bar{\mathbf{x}}_{\text{new}}$ and the cluster covariance is $\mathbf{S}_{\text{new}}$. We assume for simplicity that the data point is $\mathbf{x}_{N+1}$.

Updating the cluster mean

In the case of updating the cluster mean, we have that

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{A.8}$$

$$\bar{\mathbf{x}}_{\text{new}} = \frac{1}{N+1}\sum_{i=1}^{N+1}\mathbf{x}_i \tag{A.9}$$

$$= \frac{1}{N+1}(\sum_{i=1}^{N}\mathbf{x}_i + \mathbf{x}_{N+1}). \tag{A.10}$$

Hence, the mean of the new cluster is $\bar{\mathbf{x}}_{\text{new}} = \frac{1}{N+1}(\sum_{i=1}^{N}\mathbf{x}_i + \mathbf{x}_{N+1})$.

Updating the cluster covariance

In the case of updating the cluster covariance, we have that

$$\mathbf{S}_{\text{new}} = \frac{1}{N}\sum_{i=1}^{N+1}(\mathbf{x}_i - \bar{\mathbf{x}}_{\text{new}})^{\intercal}(\mathbf{x}_i - \bar{\mathbf{x}}_{\text{new}}) \tag{A.11}$$

$$= \frac{1}{N}(\sum_{i=1}^{N}\mathbf{x}_i^{\intercal}\mathbf{x}_i + \mathbf{x}_{N+1}^{\intercal}\mathbf{x}_{N+1} - (N+1)\bar{\mathbf{x}}_{\text{new}}^{\intercal}\bar{\mathbf{x}}_{\text{new}}) \tag{A.12}$$

$$= \frac{1}{N}(\sum_{i=1}^{N}\mathbf{x}_i^{\intercal}\mathbf{x}_i + \mathbf{x}_{N+1}^{\intercal}\mathbf{x}_{N+1} - \frac{1}{N+1}\sum_{i=1}^{N+1}\mathbf{x}_i\sum_{i=1}^{N+1}\mathbf{x}_i). \tag{A.13}$$

These results imply that we can efficiently update the sufficient statistics for multivariate Gaussian distribution by caching the sum of the observations and the sum of squares of the observations.

Appendix B

Derivation of model log posterior (Chapter 2)

B.1 Continuous BayesCluster

Using the graphical model in Figure 2.1, we can derive the posterior of the continuous BayesCluster, which we need for the computation of the acceptance probability (after we integrate out the mixing proportions π and assume that α is fixed) as follows:

$$\begin{aligned} p(\mathbf{Z}, \mathbf{W}, \epsilon, \boldsymbol{\mu}, C|\mathbf{X}, \alpha) &= \frac{p(\mathbf{X}, \mathbf{Z}, \mathbf{W}, \epsilon, \boldsymbol{\mu}, C|\alpha)}{p(\mathbf{X}|\alpha)} \\ &\propto p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \epsilon) p(\mathbf{Z}|\boldsymbol{\mu}, C) p(\mathbf{W}) \\ &= \Big[\prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I})\Big[\prod_{i=1}^{N} \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I})\Big]\Big[\prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_d|\mathbf{0}, \mathbf{I})\Big] \\ &\times \Big[\prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I})\Big]\Big[\prod_{i=1}^{N} \mathrm{Cat}(C|\alpha)\Big]. \end{aligned}$$

Hence, we can write the log posterior as follows:

$$
\begin{aligned}
\log p(\mathbf{Z}, \mathbf{W}, \epsilon, \boldsymbol{\mu}, C|\mathbf{X}, \alpha) = & \Big[\sum_{i=1}^{N} \log \mathcal{N}(\mathbf{x}_i|\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}) + \Big[\sum_{i=1}^{N} \log \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I})\Big] \\
& + \Big[\sum_{d=1}^{D} \log \mathcal{N}(\mathbf{w}_d|\mathbf{0}, \mathbf{I})\Big] + \Big[\sum_{k=1}^{K} \log \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I})\Big] \\
& + \Big[\sum_{i=1}^{N} \log \mathrm{Cat}(C|\alpha)\Big].
\end{aligned}
$$

B.2 Discrete BayesCluster

Using the graphical model in Figure 2.2, we can derive the posterior of the discrete BayesCluster, which we need for the computation of the acceptance probability (after we integrate out the mixing proportions π and assume that α is fixed) as follows:

$$
\begin{aligned}
p(\mathbf{Z}, \mathbf{W}^D, \mathbf{w}_0^D, \boldsymbol{\mu}, c|\mathbf{X}, \alpha) = & \frac{p(\mathbf{X}, \mathbf{Z}, \mathbf{W}^D, \mathbf{w}_0^D, \boldsymbol{\mu}, c|\alpha)}{p(\mathbf{X}|\alpha)} \\
\propto & p(\mathbf{X}, \mathbf{Z}, \mathbf{W}^D, \mathbf{w}_0^D, \boldsymbol{\mu}, c|\alpha) \\
= & p(\mathbf{X}|\mathbf{Z}, \mathbf{W}^D, \mathbf{w}_0^D)p(\mathbf{Z}|\boldsymbol{\mu}, c)p(\mathbf{W}^D)p(\mathbf{w}_0^D)p(\boldsymbol{\mu})p(c|\alpha).
\end{aligned}
$$

Hence, we can write the log posterior as follows:

$$\begin{aligned}
\log p(\mathbf{Z}, \mathbf{W}^D, \mathbf{w}_0^D, \boldsymbol{\mu}, c|\mathbf{X}, \alpha) =& \log \prod_{i=1}^{N} \prod_{r=1}^{R} \mathrm{Cat}(x_{ir}|\mathcal{S}(\mathbf{W}^\intercal \mathbf{z}_i + \mathbf{w}_{0r}) + \log \prod_{r=1}^{R} \prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_{dr}|\mathbf{0}, \mathbf{I}) \\
&+ \log \prod_{r=1}^{R} \mathcal{N}(\mathbf{w}_{0r}) + \log \prod_{i=1}^{N} \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \\
&+ \log \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I}) + \log \prod_{i=1}^{N} \mathrm{Mult}(c_i|\alpha) \\
=& \sum_{i=1}^{N} \sum_{r=1}^{R} \log \mathrm{Cat}(x_{ir}|\mathcal{S}(\mathbf{W}^\intercal \mathbf{z}_i + \mathbf{w}_{0r}) + \sum_{r=1}^{R} \sum_{d=1}^{D} \log \mathcal{N}(\mathbf{w}_{dr}|\mathbf{0}, \mathbf{I}) \\
&+ \sum_{r=1}^{R} \log \mathcal{N}(\mathbf{w}_{0r}) + \sum_{i=1}^{N} \log \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \\
&+ \sum_{k=1}^{K} \log \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I}) + \log \Gamma(\alpha) - \log \Gamma(\alpha + N) \\
&+ \sum_{k=1}^{K} \Big(\log(\Gamma(N_k + \frac{\alpha}{K})) - \log(\Gamma(\frac{\alpha}{K})) \Big)
\end{aligned}$$

B.3 Mixed BayesCluster

Using the graphical model in Figure 2.3, we can derive the posterior of the mixed BayesCluster, which we need for the computation of the acceptance probability (after we integrate out the mixing proportions π and assume that α is fixed) as follows:

$$\begin{aligned}
p(\mathbf{Z}, \mathbf{W}^D, \mathbf{w}_0, \mathbf{W}^C, \epsilon, \boldsymbol{\mu}, C|\mathbf{X}, \alpha) =& \frac{p(\mathbf{X}, \mathbf{Z}, \mathbf{W}^D, \mathbf{w}_0, \mathbf{W}^C, \epsilon, \boldsymbol{\mu}, C|\alpha)}{p(\mathbf{X}|\alpha)} \\
&\propto p(\mathbf{X}|\mathbf{Z}, \mathbf{W}^D, \mathbf{w}_0, \mathbf{W}^C, \epsilon) p(\mathbf{Z}|\boldsymbol{\mu}, C) p(\mathbf{W}^D) p(\mathbf{W}^C) \\
&\times p(\mathbf{w}_{0D}) p(\epsilon) p(\boldsymbol{\mu}) \\
&= \Big[\prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i^C|\mathbf{W}^C \mathbf{z}_i, \sigma^2 \mathbf{I}) \prod_{r=1}^{R} \mathrm{Cat}(x_{ir}^D|\mathcal{S}(\mathbf{W}_r^\intercal \mathbf{z}_i + \mathbf{w}_{0r})) \Big] \\
&\times \Big[\prod_{i=1}^{N} \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I}) \Big] \Big[\prod_{r=1}^{R} \prod_{j=1}^{J_r} \mathcal{N}(\mathbf{w}_{jr}|\mathbf{0}, \mathbf{I}) \Big] \Big[\prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_d^C|\mathbf{0}, \mathbf{I}) \Big] \\
&\times \Big[\prod_{r=1}^{R} \mathcal{N}(\mathbf{w}_{0r}|\mathbf{0}, \mathbf{I}) \Big] \Big[\prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I}) \Big] \Big[\prod_{i=1}^{N} \mathrm{Cat}(C|\alpha) \Big].
\end{aligned}$$

Hence, we can write the log posterior as follows:

$$
\begin{aligned}
\log p(\mathbf{Z}, \mathbf{W}^{D}, \mathbf{w}_0, \mathbf{W}^{C}, \epsilon, \boldsymbol{\mu}, C|\mathbf{X}, \alpha) = \Big[&\sum_{i=1}^{N} \sum_{r=1}^{R} \log \mathcal{N}(\mathbf{x}_i^{C}|\mathbf{W}^{C}\mathbf{z}_i, \sigma^2\mathbf{I}) \mathrm{Cat}(x_{ir}^{D}|\mathcal{S}(\mathbf{W}_r^{\intercal}\mathbf{z}_i + \mathbf{w}_{0r}))\Big] \\
&+ \Big[\sum_{i=1}^{N} \log \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_k, \mathbf{I})\Big] + \Big[\sum_{r=1}^{R} \sum_{j=1}^{J_r} \log \mathcal{N}(\mathbf{w}_{jr}|\mathbf{0}, \mathbf{I})\Big] \\
&+ \Big[\sum_{d=1}^{D} \log \mathcal{N}(\mathbf{w}_d^{C}|\mathbf{0}, \mathbf{I})\Big] \\
&+ \Big[\sum_{r=1}^{R} \log \mathcal{N}(\mathbf{w}_{0r}|\mathbf{0}, \mathbf{I})\Big] + \Big[\sum_{k=1}^{K} \log \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{0}, \mathbf{I})\Big] \\
&+ \Big[\sum_{i=1}^{N} \log \mathrm{Cat}(C|\alpha)\Big]
\end{aligned}
$$

Appendix C

Determining the number of clusters Chapter 3

C.1 Continuous data

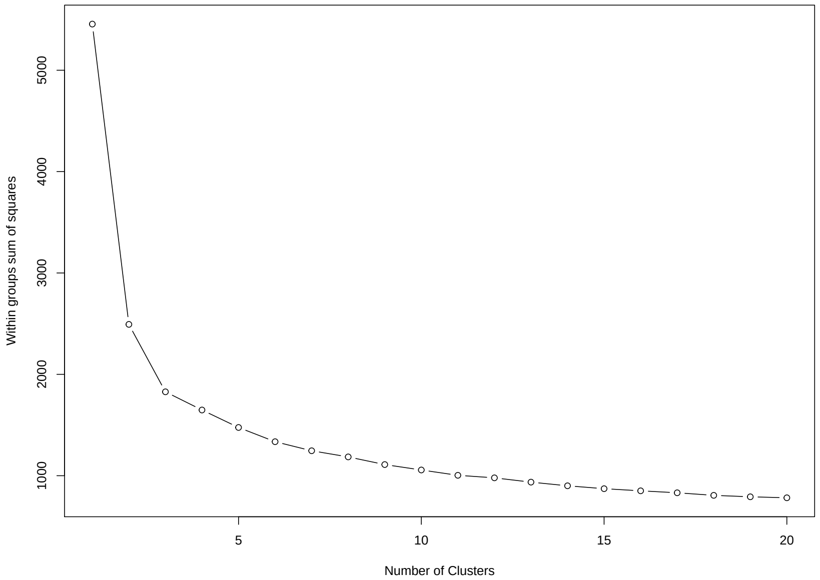


Figure C.1: Determining the number of clusters K for k-means clustering ($K = 3$)

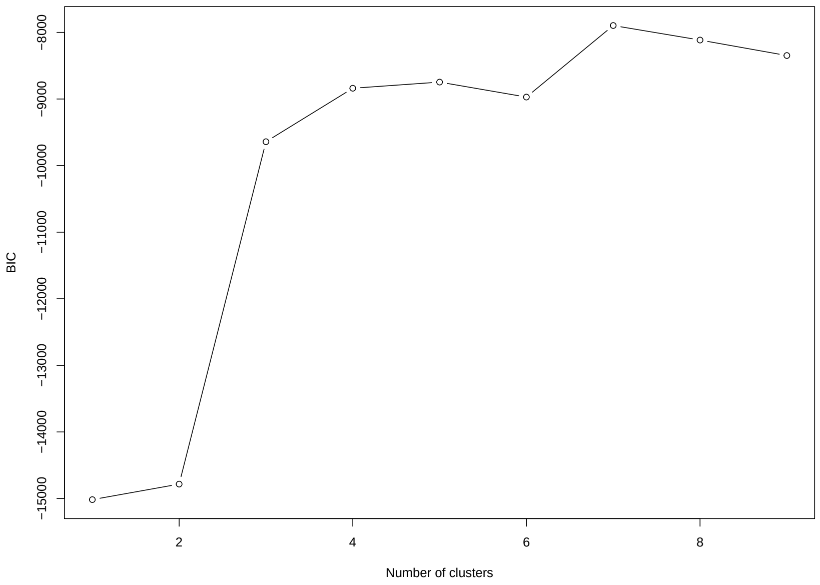


Figure C.2: Determining the number of clusters K for Gaussian mixture model ($K = 7$)

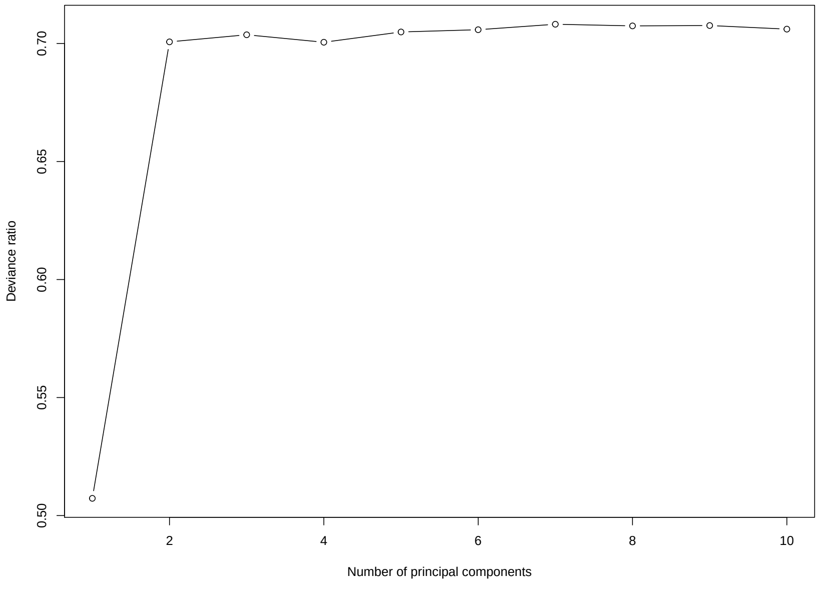


Figure C.3: Determining the latent dimensionality P in the case of the iClusterPlus model for the dataset presented on Figure 3.2iClusterPlus ($P = 2$)

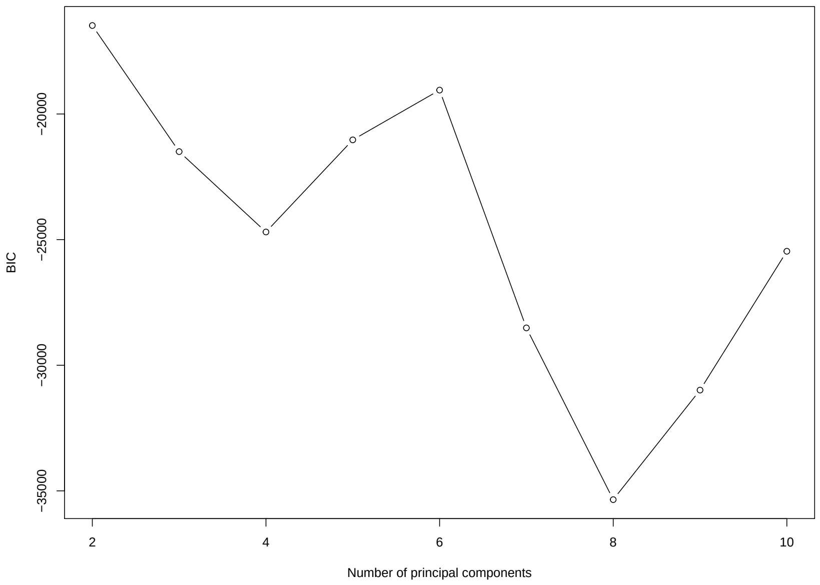


Figure C.4: Determining the latent dimensionality P in the case of BayesCluster for the dataset presented on Figure 3.2 ($P = 2$)

C.2 Discrete data

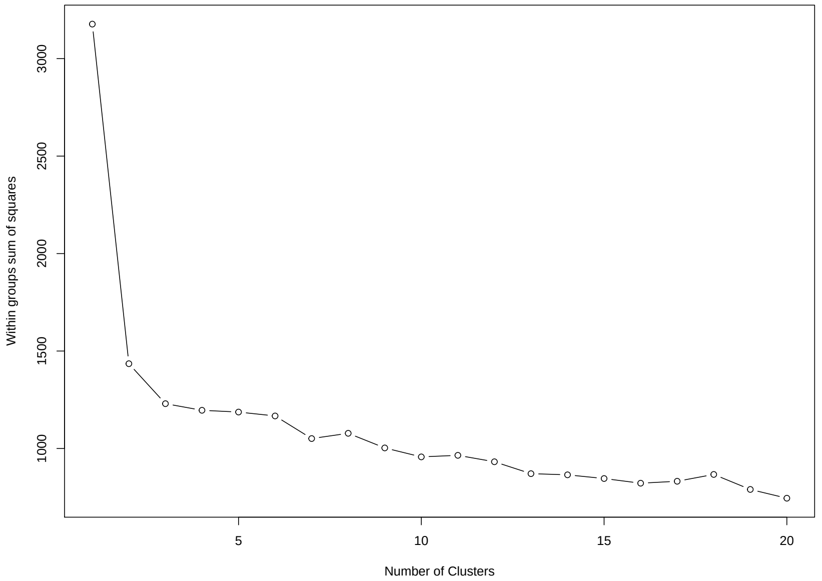


Figure C.5: Determining the number of clusters K for the dataset presented on Figure 3.5 for k-modes clustering. The number of clusters is chosen to be 2 in this case.

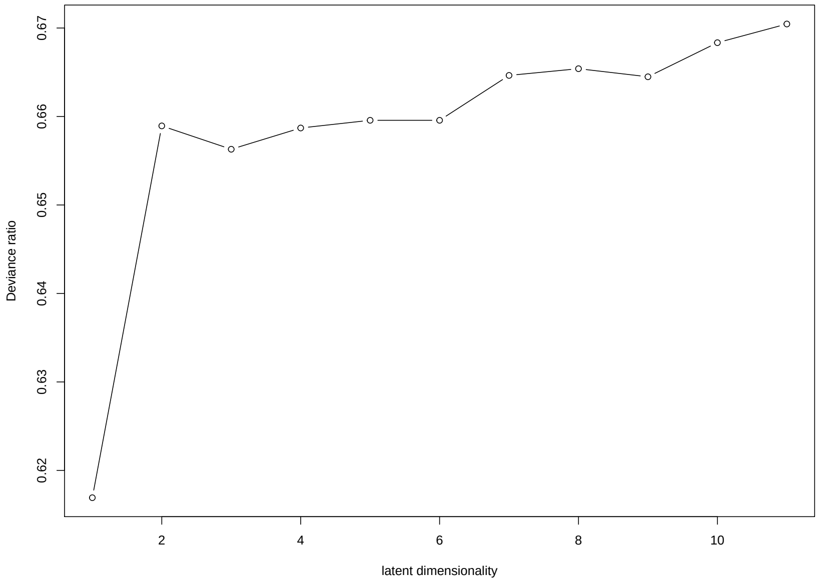


Figure C.6: Determining the latent dimensionality P in the case of iClusterPlus for the dataset presented on Figure 3.5 ($P = 2$)

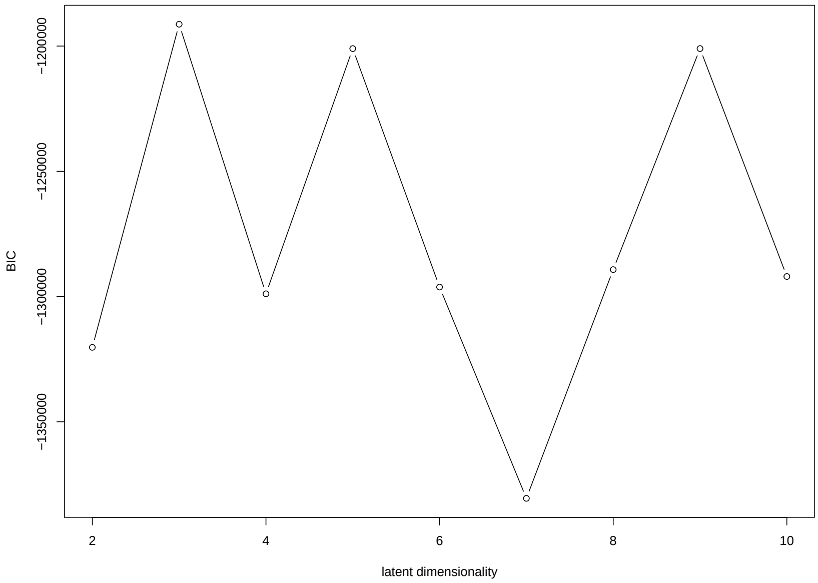


Figure C.7: Determining the latent dimensionality P in the case of BayesCluster for the dataset presented on Figure 3.5 ($P = 2$).

C.3 Mixed data

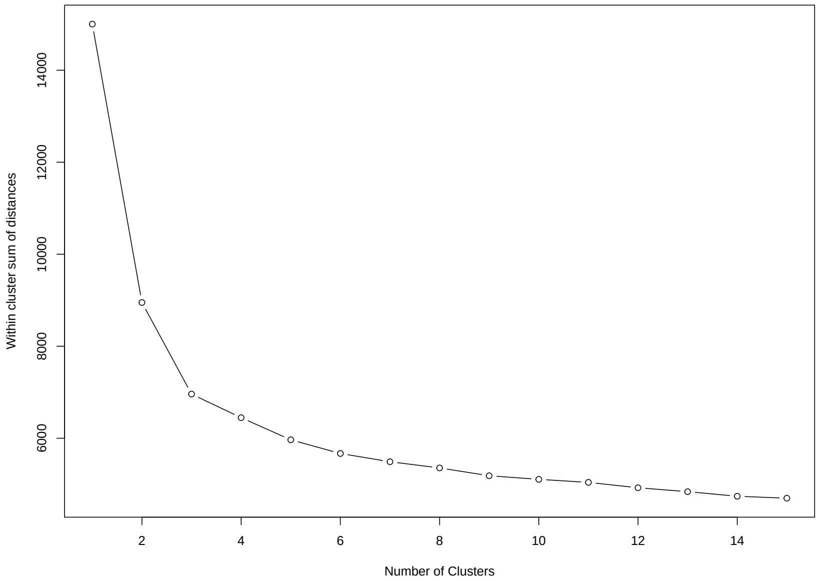


Figure C.8: Determining the number of clusters K in the case of k-prototypes for the dataset presented on Figure 3.8.

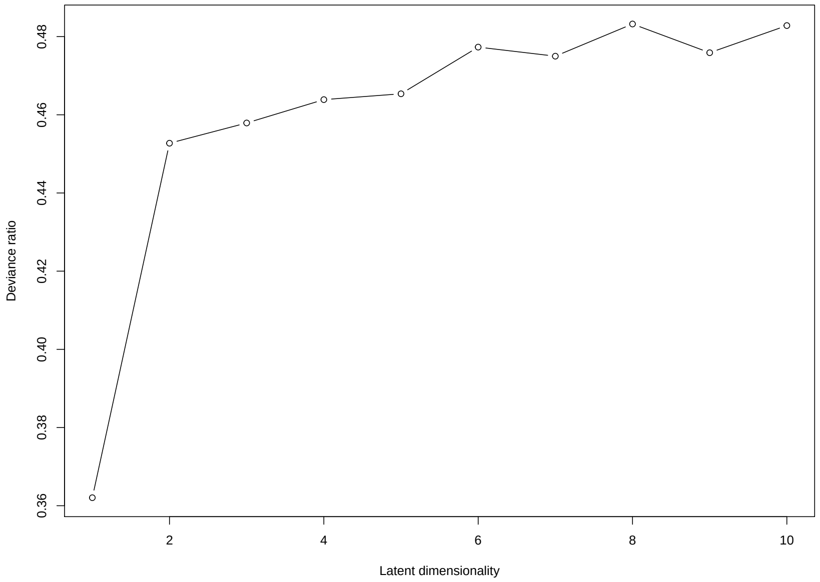


Figure C.9: Determining the latent dimensionality P in the case of iClusterPlus for the synthetic dataset presented on Figure 3.8 ($P = 2$)

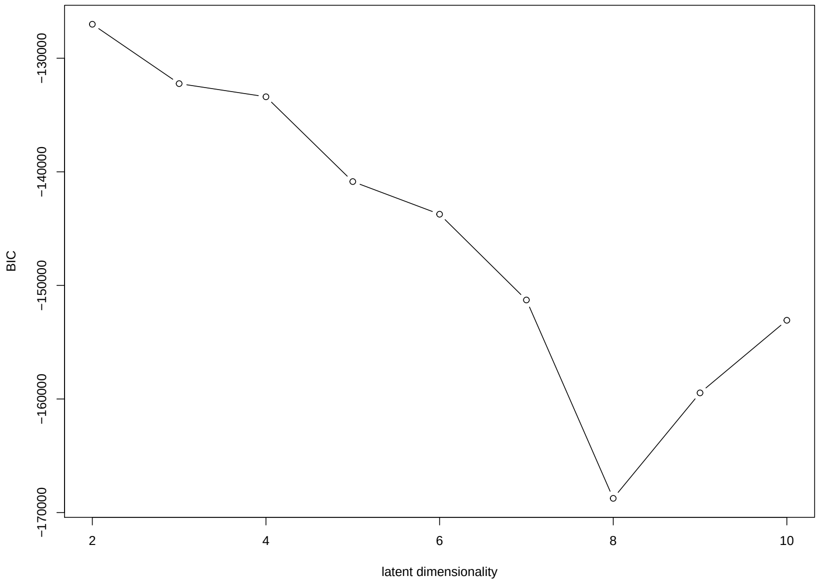


Figure C.10: Determining the latent dimensionality P in the case of BayesCluster for the synthetic dataset presented on Figure 3.8 ($P = 2$)

Appendix D

Summary of clinical data (Chapter 5)

D.1 Breast cancer

Feature	Categories	Number of patients
Age	< 50	73(33%)
	$[50, 64]$	94(44%)
	≥ 65	49(23%)
Vital status	alive	186(86%)
	dead	30(14%)
Gender	female	213(99%)
	male	3(1%)
Progesterone status	positive	141(65.3%)
	negative	74(34.3%)
	indeterminite	1(0.4%)
Estrogen status	positive	166(77%)
	negative	50(23%)
Her2 status	positive	22(10%)
	negative	109(50%)
	equivocal	43(20%)
	not available	42(20%)
Stage	T1	52(24%)
	T2	122(56%)
	T3	30(14%)

	T4	4(2%)

Table D.1: Summary of the characteristics of the clinical data for breast cancer patients.

D.2 Pancreatic cancer

Feature	Categories	Number of patients
Age	< 50	4(12%)
	$[50, 64]$	13(38%)
	≥ 65	17(50%)
Vital status	alive	18(53%)
	dead	16(47%)
Gender	female	15(44%)
	male	19(56%)
Stage	T1	2(6%)
	T2	30(88%)
	T3	1(3%)
	T4	1(3%)
Diabetes	yes	5(15%)
	no	16(47%)
	not available	13(38%)
Family history of cancer	yes	10(29%)
	no	7(21%)
	not available	17(50%)

Table D.2: Summary of the characteristics of the clinical data for pancreatic cancer patients.

D.3 Glioblastoma

Feature	Categories	Number of patients
Age	< 50	46(22%)
	$[50, 64]$	87(41%)
	≥ 65	77(36%)
Vital status	alive	64(30%)
	dead	145(69%)
Gender	female	82(39%)
	male	128(61%)
Karnofsky score	20	2(1%)
	40	8(4%)
	60	45(21%)
	80	86(41%)
	100	14(7%)
EGFR mutation	not available	126(60%)
	missence	20(10%)
	wild-type	64(30%)

Table D.3: Summary of the characteristics of the clinical data for glioblastoma patients.

D.4 Colorectal cancer

Feature	Categories	Number of patients
Age	< 50	13(6%)
	$[50, 64]$	52(24%)
	≥ 65	141(70%)
Vital status	alive	198(93%)
	dead	15(7%)
Gender	female	104(49%)
	male	109(51%)
Stage	T1	46(22%)
	T2	77(36%)
	T3	54(25%)
	T4	35(17%)

Table D.4: Summary of the characteristics of the clinical data for colorectal cancer patients.

Appendix E

Methods Chapter 5

E.0.1 Breast cancer

Using the statistical models (Section 5.2.2) and the integrative framework (Section 5.2.1), we can derive the mathematical form of the model which integrates the breast cancer gene expression (GE) and methylation (ME) datasets as follows:

$$\mathbf{X}_{brca,ge} = \mathbf{W}_{brca,ge}\mathbf{Z}_{brca} + \varepsilon_{brca,ge} \tag{E.1}$$

$$\mathbf{X}_{brca,me} = \mathbf{W}_{brca,me}\mathbf{Z}_{brca} + \varepsilon_{brca,me}, \tag{E.2}$$

where $\mathbf{W}_{brca,ge}$ and $\mathbf{W}_{brca,me}$ are the loading matrices which map the corresponding data onto a lower dimensional space, $\mathbf{Z}_{brca}$ are the latent variables corresponding to the underlying breast cancer subtypes, $\varepsilon_{brca,ge}$ and $\varepsilon_{brca,me}$, are the remaining variances unique to each data type after accounting for correlation between data types.

E.0.2 Pancreatic cancer

We can similarly derive the model which integrates the pancreatic cancer gene expression (GE) and methylation (ME) datasets:

$$\mathbf{X}_{pdac,ge} = \mathbf{W}_{pdac,ge}\mathbf{Z}_{pdac} + \varepsilon_{pdac,ge} \tag{E.3}$$

$$\mathbf{X}_{pdac,me} = \mathbf{W}_{pdac,me}\mathbf{Z}_{pdac} + \varepsilon_{pdac,me}, \tag{E.4}$$

where $\mathbf{W}_{pdac,ge}$ and $\mathbf{W}_{pdac,me}$ are the loading matrices which map the corresponding data onto a lower dimensional space, $\mathbf{Z}_{pdac}$ are the latent variables corresponding to the underlying pancreatic cancer subtypes, $\varepsilon_{pdac,ge}$ and $\varepsilon_{pdac,me}$, are the remaining variances unique to each data type after accounting for correlation between data types.

E.0.3 Glioblastoma

The integrative model which uses the information from the glioblastoma gene expression (GE), copy number variation (CNV), methylation (ME) and microRNA (miRNA) datasets can be wrtitten as follows:

$$\mathbf{X}_{gbm,ge} = \mathbf{W}_{gbm,ge}\mathbf{Z}_{gbm} + \varepsilon_{gbm,ge} \tag{E.5}$$

$$\mathbf{X}_{gbm,cnv} = \mathbf{W}_{gbm,cnv}\mathbf{Z}_{gbm} + \varepsilon_{gbm,cnv} \tag{E.6}$$

$$p(\mathbf{X}_{gbm,me}|\mathbf{W}_{gbm,1:R}, \mathbf{w}_{gbm,01:0R}) = \prod_{i=1}^{N}\prod_{r=1}^{R} \mathrm{Cat}(X_{me,ir}|\mathcal{S}(\mathbf{W}_{gbm,r}\mathbf{z}_{gbm,i} + \mathbf{w}_{gbm,0r})) \tag{E.7}$$

$$\mathbf{X}_{gbm,miRNA} = \mathbf{W}_{gbm,miRNA}\mathbf{Z}_{gbm} + \varepsilon_{gbm,miRNA} \tag{E.8}$$

where $\mathbf{W}_{gbm,ge}$, $\mathbf{W}_{gbm,cnv}$ and $\mathbf{W}_{gbm,miRNA}$ are the loading matrices which map the corresponding data onto a lower dimensional space, $\mathbf{W}_{gbm,1}, \ldots, \mathbf{W}_{gbm,R}$ are the loading matrices associated with the methylation dataset, $\mathbf{w}_{gbm,01}, \ldots \mathbf{w}_{gbm,0R}$ are the offsets, $\mathbf{Z}_{gbm}$ are the latent variables corresponding to the underlying glioblastoma subtypes, $\varepsilon_{gbm,ge}$, $\varepsilon_{gbm,cnv}$ and $\varepsilon_{gbm,miRNA}$, are the remaining variances unique to each data type after accounting for correlation between data types.

E.0.4 Colorectal cancer

The integrative model which uses the information from the colorectal cancer gene expression (GE), copy number variation (CNV) and methylation methylation (ME) datasets can be expressed as follows:

$$\mathbf{X}_{crc,ge} = \mathbf{W}_{crc,ge}\mathbf{Z}_{crc} + \varepsilon_{crc,ge} \tag{E.9}$$

$$\mathbf{X}_{crc,cnv} = \mathbf{W}_{crc,cnv}\mathbf{Z}_{crc} + \varepsilon_{crc,cnv} \tag{E.10}$$

$$\mathbf{X}_{crc,me} = \mathbf{W}_{crc,me}\mathbf{Z}_{crc} + \varepsilon_{crc,me}, \tag{E.11}$$

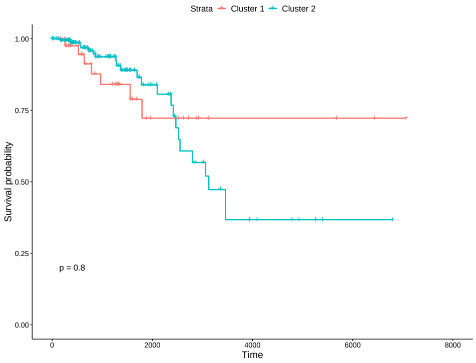
where $\mathbf{W}_{crc,ge}$, $\mathbf{W}_{crc,cnv}$ and $\mathbf{W}_{crc,me}$ are the loading matrices which map the corresponding data onto a lower dimensional space, $\mathbf{Z}_{crc}$ are the latent variables corresponding to the underlying colorectal cancer subtypes, $\varepsilon_{gbm,ge}$, $\varepsilon_{gbm,cnv}$ and $\varepsilon_{gbm,miRNA}$, are the remaining variances unique to each data type after accounting for correlation between data types.

Appendix F

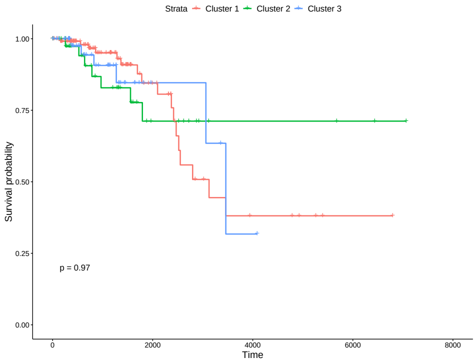
Kaplan-Meier curves Chapter 5

F.1 Breast cancer

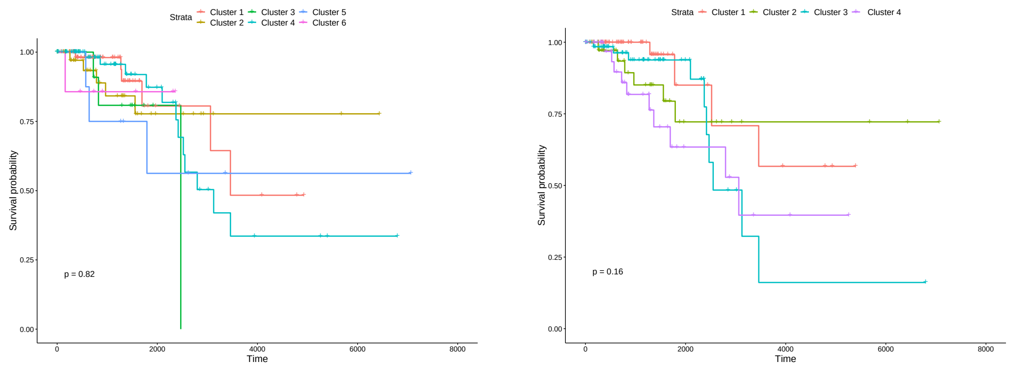
F.1.1 Gene expression



(a) k-means



(b) GMM

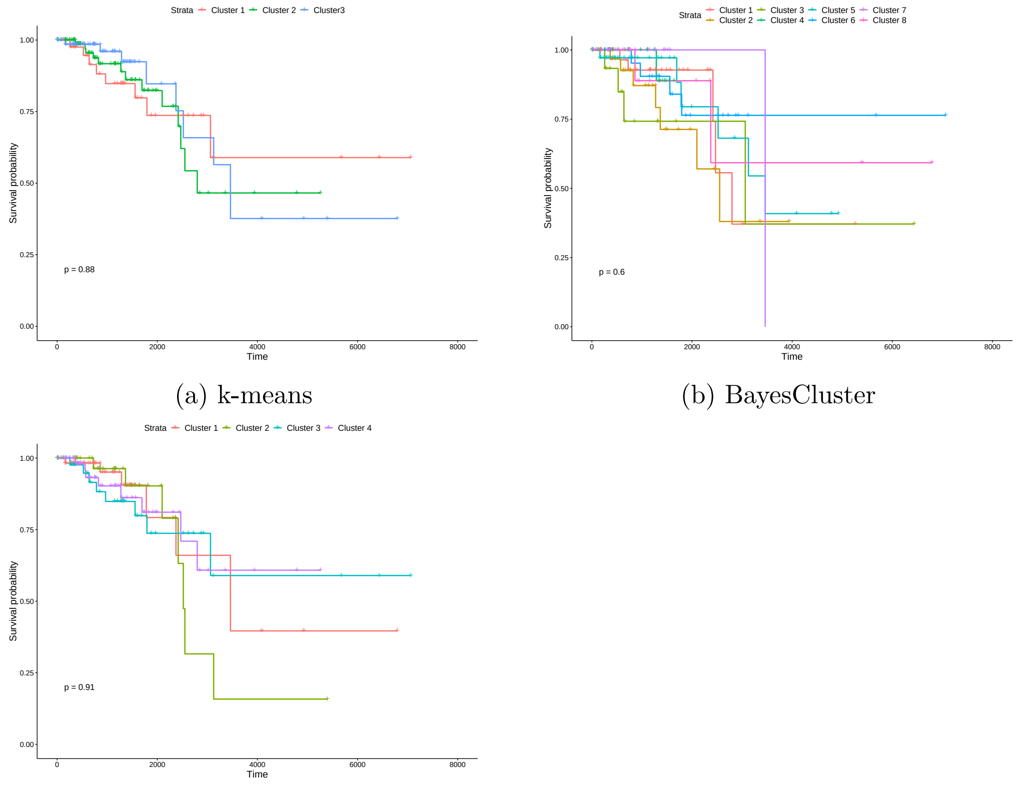


(c) BayesCluster

(d) iClusterPlus

Figure F.1: Breast cancer gene expression subtypes, identified by k-means, Gaussian mixture model, BayesCluster, iClusterPlus

F.1.2 Methylation



(a) k-means

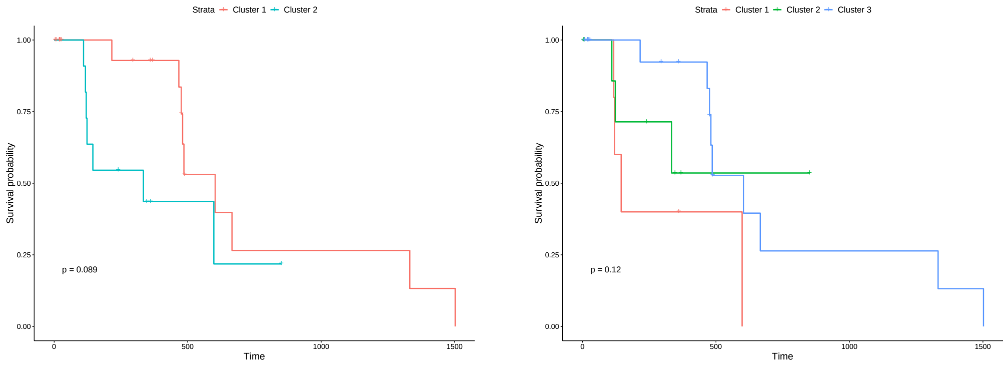
(b) BayesCluster

(c) iClusterPlus

Figure F.2: Breast cancer methylation subtypes, identified by k-means, BayesCluster, iClusterPlus

F.2 Pancreatic cancer

F.2.1 Gene expression

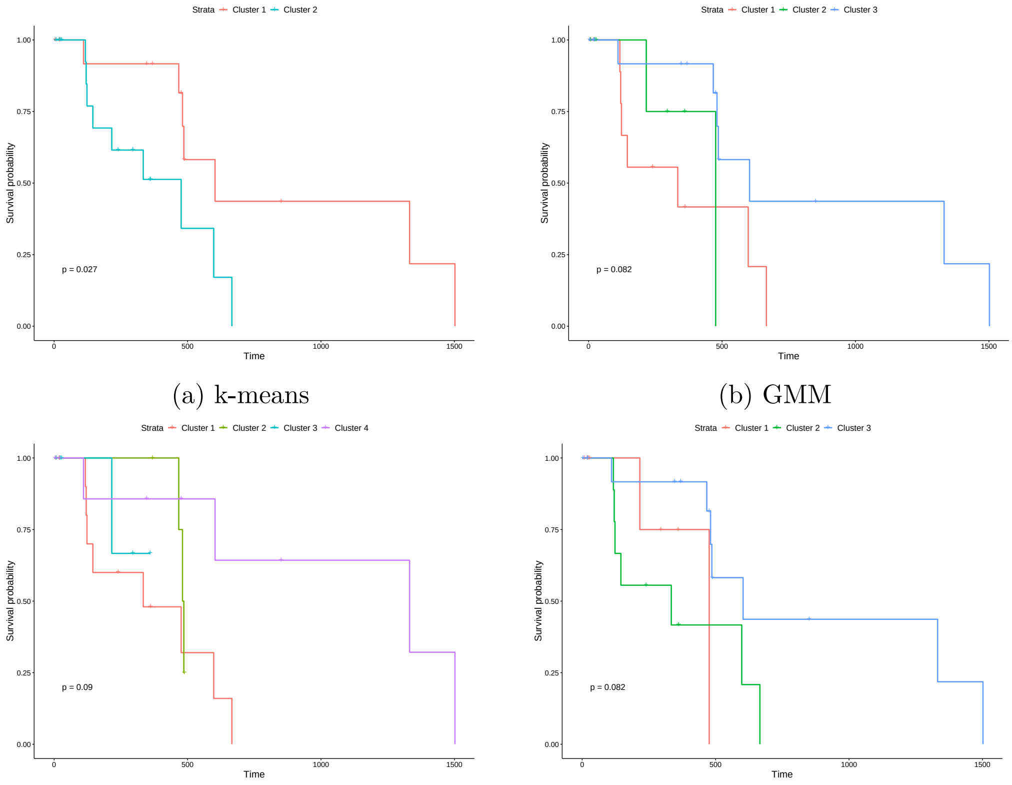


(a) k-means

(b) iClusterPlus

Figure F.3: Pancreatic cancer gene expression subtypes, identified by k-means and iClusterPlus

F.2.2 Methylation



(a) k-means

(b) GMM

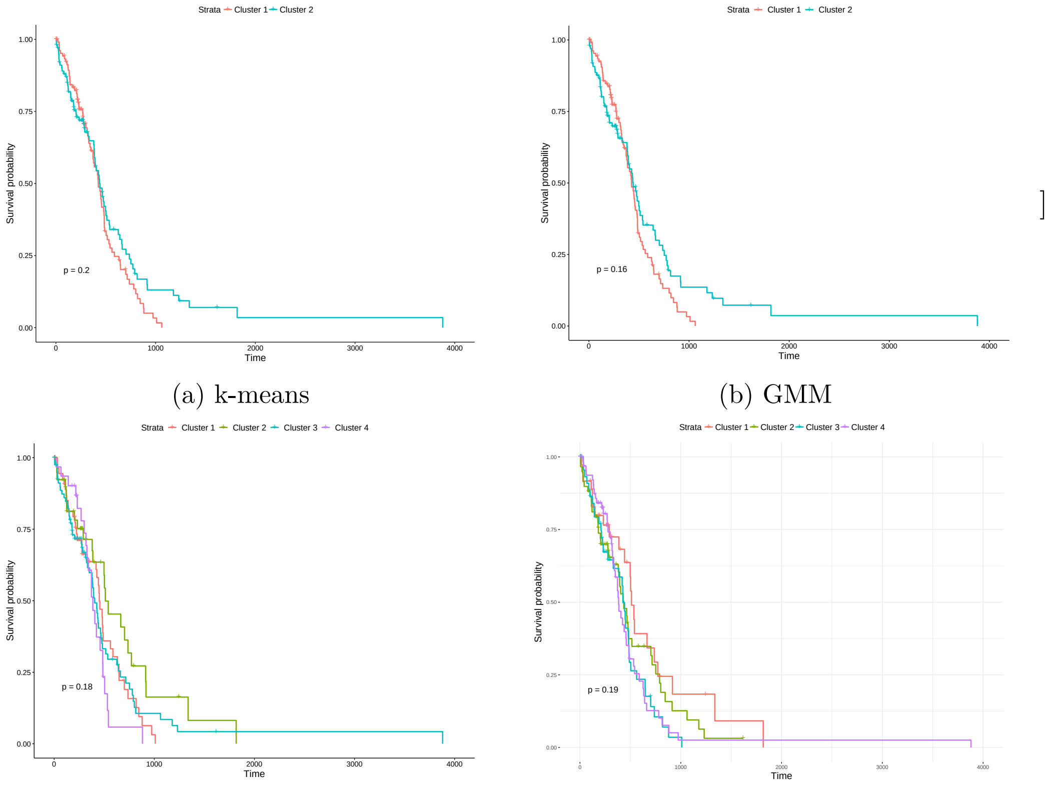
(c) BayesCluster

(d) iClusterPlus

Figure F.4: Pancreatic cancer methylation subtypes, identified by k-means, Gaussian mixture model, BayesCluster, iClusterPlus

F.3 Glioblastoma

F.3.1 Gene expression



(a) k-means

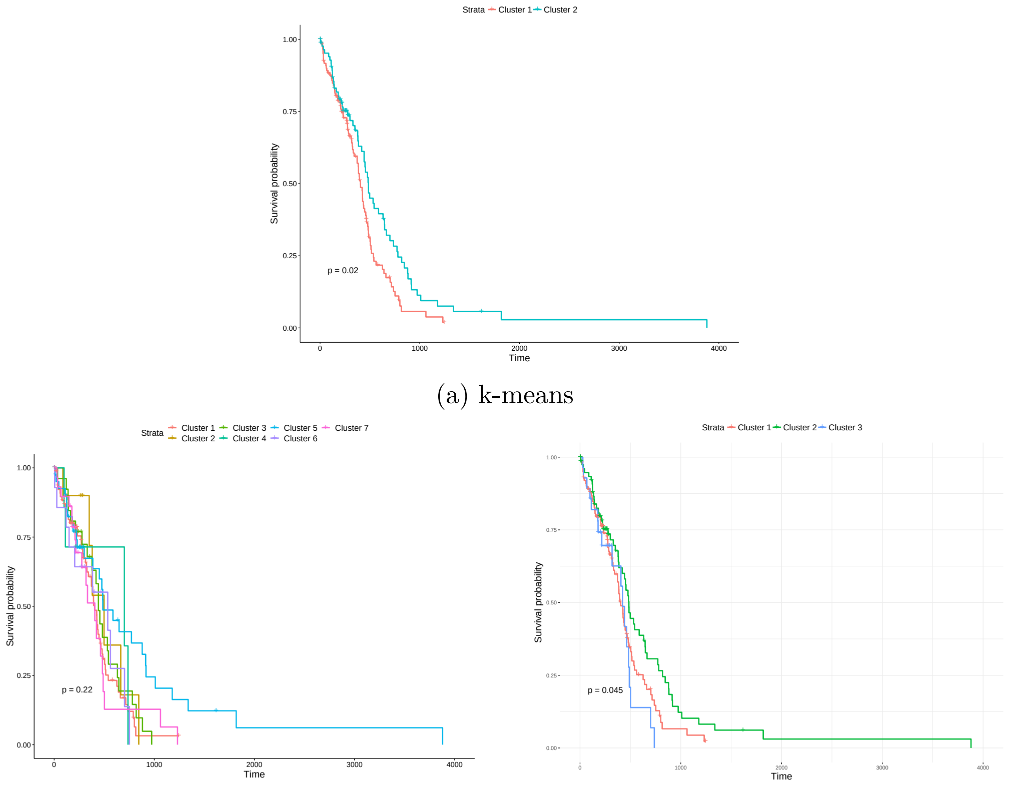
(b) GMM

(c) BayesCluster

(d) iClusterPlus

Figure F.5: Glioblastoma gene expression subtypes, identified by k-means, Gaussian mixture model, BayesCluster, iClusterPlus

F.3.2 Copy number variation



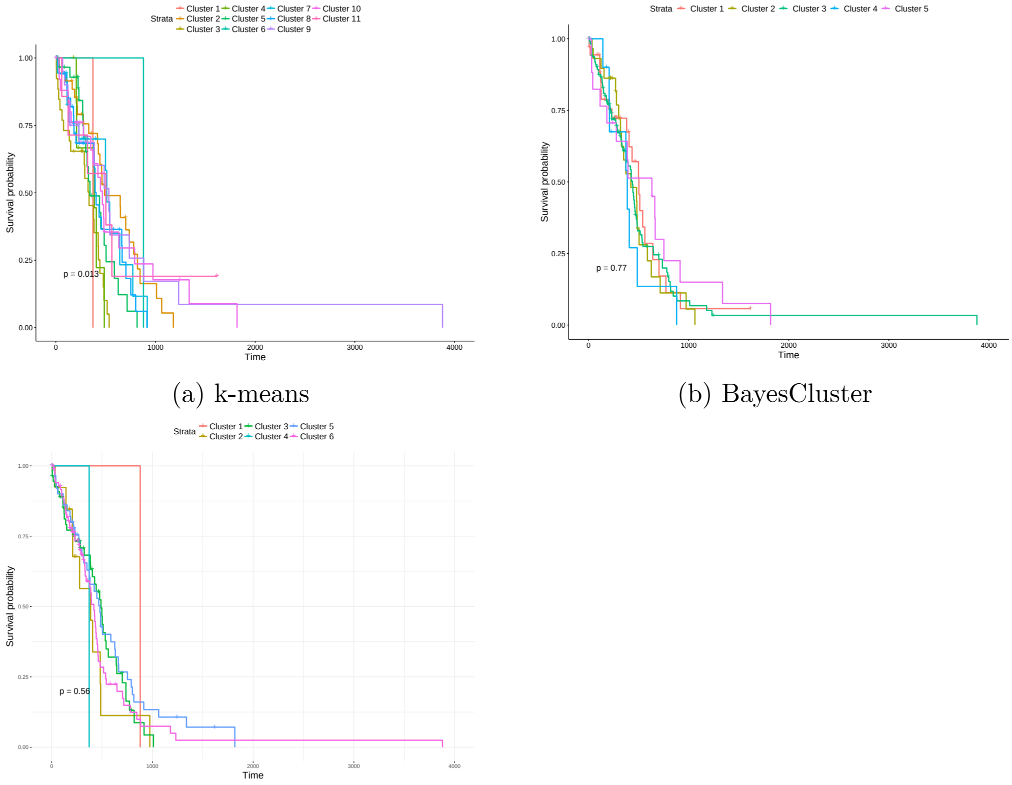
(a) k-means

(b) BayesCluster

(c) iClusterPlus

Figure F.6: Glioblastoma copy number variation subtypes, identified by k-means, Gaussian mixture model, BayesCluster, iClusterPlus

F.3.3 MicroRNA



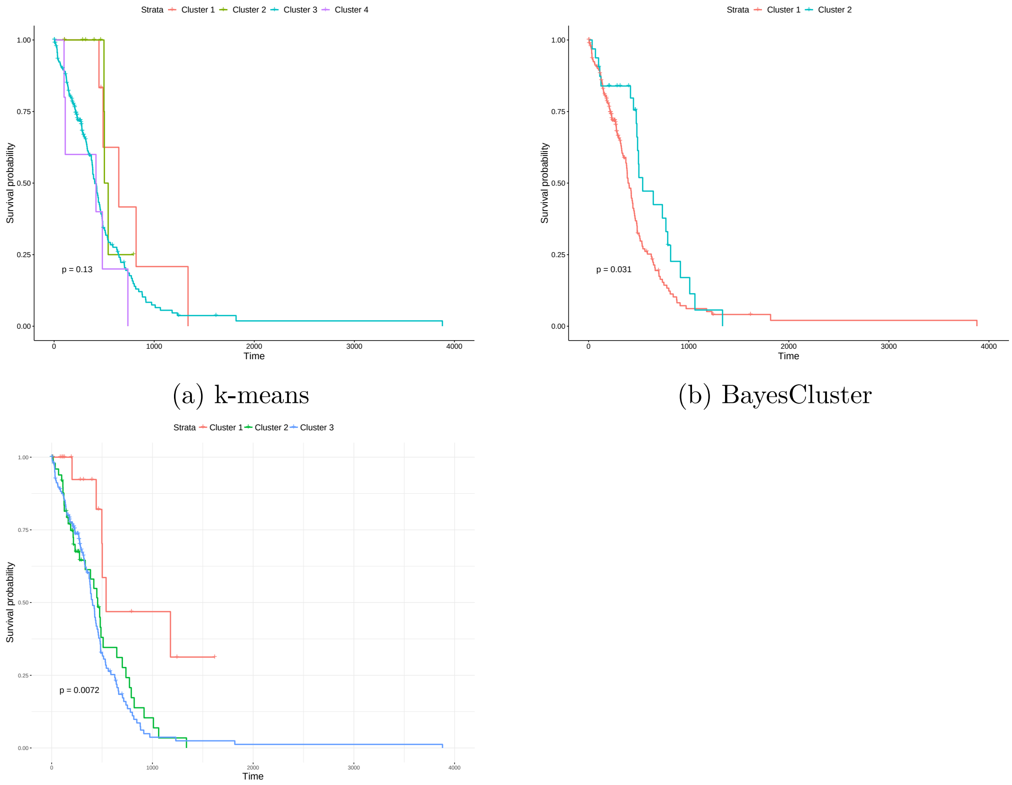
(a) k-means

(b) BayesCluster

(c) iClusterPlus

Figure F.7: Glioblastoma microRNA subtypes, identified by k-means, BayesCluster, iClusterPlus

F.3.4 Methylation



(a) k-means

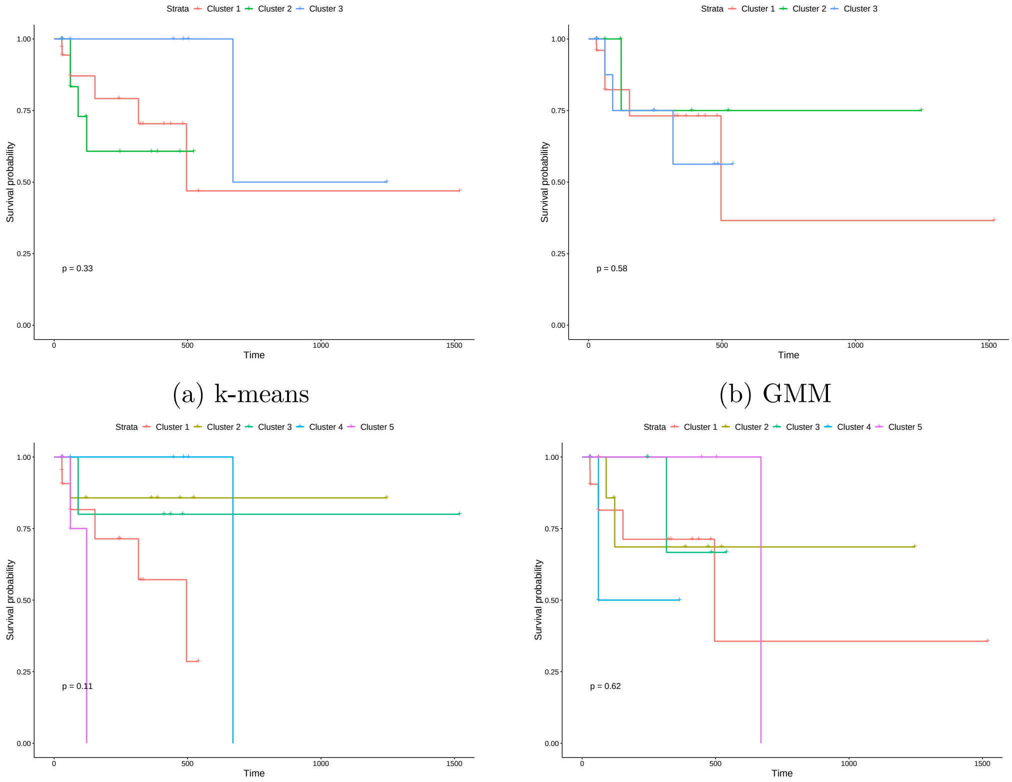
(b) BayesCluster

(c) iClusterPlus

Figure F.8: Glioblastoma metylation subtypes, identified by k-modes, BayesCluster, iClusterPlus

F.4 Colorectal cancer

F.4.1 Gene expression



(a) k-means

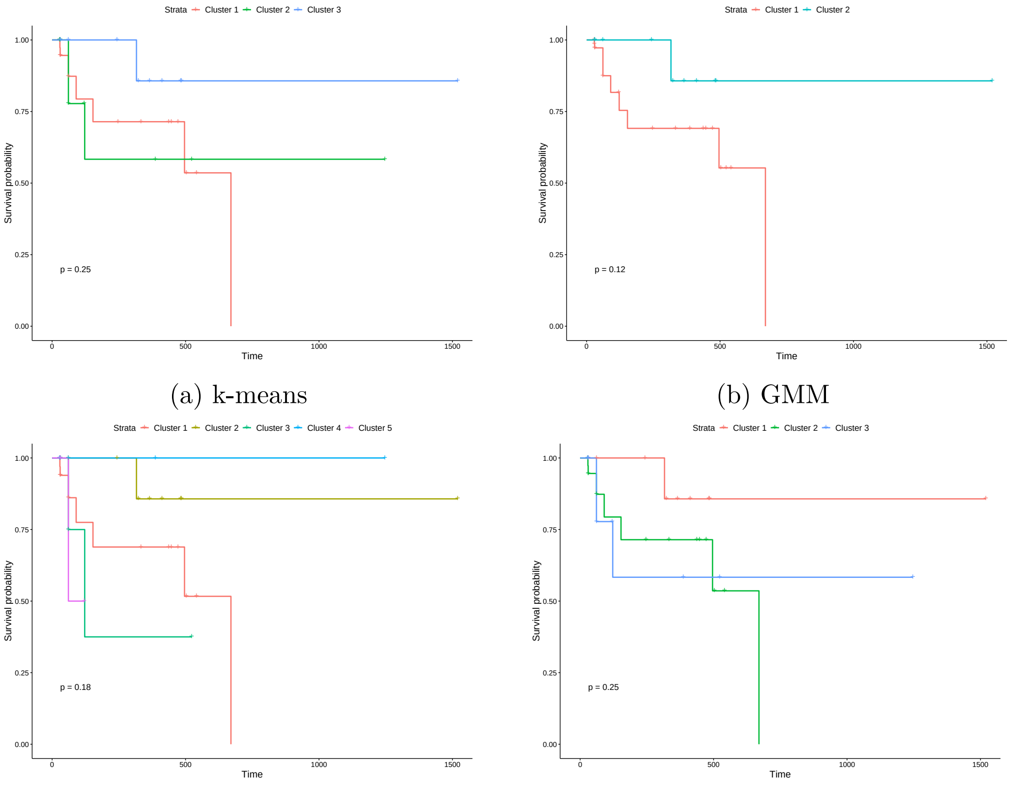
(b) GMM

(c) BayesCluster

(d) iClusterPlus

Figure F.9: Colorectal cancer gene expression subtypes, identified by k-means, Gaussian mixture model, BayesCluster, iClusterPlus

F.4.2 Copy number variation



(a) k-means

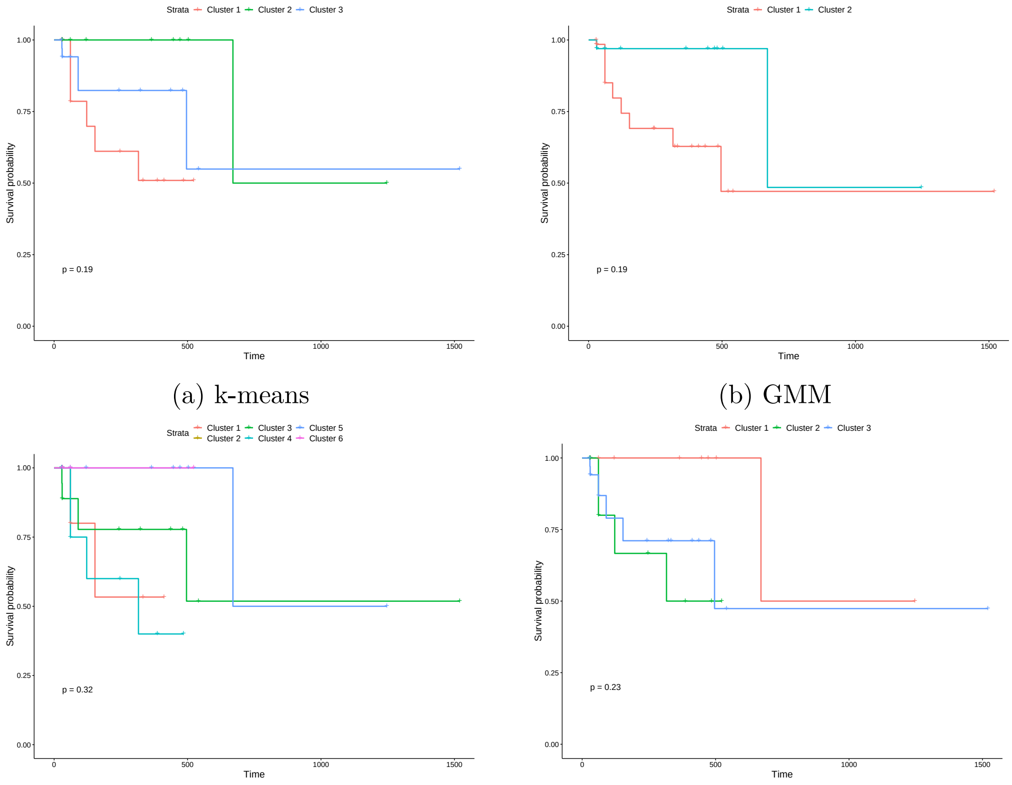
(b) GMM

(c) BayesCluster

(d) iClusterPlus

Figure F.10: Colorectal cancer copy number variation subtypes, identified by k-means, Gaussian mixture model, BayesCluster, iClusterPlus

F.4.3 Methylation



(a) k-means

(b) GMM

(c) BayesCluster

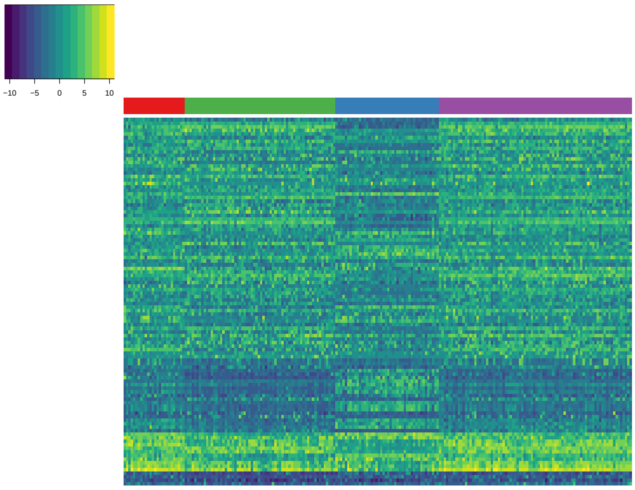
(d) iClusterPlus

Figure F.11: Colorectal cancer methylation subtypes, identified by k-means, Gaussian mixture model, BayesCluster, iClusterPlus

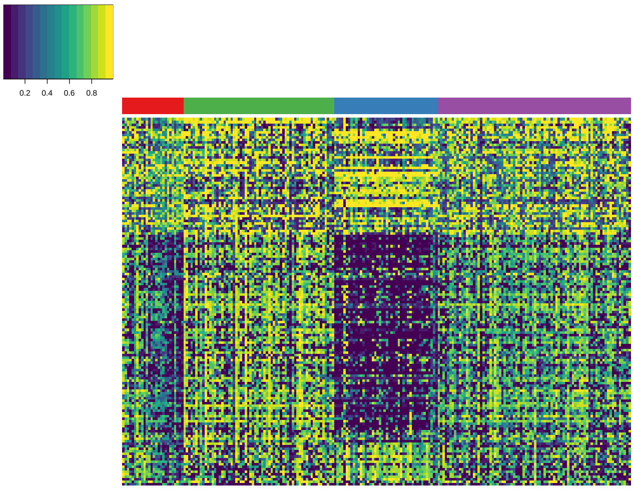
Appendix G

Heatmaps from Chapter 5

G.1 Breast cancer



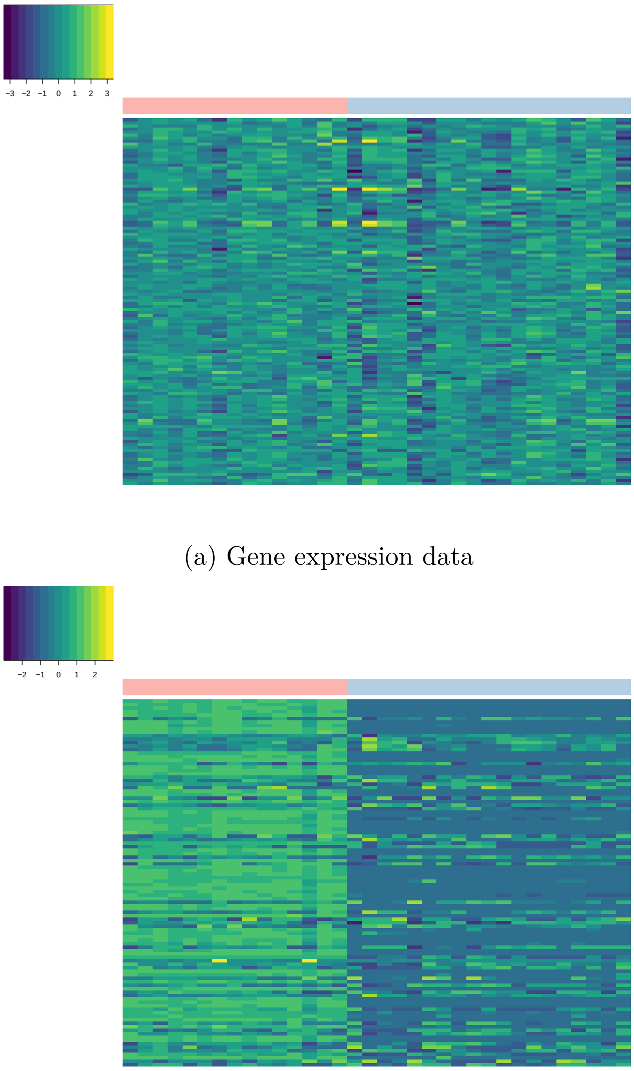
(a) Gene expression data



(b) Methylation data

Figure G.1: Heatmaps of the breast cancer gene expression and methylation data. The patients are on the x-axis, sorted by the integrative clustering partition. The y-axis gives the selected features for each data type, sorted using hierarchical clustering with average linkage.

G.2 Pancreatic cancer

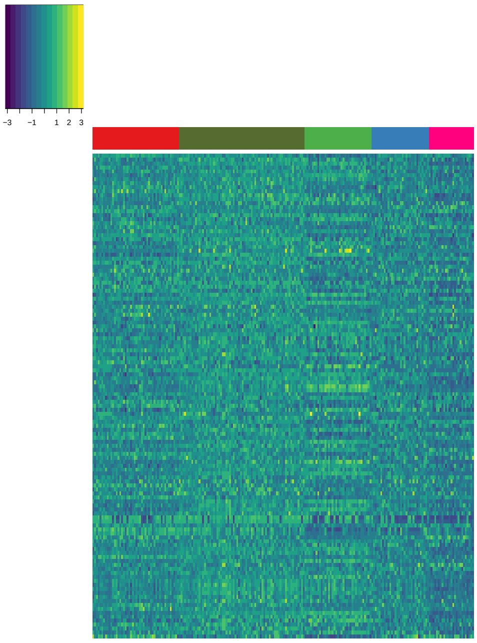


(a) Gene expression data

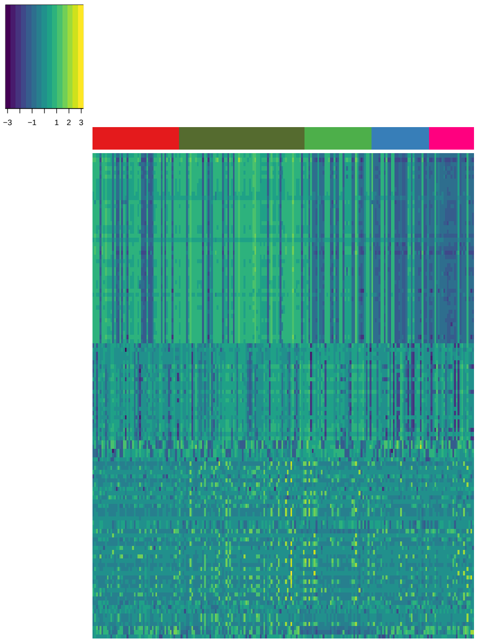
(b) Methylation data

Figure G.2: Heatmaps of the pancreatic cancer gene expression and methylation data. The patients are on the x-axis, sorted by the integrative clustering partition. The y-axis gives the selected features for each data type.

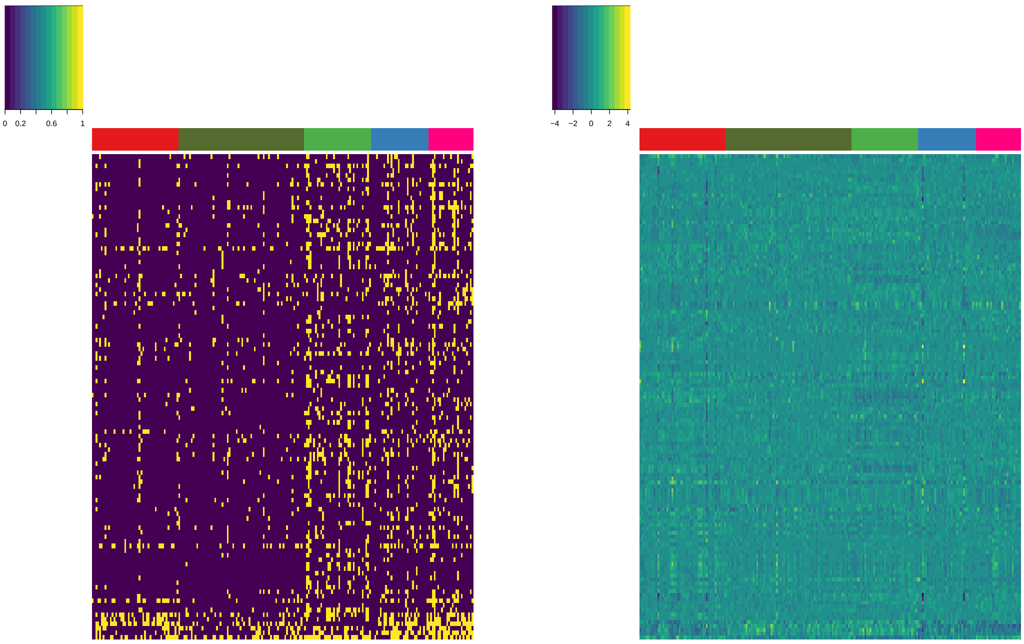
G.3 Glioblastoma



(a) Gene expression data



(b) Copy number variation data

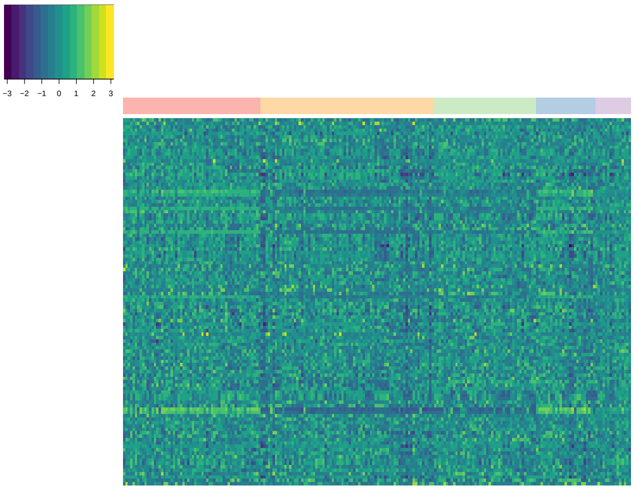


(c) Methylation data

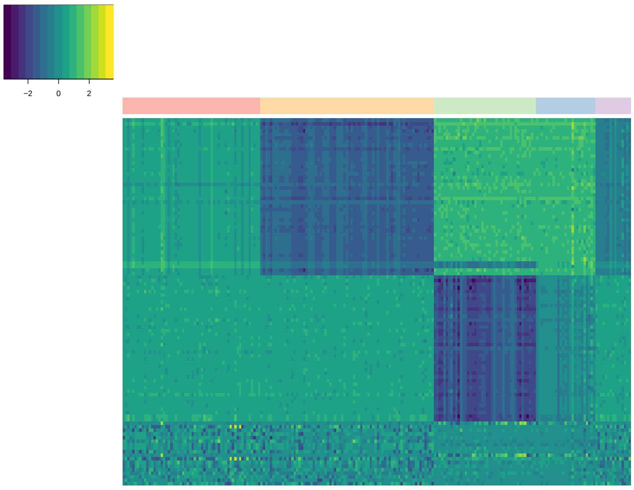
(d) MicroRNA data

Figure G.3: Heatmaps of the glioblastoma cancer gene expression, copy number variation, methylation and microRNA data. The patients are on the x-axis, sorted by the integrative clustering partition. The y-axis gives the selected features for each data type. The methylation data has been binarised.

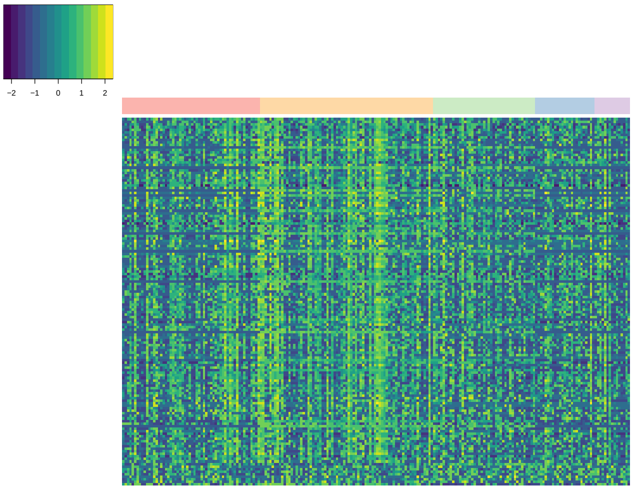
G.4 Colorectal cancer



(a) Gene expression data



(b) Copy number variation data



(c) Methylation data

Figure G.4: Heatmap of the colorectal cancer gene expression, copy number variation and methylation data. The patients are on the x-axis, sorted by the integrative clustering partition. The y-axis gives the selected features for each data type. The methylation data has been binarised.

Appendix H

Cluster specification (Chapter 5)

H.1 Breast cancer

Table H.1 presents the differences between the clusters identified by BayesCluster using the information from the gene expression and methylation data.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of patients	26	64	44	82
Median age at diagnosis	54.5	58	50	59
Age range	(38,71)	(26,83)	(29,83)	(29,83)
Progesterone receptor:				
negative	11	11	40	12
positive	15	53	3	70
indeterminite	0	0	1	0
Estrogen receptor:				
negative	6	4	38	2
positive	20	60	6	80
Her2 status:				
negative	12	24	28	45
positive	7	4	1	10
equivocal	3	19	6	15
indeterminite	0	1	1	0
not evaluated	3	12	8	8

not available	1	1	0	1
Triple negative subtype	0	0	23	1

Table H.1: Summary of the characteristics of the four integrative clusters identified by BayesCluster

Table H.2 presents the differences between the clusters identified by iClusterPlus using the information from the gene expression and methylation data.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of patients	42	43	67	44	20
Median age at diagnosis	59.5	55	58.5	49	57
Age range	(30,82)	(29,83)	(27,81)	(29,83)	(26,79)
Progesterone receptor:					
negative	16	40	5	8	5
positive	26	2	62	36	15
indeterminite		1			
Estrogen receptor:					
negative	7	37	3	2	1
positive	35	6	64	42	19
Her2 status:					
negative	16	29	39	17	8
positive	13	0	1	4	0
equivocal	6	5	10	17	5
indeterminite	0	1	0	1	0
not evaluated	6	8	8	5	4
not available	0	0	3	0	0
Triple negative subtype	1	25	0	0	0

Table H.2: Summary of the characteristics of the four integrative clusters identified by iClusterPlus.

H.2 Pancreatic cancer

Table H.3 presents the differences between the two clusters identified by BayesCluster using the information from the gene expression and methylation data.

Feature	Cluster 1	Cluster 2
Number of patients	15	19
Median age at diagnosis	65	64
Age range	(49,81)	(41,85)
Tumour stage:		
T2	0	3
T3	14	16
T4	1	0
Diabetes:		
no	5	11
yes	2	3
Family history of cancer:		
no	3	4
yes	3	7
Smoking (number of pack years):		
≤ 20	2	2
$(21, 30]$	1	1
> 30	1	1

Table H.3: Summary of the characteristics of the two integrative clusters identified by BayesCluster.

H.3 Glioblastoma

Table H.4 presents the differences between the five clusters identified by BayesCluster using the information from the gene expression, copy number variation, methylation and microRNA data. We have included information for all patients with follow-up information.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of patients	48	70	37	32	25
Median age at diagnosis	60.5	60	59	59.5	60
Age range	(30,88)	(10,89)	(23,85)	(21,83)	(36,83)
Gender					
female	18	24	15	13	12
male	30	45	22	17	13
Karnofsky score:					
20	0	0	0	1	1
40	2	2	3	1	0
60	11	20	5	2	7
80	15	28	7	13	13
100	4	7	1	2	0
EGFR mutation:					
wild-type	14	19	13	9	8
silent	2	0	0	1	0
missense	0	0	1	0	0
(silent)					
missense	7	6	2	0	4
not available	27	44	21	20	13

Table H.4: Summary of the characteristics of the five integrative clusters identified by BayesCluster.

H.4 Colorectal cancer

Table H.5 presents the differences between the five clusters identified by BayesCluster using the information from the gene expression, copy number variation and methylation data. We have included information for all patients (100) with follow-up information.

Feature	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Number of patients	22	35	23	15	5
Median age at diagnosis	69.5	74	63	68	77
Age range	(50,83)	(43,89)	(41,87)	(35,77)	(62,82)
Gender					
female	0	34	15	0	0
male	22	1	8	15	5
Tumour stage:					
T1	0	2	3	0	1
T2	7	3	5	7	0
T3	14	26	13	8	4
T4	1	4	2	0	0

Table H.5: Summary of the characteristics of the five integrative clusters identified by BayesCluster.

Bibliography

Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, volume 55. Courier Corporation, 1965.

Ken Aho, DeWayne Derryberry, and Teri Peterson. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95(3):631–636, 2014.

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

Khalid Al-Thihli, Teresa Rudkin, Nancy Carson, Chantal Poulin, Serge Melançon, and Vazken M Der Kaloustian. Compound heterozygous deletions of PMP22 causing severe Charcot-Marie-Tooth disease of the Dejerine-Sottas disease phenotype. *American Journal of Medical Genetics Part A*, 146(18):2412–2416, 2008.

Orly Alter, Patrick O Brown, and David Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

Alessia Amelio and Clara Pizzuti. Is normalized mutual information a fair measure for comparing community detection methods? In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1584–1585. ACM, 2015.

American Society of Anesthesiologists. ASA Physical Status Qualification System. URL `https://www.asahq.org/resources/clinical-information/asa-physical-status-classification-system`. [Online; accessed 10-May-2018].

L. N. Fred Ana and Anil K Jain. Robust data clustering. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2003.

Charles E Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974.

Chandrakanth Are, Chantal Afuh, Lavanya Ravipati, Aaron Sasson, Fred Ullrich, and Lynette Smith. Preoperative nomogram to predict risk of perioperative mortality following pancreatic resections for malignancy. *Journal of Gastrointestinal Surgery*, 13(12):2152, 2009.

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C. Marioni, Wolfgang Huber, Florian Buettner, and Oliver Stegle. Multi-omics factor analysis - a framework for unsupervised integration of multi-omic data sets. *bioRxiv*, 2018.

Richard Arratia, Andrew D Barbour, and Simon Tavaré. *Logarithmic Combinatorial Structures: a Probabilistic Approach*, volume 1. European Mathematical Society, 2003.

F Aubin, S Gill, R Burkes, B Colwell, S Kamel-Reid, S Koski, A Pollett, B Samson, M Tehfe, R Wong, et al. Canadian expert group consensus recommendations: KRAS testing in colorectal cancer. *Current Oncology*, 18(4):e180, 2011.

Vinod P Balachandran, Mithat Gonen, J Joshua Smith, and Ronald P DeMatteo. Nomograms in oncology: more than meets the eye. *The Lancet Oncology*, 16(4): e173–e180, 2015.

Arindam Banerjee and Joydeep Ghosh. Scalable clustering algorithms with balancing constraints. *Data Mining and Knowledge Discovery*, 13(3):365–395, 2006.

Daniel Barbará, Yi Li, and Julia Couto. COOLCAT: an entropy-based algorithm for categorical clustering. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pages 582–589. ACM, 2002.

Anand M Baswade and Prakash S Nalwade. Selection of initial centroids for k-means algorithm. *IJCSMC*, 2(7):161–164, 2013.

José M Bernardo and Adrian FM Smith. Bayesian Theory, 2001.

Christopher M Bishop. Latent variable models. In *Learning in Graphical Models*, pages 371–403. Springer, 1998.

Christopher M Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems*, pages 382–388, 1999.

Christopher M Bishop. *Pattern recognition and Machine Learning.* Springer, 2006.

David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

David R Blair, Christopher S Lyttle, Jonathan M Mortensen, Charles F Bearden, Anders Boeck Jensen, Hossein Khiabanian, Rachel Melamed, Raul Rabadan, Elmer V Bernstam, Søren Brunak, et al. A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk. *Cell*, 155(1):70–80, 2013.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

David M Blei, Michael I Jordan, et al. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

Paul S Bradley and Usama M Fayyad. Refining initial points for k-means clustering. In *International Conference on Machine Learning*, volume 98, pages 91–99. Citeseer, 1998.

PS Bradley, KP Bennett, and Ayhan Demiriz. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0, 2000.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Robert Brown and Gordon Strathdee. Epigenomics and epigenetic therapy of cancer. *Trends in Molecular Medicine*, 8(4):S43–S48, 2002.

Markus W Büchler, Jörg Kleeff, and Helmut Friess. Surgical treatment of pancreatic cancer. *Journal of the American College of Surgeons*, 205(4):S81–S86, 2007.

Kerstin Bunte, Eemeli Leppäaho, Inka Saarinen, and Samuel Kaski. Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics*, 32(16): 2457–2463, 2016.

Wray Buntine. Variational extensions to EM and multinomial PCA. In *European Conference on Machine Learning*, pages 23–34. Springer, 2002.

Wray Buntine and Aleks Jakulin. Applying discrete PCA in data analysis. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 59–66. AUAI Press, 2004.

Christopher A Bush and Steven N MacEachern. A semiparametric Bayesian model for randomised block designs. *Biometrika*, 83(2):275–285, 1996.

Diana Cai, Trevor Campbell, and Tamara Broderick. Finite mixture models are typically inconsistent for the number of components. *31st Conference on Neural Information Processing Systems*, 2017.

Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

Cancer Genome Atlas Network and others. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012a.

Cancer Genome Atlas Network and others. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012b.

Cancer Genome Atlas Research Network and others. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, 513(7517):202, 2014.

Cancer Research UK. Pancreatic cancer statistics. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/pancreatic-cancer`, 2018. [Online; accessed 18-July-2018].

Gilles Celeux, Merrilee Hurn, and Christian P Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

Thomas E Clancy. Surgery for pancreatic cancer. *Hematology/Oncology Clinics*, 29 (4):701–716, 2015.

Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, pages 617–624, 2002.

Eric A Collisson, Anguraj Sadanandam, Peter Olson, William J Gibb, Morgan Truitt, Shenda Gu, Janine Cooc, Jennifer Weinkle, Grace E Kim, Lakshmi Jakkula, et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature medicine*, 17(4):500, 2011.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

David R Cox. Regression models and life-tables. In *Breakthroughs in Statistics*, pages 527–541. Springer, 1992.

Oliver M Crook, Claire M Mulvey, Paul DW Kirk, Kathryn S Lilley, and Laurent Gatto. A Bayesian mixture modelling approach for spatial proteomics. *bioRxiv*, page 282269, 2018.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346, 2012.

David B Dahl. An improved merge-split sampler for conjugate Dirichlet process mixture models. *Technical Report*, 1:086, 2003.

Bruno De Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

Roeland F de Wilde, MGH Besselink, Ingeborg van der Tweel, IHJT De Hingh, CHJ Van Eijck, CHC Dejong, Robert J Porte, Dirk J Gouma, ORC Busch, and I Quintus Molenaar. Impact of nationwide centralization of pancreaticoduodenectomy on hospital mortality. *British Journal of Surgery*, 99(3):404–410, 2012.

M Dolores Delgado and Javier León. Gene expression regulation and cancer. *Clinical and Translational Oncology*, 8(11):780–787, 2006.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

Department for Communities and Local Government. The english indices of deprivation 2015 - frequently asked questions. `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/579151/English_Indices_of_Deprivation_2015_-_Frequently_Asked_Questions_Dec_2016.pdf`, 2018. [Online; accessed 05-July-2018].

Department of Health. Improving outcomes: A strategy for cancer. 2011.

Persi Diaconis. Finite forms of de Finetti's theorem on exchangeability. *Synthese*, 36(2):271–281, 1977.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1995.

David B Dunson. Bayesian nonparametric hierarchical modeling. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(2):273–284, 2009.

Jennifer G Dy and Carla E Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5(Aug):845–889, 2004.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

Harold Ainsley Evesham. *The History and Development of Nomography.* Docent Press, 2010.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27 (8):861–874, 2006.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

Michael Fop, Thomas Brendan Murphy, et al. Variable selection methods for model-based clustering. *Statistics Surveys*, 12:18–65, 2018.

Jean-Paul Fox. *Bayesian Item Response Modeling: Theory and Applications.* Springer Science & Business Media, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jairo Fúquene, Mark Steel, and David Rossell. On choosing mixture components via non-local priors. *arXiv preprint arXiv:1604.00314*, 2016.

Jairo Fuquene, Mark Steel, and David Rossell. On choosing mixture components via non-local priors. *arXiv:1604.00314*, 2016.

Evelina Gabasova, John Reid, and Lorenz Wernisch. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLoS Computational Biology*, 13(10):e1005781, 2017.

Jeffrey Gagan and Eliezer M Van Allen. Next-generation sequencing to guide cancer therapy. *Genome Medicine*, 7(1):80, 2015.

Eric R Gamazon and Barbara E Stranger. The impact of human copy number variation on gene expression. *Briefings in Functional Genomics*, 14(5):352–357, 2015.

Venkatesh Ganti, Johannes Gehrke, and Raghu Ramakrishnan. CACTUS—clustering categorical data using summaries. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 73–83. ACM, 1999.

Simon Garnier. viridis: Default color maps from 'matplotlib'. 2018. URL `https://CRAN.R-project.org/package=viridis`. R package version 0.5.1.

Ramiro Garzon, Muller Fabbri, Amelia Cimmino, George A Calin, and Carlo M Croce. MicroRNA expression and function in cancer. *Trends in Molecular Medicine*, 12(12):580–587, 2006a.

Ramiro Garzon, Flavia Pichiorri, Tiziana Palumbo, Rodolfo Iuliano, Amelia Cimmino, Rami Aqeilan, Stefano Volinia, Darshna Bhatt, Hansjuerg Alder, Guido Marcucci, et al. MicroRNA fingerprints during human megakaryocytopoiesis. *Proceedings of the National Academy of Sciences*, 103(13):5078–5083, 2006b.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014a.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014b.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in Computer Vision*, pages 564–584. Elsevier, 1987.

JK Ghosh and RV Ramamoorthi. *Bayesian Nonparametrics.* Springer Science & Business Media, 2003.

David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Clustering categorical data: an approach based on dynamical systems. *Databases*, 1:75, 1998.

Geoffrey S Ginsburg and Kathryn A Phillips. Precision medicine: From science to value. *Health Affairs*, 37(5):694–701, 2018.

Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A

Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

GA Gooiker, VEPP Lemmens, MG Besselink, OR Busch, BA Bonsing, I Quintus Molenaar, RAEM Tollenaar, IHJT de Hingh, and MWJM Wouters. Impact of centralization of pancreatic cancer surgery on resection rates and survival. *British Journal of Surgery*, 101(8):1000–1005, 2014.

Gea A Gooiker, Lydia GM van der Geest, Michel WJM Wouters, Marieke Vonk, Tom M Karsten, Rob AEM Tollenaar, and Bert A Bonsing. Quality improvement of pancreatic surgery by centralization in the western part of the netherlands. *Annals of Surgical Oncology*, 18(7):1821–1829, 2011.

John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

Peter J Green and Sylvia Richardson. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375, 2001.

Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366, 2000.

Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien De Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*, 21(11):1350, 2015.

Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *Cell*, 100(1): 57–70, 2000.

Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

Reina Haque, Syed A Ahmed, Galina Inzhakova, Jiaxiao Shi, Chantal Avila, Jonathan Polikoff, Leslie Bernstein, Shelley M Enger, and Michael F Press. Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades. *Cancer Epidemiology and Prevention Biomarkers*, 2012.

Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996.

John A Hartigan and Manchek A Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

Manal M Hassan, Melissa L Bondy, Robert A Wolff, James L Abbruzzese, Jean-Nicolas Vauthey, Peter W Pisters, Douglas B Evans, Rabia Khan, Ta-Hsu Chou, Renato Lenzi, et al. Risk factors for pancreatic cancer: case-control study. *The American Journal of Gastroenterology*, 102(12):2696–2707, 2007.

W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Squeezer: an efficient algorithm for clustering categorical data. *Journal of Computer Science and Technology*, 17(5): 611–624, 2002.

Zengyou He, Xiaofei Xu, and Shengchun Deng. Clustering mixed numeric and categorical data: A cluster ensemble approach. *arXiv preprint cs/0509011*, 2005.

Charlotte N Henrichsen, Nicolas Vinckenbosch, Sebastian Zöllner, Evelyne Chaignat, Sylvain Pradervand, Frédéric Schütz, Manuel Ruedi, Henrik Kaessmann, and Alexandre Reymond. Segmental copy number variation shapes tissue transcriptomes. *Nature Genetics*, 41(4):424, 2009.

Joshua S Hill, Zheng Zhou, Jessica P Simons, Sing Chau Ng, Theodore P McDade, Giles F Whalen, and Jennifer F Tseng. A simple risk score to predict in-hospital mortality after pancreatic resection for cancer. *Annals of Surgical Oncology*, 17 (7):1802–1807, 2010.

Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G Walker. *Bayesian Nonparametrics*, volume 28. Cambridge University Press, 2010.

Vivian Ho and Martin J Heslin. Effect of hospital volume and experience on in-hospital mortality for pancreaticoduodenectomy. *Annals of surgery*, 237(4):509, 2003.

Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*, volume 751. John Wiley & Sons, 2013.

Neal S Holter, Madhusmita Mitra, Amos Maritan, Marek Cieplak, Jayanth R Banavar, and Nina V Fedoroff. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proceedings of the National Academy of Sciences*, 97(15):8409–8414, 2000.

David W Hosmer and Stanley Lemesbow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9 (10):1043–1069, 1980.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

Joshua Zhexue Huang, Michael K Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668, 2005.

Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. pages 21–34, 1997.

Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and Knowledge Discovery*, 2(3):283–304, 1998.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Michael C Hughes, Emily Fox, and Erik B Sudderth. Effective split-merge Monte Carlo methods for nonparametric models of sequential data. In *Advances in Neural Information Processing Systems*, pages 1295–1303, 2012.

IDS. Cancer statistics, all types of cancer, 2018. URL `http://www.isdscotland.org/Health-Topics/Cancer/Cancer-Statistics/All-Types-of-Cancer`.

International Cancer Genome Consortium and others. International network of cancer genome projects. *Nature*, 464(7291):993, 2010.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

Hemant Ishwaran and Lancelot F James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, pages 1211–1235, 2003.

Hemant Ishwaran and Glen Takahara. Independent and identically distributed Monte Carlo algorithms for semiparametric linear mixed models. *Journal of the American Statistical Association*, 97(460):1154–1166, 2002.

Hemant Ishwaran and Mahmoud Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

Eric J Jacobs, Stephen J Chanock, Charles S Fuchs, Andrea LaCroix, Robert R McWilliams, Emily Steplowski, Rachael Z Stolzenberg-Solomon, Alan A Arslan, H Bas Bueno-de Mesquita, Myron Gross, et al. Family history of cancer and risk of pancreatic cancer: a pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan). *International Journal of Cancer*, 127(6):1421–1428, 2010.

Sonia Jain and Radford M Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182, 2004.

Sonia Jain, Radford M Neal, et al. Splitting and merging components of a nonconjugate Dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472, 2007.

Shane T Jensen and Jun S Liu. Bayesian clustering of transcription factor binding motifs. *Journal of the American Statistical Association*, 103(481):188–200, 2008.

Aditya Jitta and Arto Klami. On controlling the size of clusters in probabilistic clustering. 2018.

Valen E Johnson and James H Albert. *Ordinal Data Modeling*. Springer Science & Business Media, 2006.

Valen E Johnson and David Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170, 2010.

Valen E Johnson and David Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal Component Analysis*, pages 115–128. Springer, 1986.

Alboukadel Kassambara and Marcin Kosinski. survminer: Drawing survival curves using 'ggplot2'. 2018. URL `https://CRAN.R-project.org/package=survminer`. R package version 0.4.3.

Mohammad E Khan, Guillaume Bouchard, Kevin P Murphy, and Benjamin M Marlin. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems*, pages 1108–1116, 2010.

Suleiman A Khan, Seppo Virtanen, Olli P Kallioniemi, Krister Wennerberg, Antti Poso, and Samuel Kaski. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. *Bioinformatics*, 30(17):i497–i504, 2014.

Paul Kirk, Jim E Griffin, Richard S Savage, Zoubin Ghahramani, and David L Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28 (24):3290–3297, 2012.

Scott Kirkpatrick, Mario P Vecchi, et al. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

Arto Klami and Aditya Jitta. Probabilistic size-constrained microclustering. In *UAI*, 2016.

Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9): 2136–2147, 2015.

Bruce M Koeppen and Bruce A Stanton. *Renal Physiology E-Book: Mosby Physiology Monograph Series (with Student Consult Online Access)*. Elsevier Health Sciences, 2012.

Jouni Kuha. AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research*, 33(2):188–229, 2004.

Marta Kulis and Manel Esteller. DNA methylation and cancer. In *Advances in Genetics*, volume 70, pages 27–56. Elsevier, 2010.

Ludmila I Kuncheva and Stefan Todorov Hadjitodorov. Using diversity in cluster ensembles. In *Systems, Man and Cybernetics, 2004 IEEE international conference on*, volume 2, pages 1214–1219. IEEE, 2004.

Kenichi Kurihara, Max Welling, and Yee Whye Teh. Collapsed variational Dirichlet process mixture models. In *IJCAI*, volume 7, pages 2796–2801, 2007.

Andrew J Landgraf and Yoonkyung Lee. Dimensionality reduction for binary data through the projection of natural parameters. *arXiv preprint arXiv:1510.06112*, 2015.

Damien J LaPar, Irving L Kron, David R Jones, George J Stukenborg, and Benjamin D Kozower. Hospital procedure volume should not be used as a measure of surgical quality. *Annals of Surgery*, 256(4):606–615, 2012.

Martin HC Law, Mario AT Figueiredo, and Anil K Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, 2004.

Sharon Lee and Geoffrey J McLachlan. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24(2): 181–202, 2014.

Eemeli Leppäaho, Muhammad Ammad-ud din, and Samuel Kaski. GFA: exploratory analysis of multiple data sources with group factor analysis. *The Journal of Machine Learning Research*, 18(1):1294–1298, 2017.

Shawn E Levy and Richard M Myers. Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics*, 17:95–115, 2016.

Cen Li and Gautam Biswas. Unsupervised learning with mixed numeric and nominal data. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):673–690, 2002.

Kamil A Lipinski, Louise J Barber, Matthew N Davies, Matthew Ashenden, Andrea Sottoriva, and Marco Gerlinger. Cancer evolution and the limits of predictability in precision cancer medicine. *Trends in Cancer*, 2(1):49–63, 2016.

Z Liu, IS Peneva, F Evison, S Sahdra, DF Mirza, RM Charnley, R Savage, PA Moss, and KJ Roberts. Ninety day mortality following pancreatoduodenectomy in England: has the optimum centre volume been identified? *HPB*, 2018.

Silvia Liverani, David I Hastie, Lamiae Azizi, Michail Papathomas, and Sylvia Richardson. PReMiuM: An R package for profile regression mixture models using Dirichlet processes. *Journal of Statistical Software*, 64(7):1, 2015.

Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616, 2013.

Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523, 2013.

Harvey Lodish, Arnold Berk, James E Darnell, Chris A Kaiser, Monty Krieger, Matthew P Scott, Anthony Bretscher, Hidde Ploegh, Paul Matsudaira, et al. *Molecular Cell Biology*. Macmillan, 2008.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45(6):580, 2013.

Albert B Lowenfels and Patrick Maisonneuve. Risk factors for pancreatic cancer. *Journal of Cellular Biochemistry*, 95(4):649–656, 2005.

Albert B Lowenfels and Patrick Maisonneuve. Epidemiology and risk factors for pancreatic cancer. *Best Practice & Research Clinical Gastroenterology*, 20(2): 197–209, 2006.

Steven N MacEachern and Peter Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

Leigh-Ann MacFarlane and Paul R Murphy. MicroRNA: biogenesis, function and role in cancer. *Current Genomics*, 11(7):537–561, 2010.

Patrick Maisonneuve and Albert B Lowenfels. Risk factors for pancreatic cancer: a summary review of meta-analytical studies. *International Journal of Epidemiology*, 44(1):186–198, 2014.

Ryan Martin, Surya T Tokdar, et al. Asymptotic properties of predictive recursion: robustness and rate of convergence. *Electronic Journal of Statistics*, 3:1455–1472, 2009.

Simon J Mason and Nicholas E Graham. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128(584):2145–2166, 2002.

Peter McCullagh, Jie Yang, et al. How many clusters? *Bayesian Analysis*, 3(1): 101–120, 2008.

Geoffrey McLachlan and David Peel. *Finite Mixture Models.* John Wiley & Sons, 2004.

Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogianakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

James T McPhee, Joshua S Hill, Giles F Whalen, Maksim Zayaruzny, Demetrius E Litwin, Mary E Sullivan, Frederick A Anderson, and Jennifer F Tseng. Perioperative mortality for pancreatectomy: a national perspective. *Annals of Surgery*, 246(2):246, 2007.

Marina Meilă. Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Jeffrey Miller, Brenda Betancourt, Abbas Zaidi, Hanna Wallach, and Rebecca C Steorts. Microclustering: when the cluster sizes grow sublinearly with the size of the data set. *arXiv preprint arXiv:1512.00792*, 2015.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, (just-accepted):1–31, 2018.

Jeffrey W Miller and Matthew T Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural information Processing Systems*, pages 199–206, 2013.

Boris Mirkin. Eleven ways to look at the chi-squared coefficient for contingency tables. *The American Statistician*, 55(2):111–120, 2001.

Kevin J Mitchell. What is complex about complex disorders? *Genome Biology*, 13 (1):237, 2012.

Qianxing Mo and Ronglai Shen. iClusterPlus: Integrative clustering of multi-type genomic data. 2016. R package version 1.14.0.

Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

Qianxing Mo, Ronglai Shen, Cui Guo, Marina Vannucci, Keith S Chan, and Susan G Hilsenbeck. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 2017.

Dharmendra S Modha and W Scott Spangler. Feature weighting in k-means clustering. *Machine Learning*, 52(3):217–237, 2003.

Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian exponential family PCA. In *Advances in Neural Information Processing Systems*, pages 1089–1096, 2009.

Silvia Monticelli, K Mark Ansel, Changchun Xiao, Nicholas D Socci, Anna M Krichevsky, To-Ha Thai, Nikolaus Rajewsky, Debora S Marks, Chris Sander, Klaus Rajewsky, et al. MicroRNA profiling of the murine hematopoietic system. *Genome Biology*, 6(8):R71, 2005.

Leslie C Morey and Alan Agresti. The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement*, 44(1):33–37, 1984.

Peter Müller and Riten Mitra. Bayesian nonparametric inference–why and how. *Bayesian Analysis (Online)*, 8(2), 2013.

Kevin P Murphy. Conjugate Bayesian analysis of the Gaussian distribution. *Technical report*, 2007.

Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT Press, 2012.

National Cancer Institute. Hormone therapy for breast cancer, 2018a. URL `https://www.cancer.gov/types/breast/breast-hormone-therapy-fact-sheet`.

National Cancer Institute. Cancer Staging, 2018b. URL `https://www.cancer.gov/about-cancer/diagnosis-staging/staging`.

National Human Genome Research Institute. The cost of sequencing human genome, 2018. URL `https://www.genome.gov/sequencingcosts/`. [Online; accessed 10-May-2018].

Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Michael A Newton and Yunlei Zhang. A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, 86(1):15–26, 1999.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002.

NHS Digital. Hospital Episode Statistics. `https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics`, 2018. [Online; accessed 01-June-2018].

Agostino Nobile. Bayesian analysis of finite mixture distributions. 1996.

Agostino Nobile and Alastair T Fearnside. Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Statistics and Computing*, 17(2): 147–162, 2007.

Office for National Statistics. Cancer registration statistics, england: 2017, 2017. URL `https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/2017`.

Adrian O'Hagan, Thomas Brendan Murphy, and Isobel Claire Gormley. Computational aspects of fitting mixture models via the expectation–maximization algorithm. *Computational Statistics & Data Analysis*, 56(12):3843–3864, 2012.

Akio Onogi, Masanobu Nurimoto, and Mitsuo Morita. Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinformatics*, 12(1):263, 2011.

Purvi Parikh, Mira Shiloach, Mark E Cohen, Karl Y Bilimoria, Clifford Y Ko, Bruce L Hall, and Henry A Pitt. Pancreatectomy risk calculator: an ACS-NSQIP resource. *Hpb*, 12(7):488–497, 2010.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Michael J Pencina and Ralph B D'Agostino. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, 2004.

Francesca Petralia, Vinayak Rao, and David B Dunson. Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2012.

Sonia Petrone and Adrian E Raftery. A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statistics & Probability Letters*, 36(1):69–83, 1997.

G Philip and BS Ottaway. Mixed data cluster analysis: an illustration using Cypriot hooked-tang weapons. *Archaeometry*, 25(2):119–133, 1983.

Jim Pitman et al. Combinatorial stochastic processes. 2002.

David Martin Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2011.

Zhaohui S Qin, Lee Ann McCue, William Thompson, Linda Mayerhofer, Charles E Lawrence, and Jun S Liu. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nature Biotechnology*, 21(4):435, 2003.

Fernando A Quintana and Michael A Newton. Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences. *Journal of Computational and Graphical Statistics*, 9(4):711–737, 2000.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

Lola Rahib, Benjamin D Smith, Rhonda Aizenberg, Allison B Rosenzweig, Julie M Fleshman, and Lynn M Matrisian. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Research*, 2014.

Sara Raimondi, Patrick Maisonneuve, and Albert B Lowenfels. Epidemiology of pancreatic cancer: an overview. *Nature Reviews Gastroenterology and Hepatology*, 6(12):699–708, 2009.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Carl Edward Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*, volume 12, pages 554–560, 1999.

Surajit Ray and Bruce G Lindsay. The topography of multivariate normal mixtures. *Annals of Statistics*, pages 2042–2065, 2005.

Richard Redon, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, Michael H Shapero, Andrew R Carson, Wenwei Chen, et al. Global variation in copy number in the human genome. *nature*, 444 (7118):444, 2006.

Sylvia Richardson and Peter J Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997.

Christian P Robert and George Casella. The Metropolis—Hastings algorithm. In *Monte Carlo Statistical Methods*, pages 231–283. Springer, 1999.

Keith D Robertson. DNA methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610, 2005.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, Markus Müller, Stefan Siegert, and Maintainer Xavier Robin. Package 'pROC', 2018.

Sandra Rodríguez-Rodero, Agustín F Fernández, Juan Luís Fernández-Morera, Patricia Castro-Santos, Gustavo F Bayon, Cecilia Ferrero, Rocio G Urdinguio, Rocío Gonzalez-Marquez, Carlos Suarez, Iván Fernández-Vega, et al. DNA methylation signatures identify biologically distinct thyroid cancer subtypes. *The Journal of Clinical Endocrinology & Metabolism*, 98(7):2811–2821, 2013.

David Rossell and Donatello Telesca. Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265, 2017.

David Rossell, Donatello Telesca, and Valen E Johnson. High-dimensional Bayesian classifiers using non-local priors. In *Statistical Models for Data Analysis*, pages 305–313. Springer, 2013.

Sam T Roweis. EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems*, pages 626–632, 1998.

Donald B Rubin and Dorothy T Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

Richard S Savage, Zoubin Ghahramani, Jim E Griffin, Bernard J De La Cruz, and David L Wild. Discovering transcriptional modules by Bayesian data integration. *Bioinformatics*, 26(12):i158–i167, 2010.

Richard S Savage, Zoubin Ghahramani, Jim E Griffin, Paul Kirk, and David L Wild. Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. *arXiv preprint arXiv:1304.3577*, 2013.

Markus S Schröder, Aedín C Culhane, John Quackenbush, and Benjamin Haibe-Kains. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, 27(22):3206–3208, 2011.

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

Luca Scrucca, Michael Fop, Thomas Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):205–233, 2017. URL `https://journal.r-project.org/archive/2017/RJ-2017-008/RJ-2017-008.pdf`.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

Ronglai Shen, Adam B Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

Christiana Spyrou, Rory Stark, Andy G Lynch, and Simon Tavaré. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics*, 10(1):299, 2009.

Nicolas Städler, Frank Dondelinger, Steven M Hill, Rehan Akbani, Yiling Lu, Gordon B Mills, and Sach Mukherjee. Molecular heterogeneity at the network level: high-dimensional testing, clustering and a TCGA case study. *Bioinformatics*, 33 (18):2890–2896, 2017.

Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, volume 400, pages 525–526. Boston, 2000.

PM Steiner and Marcus Hudec. Classification of large data sets with mixture models via sufficient EM. *Computational Statistics & Data Analysis*, 51(11):5416–5428, 2007.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3(Dec):583–617, 2002.

Gero Szepannek. clustMixType: k-prototypes clustering for mixed variable-type data. 2018. URL `https://CRAN.R-project.org/package=clustMixType`.

Mahlet G Tadesse, Naijun Sha, and Marina Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617, 2005.

Cong Tan and Xiang Du. KRAS mutation testing in metastatic colorectal cancer. *World journal of gastroenterology: WJG*, 18(37):5171, 2012.

Bernd Taschler, Frank Dondelinger, and Sach Mukherjee. Model-based clustering in very high dimensions via adaptive projections. *arXiv:1902.08472*, 2019.

Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, pages 1385–1392, 2005.

Yee Whye Teh. Dirichlet process. In *Encyclopedia of Machine Learning*, pages 280–287. Springer, 2011.

Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530, 2002.

William N Venables and Brian D Ripley. *Modern Applied Statistics with S-PLUS*. Springer Science & Business Media, 2013.

Raghunandan Venkat, Milo A Puhan, Richard D Schulick, John L Cameron, Frederic E Eckhauser, Michael A Choti, Martin A Makary, Timothy M Pawlik, Nita Ahuja, Barish H Edil, et al. Predicting the risk of perioperative mortality in patients undergoing pancreaticoduodenectomy: a novel scoring system. *Archives of Surgery*, 146(11):1277–1284, 2011.

Mukesh Verma. Personalized medicine and cancer. *Journal of Personalized Medicine*, 2(1):1–14, 2012.

Seppo Virtanen, Arto Klami, Suleiman Khan, and Samuel Kaski. Bayesian group factor analysis. In *Artificial Intelligence and Statistics*, pages 1269–1277, 2012.

Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 339(6127): 1546–1558, 2013.

Stefano Volinia, George A Calin, Chang-Gong Liu, Stefan Ambs, Amelia Cimmino, Fabio Petrocca, Rosa Visone, Marilena Iorio, Claudia Roldo, Manuela Ferracin, et al. A microRNA expression signature of human solid tumors defines cancer gene targets. *Proceedings of the National Academy of Sciences*, 103(7):2257–2261, 2006.

Charles Mahlon Vollmer, Norberto Sanchez, Stephen Gondek, John McAuliffe, Tara S Kent, John D Christein, Mark P Callery, Pancreatic Surgery Mortality Study Group, et al. A root-cause analysis of mortality following major pancreatectomy. *Journal of Gastrointestinal Surgery*, 16(1):89–103, 2012.

Silke Wagner and Dorothea Wagner. *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

AM Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 80–88, 1969.

Hanna Wallach, Shane Jensen, Lee Dicker, and Katherine Heller. An alternative prior process for nonparametric Bayesian clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 892–899, 2010.

Chong Wang and David M Blei. A split-merge MCMC algorithm for the hierarchical Dirichlet process. *arXiv preprint arXiv:1201.1657*, 2012.

Lianming Wang and David B Dunson. Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011.

Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, and Bill Venables. gplots: Various R programming tools for plotting data. 2016. URL `https://CRAN.R-project.org/package=gplots`. R package version 3.0.1.

Sumio Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897, 2013.

Claus Weihs, Uwe Ligges, Karsten Luebke, and Nils Raabe. klaR: Analyzing german business cycles. In D. Baier, R. Decker, and L. Schmidt-Thieme, editors, *Data Analysis and Decision Support*, pages 335–343, Berlin, 2005. Springer-Verlag.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The Cancer Genome Atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

Welsh Cancer Intelligence and Surveillance Unit. Cancer in Wales, 2018. URL `https://www.wcisu.wales.nhs.uk/sitesplus/documents/1111/CANCERinWALESapril2014FINAL%28Eng%29.pdf`.

Mike West. *Hyperparameter Estimation in Dirichlet Process Mixture Models*. Duke University ISDS Discussion Paper# 92-A03, 1992.

Mike West and Michael D Escobar. *Hierarchical priors and mixture models, with application in regression and density estimation.* Institute of Statistics and Decision Sciences, Duke University, 1993.

World Health Organisation. Cancer key facts. `http://www.who.int/en/news-room/fact-sheets/detail/cancer`, 2018. [Online; accessed 02-Oct-2018].

Yuefeng Wu and Subhashis Ghosal. The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101 (10):2411–2419, 2010.

Dhiraj Yadav and Albert B Lowenfels. The epidemiology of pancreatitis and pancreatic cancer. *Gastroenterology*, 144(6):1252–1261, 2013.

Yinyin Yuan, Richard S Savage, and Florian Markowetz. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Computational Biology*, 7(10):e1002227, 2011.

Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, 13(Apr):1059–1062, 2012.

Yong Zhao, Eva Samal, and Deepak Srivastava. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature*, 436 (7048):214, 2005.

Hui Zhou, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic journal of statistics*, 3:1473, 2009.

Shunzhi Zhu, Dingding Wang, and Tao Li. Data clustering with size constraints. *Knowledge-Based Systems*, 23(8):883–889, 2010.

Mark H Zweig and Gregory Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39 (4):561–577, 1993.