A Bayesian Nonparametric Model for RNA-Sequencing Data

by

Matthieu Lepur

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Statistics)

The University of British Columbia

(Vancouver)

October 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

A Bayesian Nonparametric Model for RNA-Sequencing Data

submitted by **Matthieu Lepur** in partial fulfillment of the requirements for the degree of **Master of Science** in **Statistics**.

Examining Committee:
Alexandre Bouchard-Côté, Statistics, UBC
Supervisor
Andrew Roth, Computer Science, UBC
Supervisory Committee Member

Abstract

Cancers from different tissue types can share a latent structure reflecting commonly altered gene pathways. It is difficult to cluster cancer patients based on this latent structure because the tissue of origin often dominates the latent structure effect. We propose a Bayesian nonparametric model that accounts for the tissue effect and clusters based on a latent structure using a Dirichlet Process prior. More specifically, we use an infinite Gaussian mixture model where the mean parameter is modelled as the linear combination of tissue, gene, and latent cluster effects. The choice of the Dirichlet Process prior allows us to side-step a model selection problem as the number of latent clusters is unknown apriori. Our approach learns the tissue effect by using tissue parameters in a supervised learning setting, while simultaneously learning the latent structure based on the residuals in an unsupervised setting. These so-called residuals result from subtracting out the inferred tissue and gene parameters from the observations and can be interpreted as the cluster effect. A key component of the model is its ability to leverage conjugacy between the likelihood model and cluster parameters. The Gaussian form of the model is not effect by our choice of mean parameter therefore conjugacy is preserved. Indeed, the model has the intuitive interpretation of clustering on the cluster effect signal that remains subtracting out the tissue and gene effects. Conjugacy allows for the use of sophisticated Markov chain Monte Carlo techniques used in Bayesian mixture models such as *Split-Merge* samplers. We demonstrate our model by showing results on synthetic data, semi-synthetic data generated using a publicly available dataset from the Genome-Tissue Expression (GTEx) portal, and another publicly available dataset from the International Cancer Genome Consortium (ICGC).

Lay Summary

Cancers from different tissue types can share biological similarities that may be hidden by the tissue of origin. For example, there may be an underlying similarity between lung and liver cancers that are not obvious because the data reflects the tissue types. We propose a Bayesian statistical model to understand and learn these hidden relationships between different cancer types. Our model first learns the tissue effect as it is the most obvious then learns the hidden effect. Using our model we are able to group cancer patients based on these hidden similarities and learn meaningful biological information. We demonstrate our model by showing results on generated data, semi-generated data, and a publicly available real-dataset from the International Cancer Genome Consortium (ICGC).

Preface

This dissertation is original and unpublished work by the author, Matthieu Lepur, supervised by Professor Alexandre Bouchard-Côté and Professor Andrew Roth. The model introduced in Chapter 3 and approximate inference methodology outlined in Chapter 4 was jointly designed with A. Bouchard-Côté and A. Roth. Software developed for experiments conducted in Chapter 5 was contributed by the author.

Table of Contents

List of Figures

Chapter 1

Introduction

The past 20 years have seen an explosion in high throughput biological data. Microarray and sequencing based technologies have created the opportunity to study many complex biological systems in detail. One of the most successful areas of application of these technologies has been in the study of cancer biology. Early studies using microarrays to measure RNA expression of tumour tissue provided numerous insights into the mechanisms of cancer development and have allowed clinicians to develop prognostic tools for assessing patient risk, and informing treatment strategies (Tinker et al., 2006). More recently the development of high throughput sequencing assays has allowed cancer researchers to study the transcriptional state of tumours in an unbiased way (Wang et al., 2009; Desmedt et al., 2012). As these sequencing technologies have matured and costs have fallen, large scale sequencing projects such as The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al., 2013) and International Cancer Genome Consortium (ICGC) (Zhang et al., 2019) have generated massive multi-modal datasets from thousands of patients across a range of cancer types. These massive "pan-cancer" datasets have the potential to allow cancer researchers to identify key drivers of cancer development and progression. Ultimately these insights could lead to improved predictors of cancer risk and the develop-

ment of better therapeutic agents.

The goal of this thesis is to develop a statistical method for performing pan-cancer analysis of gene expression data. We hypothesize that by analyzing the gene expression data from thousands of patients we will increase our statistical power to detect groups of tumours which behave similarly. As we discuss later, existing approaches based on clustering gene expression data are not suitable to solve this problem. The core challenge for such a pan-cancer clustering analysis is that tumours will predominantly cluster by the tissue of origin. This can be seen by hierarchical clustering 3.2 where the coarsest separation of tumours is driven by tissue of origin (Chen et al., 2018). As result, classical clustering effectively degenerates into tissue specific analysis, removing the potential to share statistical strength. To address this problem we will develop a model which, informally, removes the tissue effect and clusters the residual expression. A secondary problem we also address is that of model selection i.e. how many clusters should be used. To achieve this we develop a hierarchical Bayesian model based on a Dirichlet Process mixture model. As exact inference for this model is intractable, we use Markov Chain Monte Carlo (MCMC) methods to perform approximate posterior inference. We benchmark our approach using several synthetic datasets. Finally, we apply our method to perform a pilot analysis of 125 patients from the ICGC project.

1.1 Domain background

In this section we briefly outline information relevant to the domain application. Specifically, we outline basic biological concepts necessary to understand the work and the details of the measurement technologies and analysis strategies currently employed. A discussion of the statistical background is presented in Chapter 2.

1.1.1 Basic biology

We briefly review some core biological concepts for the reader. The most important biological concept to understand for this thesis is the central dogma of biology. The central dogma is a model of information flow in biological systems. In this model DNA is a stable and heritable molecule responsible for encoding all of the information a cell needs to survive. The complete set of DNA in a cell (genome) is copied when cells divide and identical copies are passed on to descendants. The information DNA encodes are the instructions for producing proteins. Proteins can be thought of as molecular machines which perform biological functions critical to the survival of the cell. The set and quantity of proteins produced by a cell dictate its behaviour or phenotype. Though a cell contains the instruction to produce a vast set of proteins, only a subset is typically produced. By varying the selection of proteins produced, multicellular organisms such as humans can thus produce cells with diverse functions which can then lead to the formation of complex organs.

To produce proteins, DNA must first be transcribed into RNA, an intermediary information molecule. RNA molecules are then read by special proteins which translate the information and produce other proteins. Cells can thus control the level of protein synthesis by controlling the level of RNA synthesis. As an analogy the reader can imagine DNA as the central server for the cell. RNA molecules act like USB sticks which can download specific files which can then be uploaded to the machines responsible for building proteins. Without RNA, only one protein synthesis machine could read the required information to produce a protein. The core point to take away from this analogy is that the levels of RNA expression provide a proxy measurement for protein levels, which in turn inform on the state of the cell.

In this work we are interested in studying gene expression data from cancerous tumours from humans. Cancer is a disease which occurs when healthy cells acquire mutations in their genomes. The mutations lead to aberrant behaviour such as uncontrolled growth and evasion of typical signals

that cause damaged cells to die. As a cancer develops the population of malignant (cancerous) cells acquire additional mutations which can allow for more complex behaviours, such as the ability to avoid destruction by immune cells. The genomic mutations change the composition of RNA and ultimately proteins produced by cancer cells, which we will refer to as a gene expression profile. We will use gene expression analysis to group tumours which we hypothesize have similar behaviour.

1.1.2 Gene expression analysis

Gene expression data, that is measurement of RNA abundance from a tissue sample, is a powerful approach to studying many biological processes. Gene expression is commonly performed using microarrays or more recently high throughput sequencing. There are many types of analyses and experimental designs that can be used with gene expression data. In this work we focus on unsupervised analysis of gene expression data from cancer patient tumours to identify latent group structures. As discussed later we are specifically interested in bulk expression analysis from tumour tissues.

The classic approach for discovering latent group structure in expression datasets is to cluster the data using approaches such hierarchical clustering, k-means or Gaussian mixture models. In the cancer context these latent groups are usually referred to as subtypes. Identifying subtypes is important because tumours which show similar gene expression profiles will likely behave in a similar way. One of the best examples of this type of analysis and its clinical utility comes from breast cancer. Using early gene expression microarray technologies, researchers were able to identify subtypes of breast cancer with different clinical characteristics and outcome (Sørlie et al., 2001; Van't Veer et al., 2002). Ultimately a panel of 50 genes was identified to stratify breast cancer patients into different subtypes and developed into a clinical assay (Parker et al., 2009). This stratification has been used to inform the treatment strategies for breast cancer patients.

Biologically grouping patients by gene expression reflects a hypothesis that there are specific biological pathways, that is interacting sets of genes, that are perturbed. Depending on the affected pathways cancer cells will behave differently in terms of growth, migration potential and response to treatment. Thus by clustering tumours with similar expression profiles we can gain insight into the perturbed pathways and behavioural similarities.

A key challenge for performing clustering analysis to identify cancer subtypes for patient stratification, is assembling a sufficiently large cohort to be adequately powered to detect groups. For common cancers such as breast, colorectal and prostate it is usually feasible due to the high number of patients presenting each year. However, many other types of cancer occur less frequently or are not routinely biopsied, thus assembling large cohorts becomes challenging. We hypothesize that by leveraging large pan-cancer datasets we can identify subtypes that span across different tissues of origin. This hypothesis is based on the observation that genes in the same pathways are frequently mutated across different cancer types, which suggests these cancers with these mutations could behave in a similar way. Identifying group structure across cancer types also has the potential to generate hypotheses about drug repurposing. For example, if a subtype spans multiple cancer types and all of the cancers from one tissue of origin have a drug approved for use, we could hypothesize the same drug may be useful for treating the other tumours from different tissues assigned to that subtype.

1.1.3 Pan-cancer analysis

Pan-cancer analysis refers to any analysis which considers tumours from multiple tissues of origin i.e. breast, blood, liver etc. Pan-cancer analysis of DNA or genomic sequencing data is now well established. Recent pan-cancer studies have identified recurrent mutations across cancers types, common mutational patterns and modes of evolution. While mutational status provides one way to define subtypes, it does not always reflect similarities in tumour

behaviour. Gene expression profiles can be useful in these cases to get a more accurate functional read out of tumour state. Pan-cancer analysis of RNA expression data has received less attention however. This is largely due to the challenge of cancer type heterogeneity and in particular the strong effect that the tissue of origin has on the observed expression profile of a tumour. This tissue effect, as we will refer to it, stems from two dominant factors.

The first factor leading to tissue effect is related to how gene expression measurements are performed. Projects such as the TCGA and ICGC have predominantly used so-called bulk expression assays. A bulk assay consists of taking a piece of tumour tissue consisting of tens of thousands to millions of cells, breaking the cells apart and releasing their RNA for subsequent measurement. As a result, these assays measure mixture of expression profiles from the constituent cells within the tumour tissue. Within this mixture there will frequently be a significant proportion of cells specific to a given tissue. There will also be additional cells from elsewhere in the body, in particular immune cells. The composition and activity of these cells is highly relevant for predicting tumour growth and response to treatment. This is one reason bulk gene expression analysis may provide more refined patient stratification than mutation based approaches.

The second factor which contributes to the tissue effect relates to how cancer cells develop. Cancer cells develop from a healthy cell which has acquired mutations which allow for abnormal growth and survival. While the mutations tend to shift the expression profile of the malignant cells away from the progenitor normal cell, a residual imprint of the progenitor expression usually remains. Thus even if we could isolate individual cancer cells and measure their expression, this residual expression would mask the pathways which are perturbed leading to cancer.

1.2 Related work

Patient stratification of cancer patients using gene expression profiles is a well studied problem. A range of stratification schemes have been developed in cancers such as breast, colorectal and prostate (Colombo et al., 2011; Marisa et al., 2013; Lapointe et al., 2004). Several studies have also explored incorporating additional features such a mutational status along with gene expression to provide more refined subtypes (Curtis et al., 2012). Much of the work to date using gene expression data has focused on a single tissue of origin for reasons outlined in Section 1.1.3.

Pan-cancer analysis of gene expression remains under explored. Most pan-cancer approaches focus on measurements which are not affected by tissue of origin such as mutations (Sharma et al., 2019; Campbell et al., 2020; Alexandrov et al., 2020) and immune cell composition (Chakravarthy et al., 2018; Thorsson et al., 2018). One notable exception was (Chen et al., 2018) which aimed to perform pan-cancer subtyping based on gene expression and methylation data. The authors of that study first used standard k-means clustering of the gene expression data from 32 tissue types to identify subtypes. As expected this analysis identified that tissue of origin was the dominant factor driving subtype assignment. The authors next attempted to remove the tissue of origin effect by pre-processing the data and extracting the mean expression value by tissue from each sample. This is conceptually similar to the approach we have taken. However, our approach models the tissue specific effect probabilistically allowing for uncertainty in this value to be accounted for during clustering. Another difference is our approach to selecting the number of clusters. The authors of (Chen et al., 2018) use an ad-hoc criterion to identify the optimal number clusters using k-means. By contrast we use the formalism of Dirichlet Process mixture models to automatically infer the number of clusters as part of our statistical inference procedure.

Feature allocation based approaches for analyzing gene expression data

are also conceptually similar to our model. Feature allocations models can be seen as a generalization of clustering models. To see the connection imagine each data point has an associated binary vector of length K, where K is the number of clusters or number of features respectively. In a clustering model exactly one entry of this vector is one and the rest are zero, indicating mutually exclusive group membership. For feature allocation models any number of entries in this vector can be set to one, allowing for overlapping group membership. One could imagine using a feature allocation approach to capture tissue specific effects with a subset of features and using additional features to capture non-tissue specific effects. Though we are not aware of anyone having done this, feature allocations have been used to analyze gene expression data from breast cancer (Xu et al., 2016). We believe that further application in the pan-cancer context is an exciting avenue for future work.

Chapter 2

Background

The aim of this chapter is to outline the mathematical notation for the remainder of this thesis and provide the necessary background of Bayesian Nonparametric models to understand the proposed model. Section 2.1 discusses a specific case of Bayesian mixture models. That is, the case in which the cluster parameters θ can be integrated out. Section 2.2 discusses the Dirichlet Process as a prior distribution on possible partitions of our observations and its properties. Followed by one possible construction of the Dirichlet Process called the Chinese Restaurant Process. We conclude by bringing together Bayesian mixture models and the Dirichlet Process to construct the so-called Dirichlet Process mixture model in Section 2.3.

2.1 Bayesian mixture models

Suppose we have N observations denoted by $\mathbf{x} = \{x_1, x_2, ..., x_N\}$. A mixture model assumes that the data is generated by a mixture of distributions. In other words, a mixture model assumes that observations can be grouped into subsets in which members from the same subset are generated by the same constituent distribution. More precisely, the observation indices $[N] = \{1, 2, ..., N\}$ are partitioned into subsets $c_k \subseteq [N]$ called *clusters* where $|c_k|$

denotes the cardinality of cluster k. A partition of observations is a collection of clusters and is referred to as a *clustering* denoted by $c = \{c_1, c_2, ..., c_K : c_k \subseteq [N]\}$ where $\bigcup_{c_k \in c} c_k = [N]$ and K denotes the total number of clusters. Following the notation of Bouchard-Côté et al. (2017), we condition on a clustering and we define the likelihood of the data as

$$L(\mathbf{x}|c) = \prod_{k=1}^{K} L(\mathbf{x}_k) \tag{2.1}$$

where $L(\mathbf{x}_k)$ is the likelihood of observations assigned to cluster k and defined by

$$L(\mathbf{x}_k) = \int \left(\prod_{i \in c_k} L(x_i|\theta) \right) H(d\theta) \tag{2.2}$$

where $L(x_i|\theta)$ is the likelihood of observation x_i parameterized by θ. Notice we focus on the case of Bayesian mixture models where the cluster parameters θ can be integrated out. This should be intuitive as the likelihood of the data is simply the product of the likelihoods from each cluster.

We can combine the likelihood $L(\mathbf{x}|c)$ and prior $p(c)$ to obtain the following posterior on the clustering

$$p(c|\mathbf{x}) \propto L(\mathbf{x}|c)p(c). \tag{2.3}$$

This prior distribution will be selected as the Dirichlet Process prior on the set of possible clusterings of our observations. We discuss this prior distribution in the following section.

2.2 Dirichlet Process

In the Bayesian paradigm unknown quantities are treated as random variables. Therefore, we treat the clustering c as a random variable and place

a prior distribution $p(c)$ on its domain. A common problem encountered in mixture models is the number of components K is unknown. Consequently, we also treat K as a random variable. We can do so by selecting a prior distribution that places non-zero probability on clusterings of all possible sizes. This includes cluster sizes in the range from one to N where one would indicate all observations clustered together and N would indicate each observation clustered in their own cluster, called the singleton clusters. To this end, the Dirichlet Process achieves these objectives as a prior distribution. It is a prior distribution on the set of all possible clusterings of our observations of any size.

We outline one possible construction of the Dirichlet Process called the Chinese Restaurant Process. We begin with the Chinese Restaurant Process for pedagogical purposes as it provides an intuitive analogy to construct the Dirichlet Process. There does exist other constructions of the Dirichlet process such as the the Stick-Breaking Process construction (Teh, 2010). In fact, the stick breaking construction is advantageous as it can be simply modified to construct other prior distributions such as the Pitman-Yor Process (Orbanz, 2015). Therefore, we suggest constructing the Dirichlet process through the lens of the Stick-Breaking Process if the reader is interested in other prior distributions.

The Chinese Restaurant Process can be best explained through an analogy involving a chinese restaurant. In this analogy, observations and clusters will be referred to by customers and tables, respectively. Consider N customers (observations) sequentially entering a chinese restaurant. To begin the stochastic process, the first customer sits at a table (cluster) by themselves. Followed by the second customer, that joins the table with the first customer or sits at a new table by themselves. The n^{th} customer enters the restaurant and decides to join any existing tables with probability proportional to their popularity or it's own table with probability proportional to a hyper-parameter α. Notice, if we stop this stochastic process at any iteration

a random partition is defined for the number of customers currently seated in the restaurant. That is, we have a random partition defined for n observations where n is the number of customers seated in the restaurant and the cluster assignments are indicated by table assignments. We can write the distribution of the table assignment for the n^{th} observation as follows

$$z_n \sim \sum_{k=1}^{K} \frac{|c_k|}{n-1+\alpha}\delta_k + \frac{\alpha}{n-1+\alpha}\delta_{K+1} \tag{2.4}$$

where K denotes the current number of existing clusters, $|c_k|$ denotes the number of observations assigned to table k, and δ_k is the point mass at cluster k. The Chinese Restaurant Process induces a probability distribution on the possible clusterings of N data points. More precisely, the probability of any clustering c of N observations is given by

$$p(c) = \frac{\alpha^{|c|}\prod_{k=1}^{K}(|c_k|-1)!}{\prod_{n=1}^{N}(n-1+\alpha)}.$$

This can be seen as a specific case of a prior distribution taking the following form (Bouchard-Côté et al., 2017)

$$p(c) \propto \tau_1(|c|)\prod_{k=1}^{K}\tau_2(|c_k|) \tag{2.5}$$

where $\tau_1(|c|) = \alpha^{|c|}$ and $\tau_2(|c_k|) = (|c_k|-1)!$. From this perspective, one can see that prior distribution of this form depend on two main components: the number and size of each cluster. The Pitman-Yor Process and Finite Dirichlet distribution are both special cases of prior distribution that take the form of Equation 2.5.

The Dirichlet Process has a several properties that make it useful as a prior distribution. Namely, the Dirichlet Process exhibits a "rich get richer" phenomenon, is exchangeable, and learns the number of clusters in a data

informed manner. The "rich get richer" property implies that the larger a cluster gets the more likely a new observation will join that cluster. This can be seen in Equation 2.4 as the probability of joining an existing cluster is directly proportional to the size of a cluster. Conversely, every customer has the opportunity to create a new cluster as there is a non-zero probability of creating a new cluster that is directly proportional to the hyper-parameter α. That is, as we increase α the more clusters we will observe and vice-versa. Intuitively, these two observations suggest that the structure imposed by a Dirichlet Process is a few larger clusters with some smaller clusters where some is related to the value of the hyper-parameter α. More precisely, the expected number of clusters is $O(\alpha \log(N))$ (Teh, 2010). Despite the sequential nature of the Chinese Restaurant Process the Dirichlet Process is exchangeable. That is, the probability distribution for the Dirichlet Process does not take into account the order of cluster assignments. This becomes particularly useful when sampling the posterior. In a nutshell, we are able to treat the cluster assignment being updated as the new observation given all other cluster assignments as seen in Section 4.2. Perhaps the most appealing feature of the Dirichlet Process prior is that the number of clusters is learned from the data. That is, the number of clusters changes throughout inference depending on signal in the data. In this way, the Dirichlet Process is able to side-step a model selection problem through the use of a fully Bayesian treatment of modelling.

2.3 Dirichlet Process mixture models

The Dirichlet Process is often used as a prior distribution on cluster configuration in Bayesian mixture models for unsupervised learning problems such as clustering. The ability of the Dirichlet Process to allow for a countably infinite number of clusters manifests itself in a Bayesian mixture model by allowing for a countably infinite number of component distributions. In

this regard, the Dirichlet Process mixture model is referred to as an *infinite* mixture model. The possibility of an infinite number of clusters allows a Dirichlet Process mixture model to learn the number of clusters dynamically from the data. That is, the number of clusters changes throughout posterior inference depending on signal in the data. In this way, the Dirichlet Process is able to side-step a model selection problem through the use of a fully Bayesian treatment of modelling. In contrast, finite mixture models specify the number of clusters and model selection is often performed to determine the number of clusters.

Chapter 3

The Gene Expression Model

3.1 Problem statement

We consider the problem of performing pan-cancer gene expression subtyping. As discussed in Section 1.1.3, cancer subtyping can be seen as a clustering problem. By clustering gene expression data we can identify groups of tumours with similar gene expression profiles. In Section 1.1.3 we argue that while straightforward when considering tumours from a single tissue of origin, the problem becomes challenging in a pan-cancer setting with tumours from multiple tissue types. The core issue is that the tissue of origin has a dramatic effect on gene expression. Thus tumours from the same tissue tend to cluster together regardless of the changes driving the cancer. While the tissue effect is biologically important, it masks potentially more interesting expression patterns related directly to the drivers of a cancer. Our hypothesis is that if the tissue effect could be controlled for, then we could identify subtle latent structures which can be used to group cancers from different tissues. This grouping may better reflect the underlying biology of the tumours and provide insights into the shared mechanisms driving the tumours within a group. Ultimately such information could also be used to inform treatment strategies which were not dependent on the tissue of origin.

To formalize the problem let $\boldsymbol{X}$ denote the dataset where observations (tumours) are rows and features (genes) are columns. Further suppose there is a known obvious cluster structure (tissue effect) in $\boldsymbol{X}$ that dominates the signal in the data. We will refer to this clustering as the *primary* clustering. We hypothesize that there exists a latent structure that is hidden or masked by the primary clustering. We will refer to this latent clustering as the *secondary* clustering. Once we control for the primary clustering a meaningful secondary clustering will be revealed. This secondary structure could contain interesting information regarding the data generating process that is not initially available. Bulk RNA-sequencing data falls into this data setting where the primary clustering is often the tissue of origin that masks secondary clustering which is another biological phenomenon such as affected gene pathways. We observe in figure 3.1 that Bulk RNA-sequencing data is dominated by the tissue of origin. This can be seen on the left most side of the data matrix as the tissue assignments are labelled with colours.

If we perform classical clustering techniques such as hierarchical clustering, finite mixture modelling, or infinite mixture modelling we will observe that these methods predict cluster assignments according to the primary clustering. This can be seen in figure 3.2 as the cluster labels directly correspond to tissue labels. This figure shows the cluster labels predicted using Scikit-learn's implementation of the Dirichlet Process Gaussian mixture model (Pedregosa et al., 2011). The implementation uses a variational approximation to the posterior of the mixture model (Blei and Jordan, 2006). The same clustering was predicted when using Scikit-learn's implementation of the k-means algorithm and a finite Gaussian mixture model. We also observe hierarchical clustering of the observations reveals a structure that corresponds to the tissue of origin.

In this data setting, the goal is to control for the primary clustering and unmask the secondary clustering. In other words, we want a model that captures the primary and secondary cluster effects. We can use the

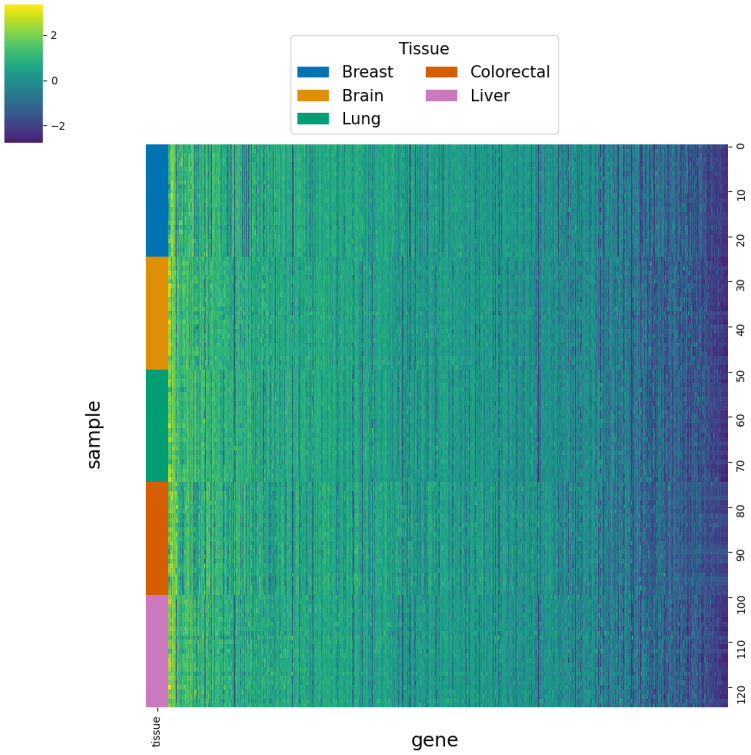


Figure 3.1: Bulk RNA-sequencing data of breast, brain, liver, colorectal, and lung cancer obtained from the ICGC portal where each cancer type has 25 samples.

inferred cluster parameters of both clusterings to better understand the data generating process. We want the primary clustering to be known apriori and added to the model as a supervised component. The inferred cluster assignments of the secondary clustering can be used to determine meaningful relationships between observations unrelated to the the primary clustering. Ultimately, we will use a Dirichlet Process Gaussian mixture model to model the gene expression matrix $\boldsymbol{X}$.

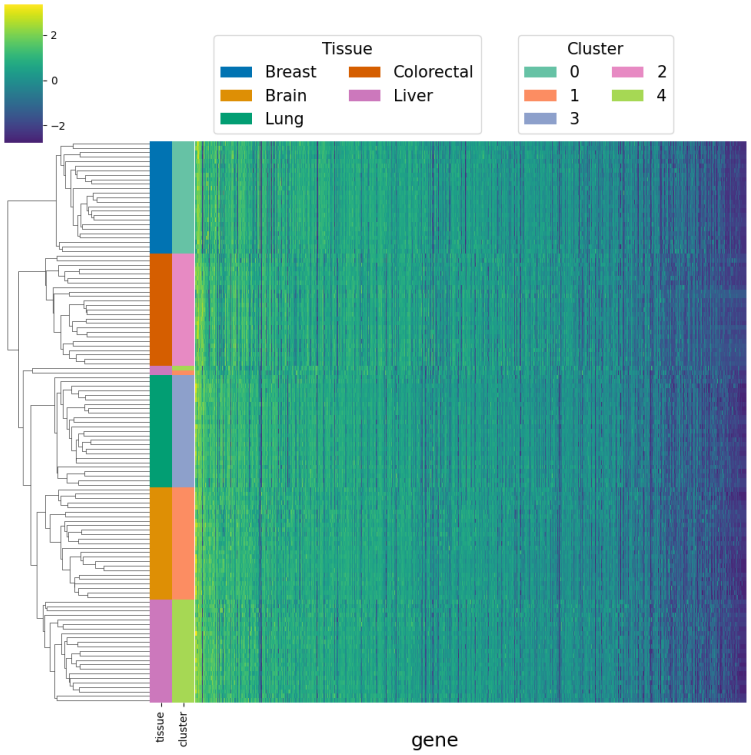


Figure 3.2: Bulk RNA-sequencing data of breast, brain, liver, colorectal, and lung cancer obtained from the ICGC portal with tissue and cluster labels. Cluster labels were predicted using Scipy's Dirichlet Process Gaussian mixture model.

3.2 Observation model

Section 3.5 discusses the normalization technique used for RNA-sequencing data. Each entry of the gene expression matrix $\mathbf{x}_{nm}$ is assumed to be distributed according to a Gaussian distribution with mean and variance parameters $(\mu_{nm}, \sigma^2_{nm})$. The mean μ_{nm} will be a function of parameters that span across samples and genes allowing information to be shared across both dimensions. The variance parameter σ^2_{nm} will solely be determined by the cluster assignment of the sample. We will observe in Section 4.1, that this

choice of variance parameter is required to obtain the convenient and necessary property of conjugacy. That is, we are required to sacrifice some model flexibility in the variance term for computational efficiency. We can write this succinctly as follows:

$$\mathbf{x}_{nm} \sim \mathcal{N}(\mu_{nm}, \sigma^2_{nm})$$

where

$$\begin{aligned} z_n &= k \\ \mu_{nm} &= \nu_m + \phi_{km} \\ \sigma^2_{nm} &= \sigma^2_{km} \end{aligned}$$

$z_n \in \{1, 2, ..., K\}$ denotes the cluster assignment for sample n, $\phi_{km} \in \mathbb{R}$ denotes the cluster mean for gene m, and $\sigma^2_{km} \in \mathbb{R}^+$ denotes the cluster variance for gene m. The mean parameter μ_{nm} can be decomposed into two main effects: a gene specific effect ν_m and cluster specific effect ϕ_{km}. This observation model is similar to a Bayesian mixture model introduced in Section 2.1 with Gaussian component distributions and added parameters to control for the gene effect. Next, we will control for more effects by simply adding to the mean parameter μ_{nm}.

We introduce a supervised component to the observation model for the tissue of origin and add tissue specific parameters to the mean parameter of the Gaussian distributions. We control for the tissue effect similar in the same way we controlled for the gene effect in the previous model. After controlling for the these two effects we can investigate the cluster configuration. This can be written succinctly as follows:

$$\mathbf{x}_{nm} \sim \mathcal{N}(\mu_{nm}, \sigma^2_{nm})$$

where

$$
\begin{aligned}
z_n &= k \\
t_n &= l \\
\mu_{nm} &= \nu_m + \phi_{km} + \psi_{lm} \\
\sigma^2_{nm} &= \sigma^2_{km}
\end{aligned}
$$

$t_n \in \{1, 2, ..., L\}$ denotes the tissue assignment for sample n and $\psi_{lm} \in \mathbb{R}$ denotes the tissue specific effect of tissue l on gene m. This model is similar to the previous observation model with the addition of tissue parameters ψ_{lm} to the mean parameters μ_{nm}. The mean parameter can now be decomposed into three effects: the gene effect, cluster effect, and tissue effect.

3.3 Prior distributions

We need to specify prior distributions for each parameter used in the observation model. Gaussian distributions will be used as prior distributions on all tissue and gene parameters ν_m, ψ_{lm}. For reasons discussed in Section 4.4, a Gamma distribution will be used as a prior distribution on the hyper-parameter α. The number of clusters in the secondary structure is unknown apriori therefore we encounter a model selection problem. As discussed in Section 2.3, the Dirichlet Process prior is able to side-step this model selection problem as it learns the number of clusters using the data. Therefore, we will use a Dirichlet Process prior on the cluster assignments Z.

3.3.1 The Normal Inverse Chi-Squared distribution

A Normal Inverse Chi-Squared distribution will be used as a prior on the cluster parameters $(\phi_{km}, \sigma^2_{km})$. The Normal Inverse Chi-Squared distribution is a special case of the Normal Inverse-Gamma distribution. This dis-

tribution is conjugate to the the Gaussian distribution with unknown mean and variance μ and σ^2. Section 4.1.1 shows this prior allows us to leverage the convenient property of conjugacy throughout posterior inference. It is referred to as a *compound distribution* as the prior placed on μ is dependant on σ^2 which is also random. More specifically, the prior decomposes as $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$. The distribution takes four hyper-parameters $(\mu_0, \kappa_0, \nu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are the prior mean and variance and κ_0 and ν_0 are levels of confidence for the prior parameters. If

$$\begin{aligned}\sigma^2 &\sim I\chi^2(\nu_0, \sigma_0^2) \\ \mu|\sigma^2 &\sim \mathcal{N}(\mu_0, \frac{\sigma^2}{\kappa_0})\end{aligned}$$

then $(\mu, \sigma^2) \sim \mathcal{N}I\chi^2(\mu_0, \kappa_0, \nu_0, \sigma_0^2)$ with probability density function

$$\begin{aligned}p(\mu, \sigma^2) &= \mathcal{N}I\chi^2(\mu, \sigma^2) \\ &= \mathcal{N}(\mu|\mu_0, \frac{\sigma^2}{\kappa_0}) \times I\chi^2(\sigma^2|\nu_0, \sigma_0^2) \\ &= \frac{1}{Z(\mu_0, \kappa_0, \nu_0, \sigma_0^2)} \frac{1}{\sigma} \frac{1}{(\sigma^2)^{1+\frac{\nu_0}{2}}} e^{-\frac{1}{2\sigma^2}(\kappa_0(\mu-\mu_0)^2+\nu_0\sigma_0^2)}\end{aligned}$$

where $Z(\mu_0, \kappa_0, \nu_0, \sigma_0^2) = \frac{\sqrt{2\pi}}{\kappa_0}\Gamma(\frac{\nu_0}{2})(\frac{2}{\sigma_0^2\nu_0})^{\frac{\nu_0}{2}}$.

3.4 Graphical models

Given the specification of the observation model and prior distributions we can now provide the full hierarchical model. The following generative process is for the gene expression model that controls for the tissue effect. One can obtain the gene expression model that only controls for the gene effect by removing the tissue parameters and tissue assignments. Figure 3.3 illustrates the corresponding probabilistic graphical models.

$$
\begin{aligned}
\alpha &\sim \text{Gamma}(0.01, 100) \\
H &\sim \mathcal{N}I\chi^2(0, 1, 3, 0.1) \\
G|\alpha, H &\sim DP(\alpha, H) \\
(\phi_{km}, \sigma^2_{km}) &\sim G \\
\psi_{lm} &\sim \mathcal{N}(0, 5) \\
\nu_m &\sim \mathcal{N}(0, 5) \\
x_{nm}|z_n = k, t_n = l, (\phi_{km}, \sigma^2_{km}), \nu_m, \psi_{lm} &\sim \mathcal{N}(\nu_m + \psi_{lm} + \phi_{km}, \sigma^2_{km})
\end{aligned}
$$

for $l \in 1, 2, ..., L$, $m \in 1, 2, ..., M$ and $n \in 1, 2, ..., N$.

We place a vague prior distribution on α that is centred at 1 with a variance of 100. The tissue and gene parameters have a Gaussian prior with mean 0 and variance 5. These priors were selected so that it is difficult for the tissue and gene parameters to fully explain an observation. The hyper-parameters for the clusters parameters constrain them to be smaller in magnitude relative to tissue and gene parameters. This is because we expect the cluster effect to be dominated by the tissue effect as described in Section 3.1.

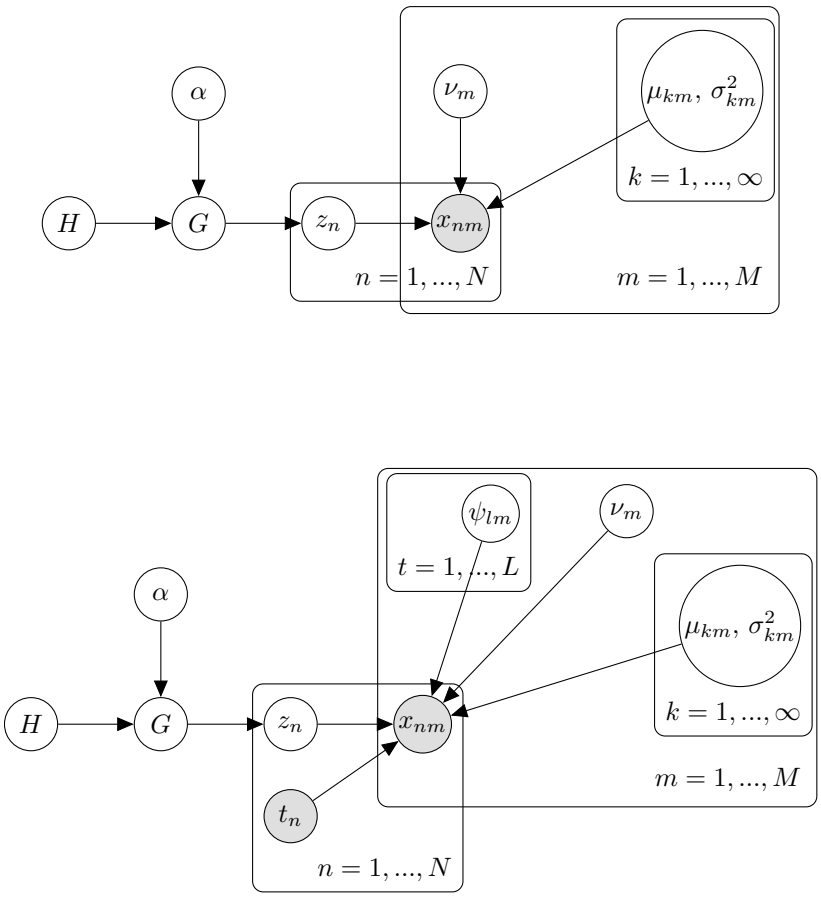


Figure 3.3: Probabilistic graphical models of proposed gene expression model with and without tissue parameters.

3.5 Preprocessing data

RNA-sequencing data is typically provided as discrete count values represented in the gene expression matrix $\boldsymbol{G}$. The entries of $\boldsymbol{G}_{nm}$ denotes the number of "reads" observed in sample n for gene m. In order to model the observations with a Gaussian distribution we need to log-transform the gene expression matrix (Soneson and Delorenzi, 2013). Sometimes RNA-sequencing technologies output highly inflated gene expression counts for specific samples. Therefore, we standardize across genes (or per sample) to control for this potential inflation. Let $\bar{\boldsymbol{x}}_n$ and $\hat{\boldsymbol{\sigma}}_n$ denote the mean and sample standard deviation of the log gene expression for sample n. We perform the log-normalization of the gene expression matrix as follows:

$$\boldsymbol{X}_{nm} = \frac{\log \boldsymbol{G}_{nm} - \bar{\boldsymbol{x}}_n}{\hat{\boldsymbol{\sigma}}_n}.$$

For the remainder of this thesis will use $\boldsymbol{X}$ to denote the log-normalized gene expression matrix and, somewhat abusively, will refer to the log-normalized gene expressions as simply the gene expression matrix.

Chapter 4

Posterior Inference

Posterior inference was performed using Markov chain Monte Carlo (MCMC) techniques for all sampled parameters. The state of our Markov chain included parameters $(\boldsymbol{\nu}, \boldsymbol{\psi}, \boldsymbol{Z}, \alpha)$. The prior distribution for the cluster parameters $(\boldsymbol{\phi}, \boldsymbol{\sigma}^2)$ was conjugate to the likelihood and consequently integrated out. Therefore these parameters are excluded from the Markov chain. Section 4.1, constructs the posterior distribution. Followed by Section 4.2, which discusses Gibbs and Split-Merge updates used for cluster assignments $\boldsymbol{Z}$. Both of these sampling techniques required analytical solutions for the predictive likelihood of a given data point. Section 4.3 outlines the updates used for gene and tissue parameters, $\boldsymbol{\nu}$ and $\boldsymbol{\psi}$. Each parameter uses a similar adaptive Metropolized-Gibbs sampling technique. Finally, the hyper-parameter for the Dirichlet process α was updated using a Gibbs sampling technique described in Section 4.4.

4.1 Constructing the posterior distribution

We construct the posterior distribution for the gene expression model parameters denoted by $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\boldsymbol{X})$. We assume conditional independence between samples and genes so that the likelihood term decomposes into a product of

likelihood terms from each dimension of the data. For completeness, we show the full posterior and its marginalized version that results after integrating out the cluster parameters $(\boldsymbol{\phi}, \boldsymbol{\sigma})$. Since we are collapsing out the cluster parameters we will refer to this latter posterior as the *collapsed* posterior denoted by $p(\boldsymbol{\nu}, \boldsymbol{\psi}, \boldsymbol{Z}, \alpha | \boldsymbol{X})$. Given the cluster assignments, we denote the likelihood of cluster k as

$$L(\mathbf{x}_k | \boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2) = \prod_{n | z_n = k} L(\mathbf{x}_n | \boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2, \boldsymbol{\psi}_{t_n}, \boldsymbol{\nu}) \tag{4.1}$$

and the collapsed likelihood of cluster k as

$$L(\mathbf{x}_k) = \int L(\mathbf{x}_k | \boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2) p(\boldsymbol{\phi}_k, \boldsymbol{\sigma^2}_k) d\boldsymbol{\phi}_k d\boldsymbol{\sigma}_k^2 \tag{4.2}$$

where the prior distribution on each dimension of $(\boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2)$ is a Normal Inverse Chi-Squared distribution. Equation 4.1 will be used in the derivation of the full posterior and Equation 4.2 will be used in the derivation of the collapsed posterior. We derive the full posterior distribution as follows:

$$\begin{aligned} p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{X}) &\propto L(\boldsymbol{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \prod_n L(\mathbf{x}_n | \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \prod_{k=1}^{K} \Big(L(\mathbf{x}_k | \boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2) p(\boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2) \Big) p(\boldsymbol{\nu}) p(\boldsymbol{\psi}) p(\boldsymbol{Z} | \alpha) p(\alpha). \end{aligned}$$

We integrate out the cluster parameters $(\boldsymbol{\phi}, \boldsymbol{\sigma^2})$ from the full posterior to derive the collapsed posterior.

$$\begin{aligned} p(\boldsymbol{\nu}, \boldsymbol{\psi}, \boldsymbol{Z}, \alpha | \boldsymbol{X}) &= \int p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{X}) d\boldsymbol{\phi} d\boldsymbol{\sigma^2} \\ &\propto \prod_{k=1}^{K} \int L(\mathbf{x}_k | \boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2) p(\boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2) d\boldsymbol{\phi}_k d\boldsymbol{\sigma}_k^2 \\ &\quad \times p(\boldsymbol{\nu}) p(\boldsymbol{\psi}) p(\boldsymbol{Z} | \alpha) p(\alpha) \\ &= \prod_{k=1}^{K} L(\mathbf{x}_k) p(\boldsymbol{\nu}) p(\boldsymbol{\psi}) p(\boldsymbol{Z} | \alpha) p(\alpha). \end{aligned}$$

The collapsed posterior is the product of likelihoods from each cluster k and the prior distribution of the remaining model parameters. Section 4.1.1 shows that there is a closed form solution to Equation 4.2.

4.1.1 Conjugate analysis

Conjugacy is a convenient property between the likelihood function and prior distribution that results in the prior and posterior distributions taking the same form. This allows for closed form solutions of otherwise potentially intractable integrals. The desired closed form solution is for the integral of the Gaussian likelihood specified in Section 3.2 against the priors on the cluster parameters. Recall, in Section 2 this setting of Bayesian mixture models was discussed. Once we evaluate Equation 4.2 we effectively integrate out the cluster parameters. Therefore the cluster parameters are excluded from the Markov chain constructed when performing approximate inference using Markov chain Monte Carlo techniques. Consequently, our inference algorithm samples less parameters at each iterations through a process called collapsed Gibbs sampling or Rao-Blackwellization (Das, 2014). We will observe in Section 4.2 that this integral is required for two Markov chain Monte Carlo

sampling procedures.

We will consider the one dimensional case then extend this to the multi-dimensional case for the gene model. Let $L(x_n|\mu, \sigma^2)$ be a Gaussian likelihood with parameters (μ, σ^2) and $p(\mu, \sigma^2)$ be a Normal Inverse Chi-Squared prior distribution. The posterior distribution given a dataset $\mathbf{x}$ is as follows (Murphy, 2007):

$$p(\mu, \sigma^2|\mathbf{x}) \sim \mathcal{N}I\chi^2(\mu_n, \kappa_n, \nu_n, \sigma_n^2)$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_n} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\sigma_n^2 &= \frac{1}{\nu_n}(\nu_0\sigma_0^2 + \sum_{i=1}^{n}(x_i - \bar{x}) + \frac{n\kappa_0}{\kappa_n}(\bar{x} - \mu_0)^2).
\end{aligned}$$

Using conjugacy, we obtain a closed form solution to Equation 4.2 as follows:

$$\begin{aligned}
L(\mathbf{x}) &= \int \left(\prod_n L(x_n|\mu, \sigma^2) \right) p(\mu, \sigma^2) d\mu d\sigma^2 \\
&= \int \left(\prod_n \mathcal{N}(x_n|\mu, \sigma^2) \right) \mathcal{N}I\chi^2(\mu, \sigma^2) d\mu d\sigma^2 \\
&= \frac{Z(\mu_n, \kappa_n, \nu_n, \sigma_n^2)}{Z(\mu_0, \kappa_0, \nu_0, \sigma_0^2)}
\end{aligned}$$

where the normalization constant of the prior Z is given in Section 3.3.

We will now consider the multi-dimensional case encountered in the gene expression model. $\mathbf{x}_n$ is a vector of gene expression values for sample n. We assume independence between samples and genes. Therefore, the integral in Equation 4.2 turns into the product of M one-dimensional integrals. Closed

form solutions to these one-dimensional integrals can be obtained by using conjugacy.

$$\begin{aligned}
L(\mathbf{x}_k) &= \int L(\mathbf{x}_k|\boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2) p(\boldsymbol{\phi}_k, \boldsymbol{\sigma}_k^2) d\boldsymbol{\phi}_k d\boldsymbol{\sigma}_k^2 \\
&= \prod_m \int L(\mathbf{x}_{km}|\phi_{km}, \sigma_{km}^2) p(\phi_{km}, \sigma_{km}^2) d\phi_{km} d\sigma_{km}^2 \\
&= \prod_m \int \Big(\prod_{n|z_n=k} L(x_{nm}|\phi_{km}, \sigma_{km}^2) \Big) p(\phi_{km}, \sigma_{km}^2) d\phi_{km} d\sigma_{km}^2 \\
&= \prod_m \frac{Z(\mu_{km}, \kappa_{km}, \nu_{km}, \sigma_{km}^2)}{Z(\mu_0, \kappa_0, \nu_0, \sigma_0^2)}.
\end{aligned}$$

4.1.2 Clustering on residuals and shifted data

Conjugacy is not affected by our choice to model $\boldsymbol{\mu}$ as a linear combination of the gene, tissue, and cluster effects. However, it is necessary to constrain the covariance matrix $\boldsymbol{\Sigma}$ to be diagonal and equivalent to cluster variance $\boldsymbol{\sigma}^2$. This can be observed in the following,

$$\begin{aligned}
L(\mathbf{x}_{km}|\phi_{km}, \sigma_{km}^2) &= \prod_{n|z_n=k} \frac{1}{\sqrt{2\pi\sigma_{km}^2}} e^{-\frac{1}{2\sigma_{km}^2}(x_{nm}-\mu_{km})^2} \\
&= \prod_{n|z_n=k} \frac{1}{\sqrt{2\pi\sigma_{km}^2}} e^{-\frac{1}{2\sigma_{km}^2}(x_{nm}-(\nu_m+\psi_{tm}+\phi_{km}))^2} \\
&= \prod_{n|z_n=k} \frac{1}{\sqrt{2\pi\sigma_{km}^2}} e^{-\frac{1}{2\sigma_{km}^2}(\tilde{x}_{nm}-\phi_{km})^2}.
\end{aligned}$$

This likelihood term is still Gaussian but with a shifted observation $\tilde{x}_{nm}$. The shifted observation $\tilde{x}_{nm} = x_{nm} - \nu_m - \psi_{tm}$ can be interpreted as the residuals after controlling for tissue and gene effects. Therefore, the model is clustering on observations after controlling for the tissue and gene effects. However, if we decided to model the variance as a function of the cluster

variance and another variance term, such as an observation specific variance, then it would be impossible to decompose the variance term and retain the Gaussian form of the likelihood. In chapter 5, the shifted observations are often used to understand what the model is clustering on.

4.2 Gibbs and Split-Merge updates for cluster assignments

Gibbs and Split-Merge samplers are used to update the cluster assignments $\boldsymbol{Z}$. We will observe both these samplers require an analytical solution to the predictive likelihood of an observation. The Gibbs sampler sequentially updates the cluster assignment for each observation in a random order. While the Split-Merge sampler either splits an existing cluster or merges two existing clusters. Hence, the aptly chosen name *Split-Merge*. This has the potential to update a large number of cluster assignments in a single iteration. In this regard, the Gibbs sampler can be thought of making more refined *local* moves in the sample space while the Split-Merge sampler makes more *global* moves. On its own the Gibbs sampler can exhibit extremely slow mixing behaviour and often gets stuck in local modes. Therefore, we want to use the Split-Merge sampler as a tool to get unstuck from a local mode the Gibbs sampler cannot traverse. This is exemplified in Section 5.1.6. In practice, we alternate between the use of the two samplers by flipping a fair coin where heads selects the Gibbs sampler and tails selects the Split-Merge sampler. We alternate the samplers because it is computationally inefficient and unnecessary in terms of Markov chain Monte Carlo mixing to use both samplers at each iteration.

4.2.1 Gibbs update

We use algorithm 3 of Neal (2000) to perform Gibbs updates of cluster assignments $\boldsymbol{Z}$. This algorithm sequentially updates each cluster assignment by sampling a new assignment according to probabilities proportional to an observations predictive likelihoods. Intuitively, one can think of a predictive likelihood as being a measure of compatibility between an observation and a cluster. This is because the predictive likelihood is the expected likelihood with respect to the posterior distribution of an existing cluster or, in the case of a new cluster, the prior distribution. Note the posterior distribution of an existing cluster is based on the prior distribution of cluster parameters. Therefore, this sampling technique makes intuitive sense as a new cluster assignment will largely be governed by an observations compatibility with each cluster. In the case where an observation is incompatible with the existing clusters, the sampler allows for the possibility of a new cluster. Conjugacy between the Gaussian likelihood and Normal Inverse Chi-Squared prior on cluster parameters allows for an analytical solution to be available for each of the predictive likelihoods. We are able to treat the current observation as new and predict its cluster assignment conditioned on other cluster assignments because the Dirichlet Process is an exchangeable stochastic process. Suppose there are a total of K existing clusters. The probabilities of a new cluster assignment for observation n denoted by z_n is calculated as follows:

$$p(z_n = k) = C\frac{N_{-n,k}}{N-1+\alpha}\int \mathcal{N}(\mathbf{x}_n|\boldsymbol{\phi},\boldsymbol{\sigma}^2)\mathcal{NI}\chi^2_{-n,k}(\boldsymbol{\phi},\boldsymbol{\sigma}^2)d\boldsymbol{\phi}d\boldsymbol{\sigma}^2 \quad (4.3)$$

$$p(z_n = K+1) = C\frac{\alpha}{N-1+\alpha}\int \mathcal{N}(\mathbf{x}_n|\boldsymbol{\phi},\boldsymbol{\sigma}^2)\mathcal{NI}\chi^2(\boldsymbol{\phi},\boldsymbol{\sigma}^2)d\boldsymbol{\phi}d\boldsymbol{\sigma}^2 \quad (4.4)$$

where $k \in \{1,...,K\}$ denotes an existing cluster, $\mathcal{NI}\chi^2_{-n,k}$ denotes the posterior of $(\boldsymbol{\phi}_k, \boldsymbol{\sigma}^2_k)$ given observations assigned to cluster k excluding observation n, $\mathcal{NI}\chi^2$ denotes the prior on $(\boldsymbol{\phi}_k, \boldsymbol{\sigma}^2_k)$ and $N_{-n,k}$ denotes the number of observations assigned to cluster k excluding observation n. Notice, the

probability of cluster assignments are weighted by the cardinality of existing clusters or the Dirichlet Process hyper-parameter α. This comes from the derivation of the predictive distribution of a cluster assignment. Therefore, this sampling technique will exhibit the "rich get richer" behaviour in the Dirichlet Process prior. The procedure for performing Gibbs sampling on the cluster assignments is summarized in Algorithm 1 found below.

Algorithm 1: Gibbs sampler to update cluster assignment of observations n

for *k = 1 to K* **do**
  calculate $p(z_n = k)$ according to (3.1)
end
calculate $p(z_n = K + 1)$ according to (3.2)
sample $z_n \sim \text{Discrete}(p)$

This sampler needs to be performed sequentially on a random permutation of observation indices. Due to the sequential nature of this sampler, a significant proportion of compute time is spent updating cluster assignments via Gibbs sampling.

4.2.2 Sequential Monte Carlo and Particle Gibbs Split-Merge

Before discussing Particle Gibbs Split-Merge (PGSM) (Bouchard-Côté et al., 2017), it is necessary to discuss Sequential Monte Carlo (SMC) (Chopin and Papaspiliopoulos, 2020) and Particle Markov Chain Monte Carlo (PMCMC) (Moral et al., 2006). This is because the PGSM sampler is constructed using both SMC and PMCMC. SMC is a Monte Carlo sampling technique that sequentially targets the posterior distribution of interest using a sequence of intermediate target distributions (Gu et al., 2015). In this setting, the

posterior distribution of interest is often intractable therefore the sampler must leverage simpler intermediate distributions in order to eventually draw samples from the target. It does so by propagating a pre-specified number of particles through a sequence of intermediate distributions. These particles provide an approximation to each intermediate distribution where the final distribution targets the posterior distribution of interest. PMCMC are a set of methods that essentially wraps a Metropolis-Hastings sampler around a SMC sampler and uses it as a proposal distribution. The SMC methodology is present in the PGSM sampler as it uses a sequence of intermediate distribution to sample the posterior of the cluster configuration and resampling steps to remedy the problem of degeneracy. More precisely, the sampler propagates particles that correspond to a split of one cluster or a merge of two unique clusters by evaluating a sequence of intermediate target distributions. The PMCMC methodology is present in the PGSM sampler as the conditional path is used. This conditional path is used to ensure the sampler is targeting the correct posterior distribution.

Figure 4.1 shows an illustration of particle construction in one iteration of the PGSM sampler involving four observations whose indices are $\{3, 4, 5, 6\}$. The algorithm begins by randomly selecting two observations called the anchors. In this example, the anchors are observations 3 and 5. Then, the PGSM sampler only updates observations currently clustered with any of the two anchors. The algorithm begins by initializing the first anchor 3 to a singleton cluster. Followed by allocating the second anchor 5 to another singleton cluster or existing cluster with observation 3. A merge decision is made when the anchors are clustered together and a split decision is made otherwise. If a merge decision is made, all subsequent observations are allocated to the merge cluster. Otherwise, subsequent observations are allocated sequentially to one of the two existing clusters. The allocations are based upon the immediate target distributions at each iteration. The final step provides a new sample of a cluster configuration for the subset of observa-

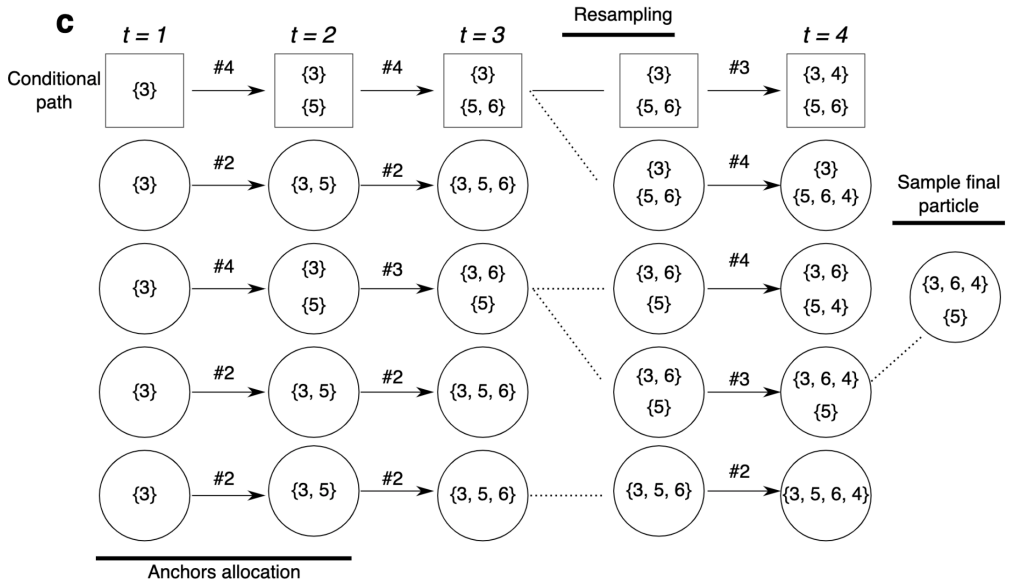


Figure 4.1: Schematic of particle construction for a single iteration of the PGSM sampler on a toy problem (Bouchard-Côté et al., 2017).

tions.

4.3 Metropolized Gibbs updates for gene and tissue parameters

Gene and tissue parameters, $\boldsymbol{\nu}$ and $\boldsymbol{\psi}$, were updated using an adaptive Metropolized-Gibbs sampling technique. That is, we perform a Metropolis-Hastings update within a Gibbs step for each individual parameter ν_m and ψ_{tm}. We assume independence between the dimensions of the gene and tissue parameters therefore the sampling technique essentially reduces to a Metropolis-Hastings update for each parameter. Consequently, a subset of sufficient statistics and likelihood terms are needed to determine the acceptance probability of an update. This leads to a more computationally efficient

sampler. More precisely, instead of re-calculating an $N \times M$ likelihood values when proposing a new parameter value at most M likelihood values are re-calculate. Sufficient statistics are discussed in Section 4.5. The procedure to perform Metropolized-Gibbs sampling for $\boldsymbol{\nu}$ and $\boldsymbol{\psi}$ is summarized by Algorithm 2.

Algorithm 2: Metropolized-Gibbs Sampler for $\gamma_m \in \{\boldsymbol{\nu}, \boldsymbol{\psi}\}$

Let θ_m be the proposal variance for γ_m
$\hat{\gamma}_m \sim \mathcal{N}(\gamma_m, \theta_m)$
Calculate new sufficient statistics using $\hat{\gamma}_m$
$\alpha = \min\left\{1, \frac{p(\hat{\boldsymbol{\mu}},\Sigma|X)}{p(\boldsymbol{\mu},\Sigma|X)} \frac{q(\gamma|\hat{\gamma})}{q(\hat{\gamma}|\gamma)}\right\}$
$u \sim \mathcal{U}[0,1]$
if $u < \alpha$ **then**
  $\gamma_m = \hat{\gamma}_m$
end
Update θ_m according to algorithm 3

Each parameter in $\boldsymbol{\nu}$ and $\boldsymbol{\psi}$ was assigned its own proposal variance parameter θ_m. We set the proposal target acceptance rate to be 0.44 and adapt the log of the proposal variance every 50 MCMC iterations as follows:

$$\log(\theta_m^{new}) = \log(\theta_m) \pm \min\left\{0.25, \frac{1}{\sqrt{N}}\right\}$$

where we increase ($\pm = +$) the variance if the acceptance rate is too low and decrease ($\pm = -$) the variance if the acceptance rate is too high. Initially, this update mechanism changes the proposal variance by a factor of $e^{\pm 0.25}$ encouraging large jumps in the posterior during the burn-in phase. Notice, however, eventually $\frac{1}{\sqrt{N}}$ will be selected as the change in proposal ensuring the proposal adaptation diminishes with the number of MCMC iterations.

Algorithm 3: Update proposal variance θ_m for γ_m

Let N be the current number of MCMC iterations.

Let r_m be the current acceptance rate for γ_m.

if $r_m < 0.44$ **then**

$\log(\theta_m^{new}) = \log(\theta_m) + \min\left\{0.25, \frac{1}{\sqrt{N}}\right\}$

else

$\log(\theta_m^{new}) = \log(\theta_m) - \min\left\{0.25, \frac{1}{\sqrt{N}}\right\}$

end

$\theta_m = e^{\log \theta_m^{new}}$

4.4 Gibbs update for Dirichlet Process hyper-parameter

The concentration parameter α of a Dirichlet Process can be thought of as a tuning parameter for the number of clusters. This is reflected in the equation for the expected number of clusters K given by $\mathbb{E}[K] = \alpha \log N$. Indeed, the expected number of clusters grows linearly with the concentration parameter. This is also reflected in Equation 2.4 describing the Chinese Restaurant Process where the probability of a new table assignment is directly proportional to α. The concentration parameter can strongly effect the resulting posterior inference on the number of clusters (West, 1992). Therefore, instead of treating this hyper-parameter as static, it is necessary to update this parameter via sampling throughout inference. Allowing the number of clusters will be more flexible allowing it to be informed by the data. We will use a Gibbs sampler proposed by West (1992) to update the concentration parameter α. This update mechanism assumes a mixture of gamma distributions as a prior distribution on α, then uses an auxiliary variable trick to obtain the two conditional distributions, $p(\alpha|x)$ and $p(x|\alpha)$, where x is the introduced an auxiliary variable. The assumption of the mixture of Gammas allows for nice form to be obtained for both conditional distributions. Using these

two conditional distribution, we target the posterior of α in a Gibbs fashion given the current number of clusters K. In practice, we simply assume a single gamma prior distribution on α instead of a mixture. We will begin with the joint distribution of α and y and derivate the conditional distribution of each parameter. Following West (1992), we assume $\alpha \sim \text{Gamma}(a, b)$.

$$p(\alpha, y|K) \propto p(\alpha)\alpha^{K-1}(\alpha + n)y^{\alpha}(1-y)^{n-1}. \tag{4.5}$$

To obtain the conditional distribution of α we plug in its prior distribution and treat y as fixed

$$\begin{aligned} p(\alpha|y, K) &\propto \alpha^{a-1}e^{-b\alpha}\alpha^{K-1}(\alpha + n)y^{\alpha} \\ &\propto (\alpha + n)\alpha^{a+K-2}e^{-\alpha(b-\alpha\log(y))} \\ &\propto \alpha^{a+K-1}e^{-\alpha(b-\alpha\log(y))} + n\alpha^{a+K-2}e^{-\alpha(b-\alpha\log(y))} \end{aligned}$$

this can be viewed as a mixture of gamma distributions with the following parameterization:

$$\alpha|y, K \sim \pi_y\text{Gamma}(a_0, b_0) + (1-\pi_y)\text{Gamma}(a_1, b_0)$$

where

$$\begin{aligned} \pi_y &= \frac{(a+K-1)}{n(b-\log(y)) + (a+K-1)} \\ a_0 &= a + K \\ b_0 &= b - \alpha(b - \log(y)) \\ a_1 &= a + K - 1. \end{aligned}$$

Following Equation 4.5, we will now obtain the conditional distribution of y by treating α as fixed

$$p(y|\alpha, K) \propto y^{\alpha}(1-y)^{n-1}. \tag{4.6}$$

Therefore,

$$y|\alpha, K \sim \text{Beta}(\alpha + 1, n). \tag{4.7}$$

We can now using Gibbs sampling because we have the conditional distributions of both α and y. More precisely, we will Gibbs sample by α and y using Equations 4.6 and 5.1 then discard the samples of y and only store the samples of α. Indeed, this is equivalent to marginalizing the auxiliary variable y. The procedure to sample α is summarized in Algorithm 4.

Algorithm 4: Gibbs sampler for concentration parameter α

Let K be the current number of clusters.
Let n be the number of observations.
$y \sim \text{Beta}(\alpha + 1, n)$
Calculate π_y using sampled auxiliary variable
$\beta \sim \text{Uniform}(0, 1)$
if $\beta \leq \pi_y$ **then**
  $\alpha \sim \text{Gamma}(a_0, b_0)$
else
  $\alpha \sim \text{Gamma}(a_1, b_0)$
end

4.5 Sufficient statistics

In this Section we discuss one possible method to reduce computational time by storing sufficient statistics. In particular, we save computation time when

sampling $\boldsymbol{Z}$ with the Gibbs sampler. We saw in Section 4.1, to calculate the collapsed posterior we need to calculate the collapsed likelihood in Equation 4.2. In turn, in order to calculate the collapsed likelihood we need to calculate the posterior parameters for each existing cluster. We can calculate posterior parameters for each clustering using two sufficient statistics. More precisely, we need the sum of data and sum of squared data for each cluster. Returning to the Chinese Restaurant Process analogy in Section 1, this can be thought of as maintaining the posterior parameters for each existing table. One is able to calculate these sufficient statistics by iterating over the observations assigned to each cluster. However, this can be a computationally expensive operation to perform for each MCMC iteration. Instead, we could update sufficient statistics based on new cluster assignments by adding or subtracting an observations or the squared of an observation. Then the posterior parameters can be updated accordingly. We are able to maintain these sufficient statistics using two K by M matrices as there are two sufficient statistics for every cluster and feature. Therefore, updating a sufficient statistic would correspond to updating rows of both the sufficient statistic matrices. We can rewrite the mean and variance of posterior parameter as function of the sufficient statistics.

$$
\begin{aligned}
\phi_{km} &= \frac{\kappa_0\mu_0 + N_k\bar{x}_{km}}{\kappa_{N_k}} \\
&= \frac{\kappa_0\mu_0}{\kappa_0 + N_k} + \frac{N_k}{\kappa_0 + N_k}\sum_{n|z_n=k} x_{nm} \\
\sigma^2_{km} &= \frac{1}{\nu_{N_k}}(\nu_0\sigma_0^2 + \sum_{n|z_n=k}(x_{nm} - \bar{x}_{km})^2 + \frac{N_k\kappa_0}{\kappa_{N_k}}(\bar{x}_{km} - \mu_0)^2) \\
&= \frac{1}{\nu_{N_k}}\nu_0\sigma_0^2 + \frac{1}{\nu_{N_k}}\Big(\sum_{n|z_n=k} x_{nm}^2\Big) - \frac{1}{\nu_{N_k}}\frac{1}{N_k}\Big(\sum_{n|z_n=k} x_{nm}\Big)^2 \\
&+ \frac{\kappa_0}{\nu_{N_k}\kappa_{N_k}N_k}\Big(\sum_{n|z_n=k} x_{nm}\Big)^2 - \frac{\kappa_0}{\nu_{N_k}\kappa_{N_k}}2\mu_0\Big(\sum_{n|z_n=k} x_{nm}\Big) - \frac{N_k\kappa_0}{\nu_{N_k}\kappa_{N_k}}\mu_0^2.
\end{aligned}
$$

Hence, if we maintain the sum of data and sum of squared data, $\sum_{n|z_n=k} x_{nm}$ and $\sum_{n|z_n=k} x_{nm}^2$, one can efficiently re-calculate the posterior parameters needed to calculated the collapsed posterior. In our model these data points are the shifted observations $\tilde{x}_{nm}$ introduced in Section 4.1.2. This is because we are clustering on the residual signal in the data after controlling for the gene and tissue parameters. Consequently, as we update the gene and tissue parameters $\boldsymbol{\nu}$ and $\boldsymbol{\psi}$ we also need to update the shifted observations and sufficient statistics.

4.6 Posterior summary

Markov chain Monte Carlo sampling techniques return samples from an intractable distribution of interest called the *target distribution*. The list of samples from the sampling algorithms is often referred to as the *chain of samples*. Using these samples we are able to obtain an approximation to the target distribution using an empirical distribution and calculate values of interest. In our case, the target distribution will be collapsed posterior

distribution specified in Section 3.2 which we can only evaluate up to a normalization constant. In other words, leveraging Markov chain Monte Carlo techniques we go from dealing with an intractable posterior distribution to a tractable approximation to the posterior distribution. As we increase the number of samples we can approximate the target distribution arbitrarily well. Recall, the collapsed posterior distribution is parameterized by four main parameters: the gene, tissue, and alpha parameters and cluster assignments. Therefore, our Markov chain Monte Carlo sampling techniques return chain of samples for each of these parameters. To perform posterior inference we need to obtain point estimates of each of the model parameters using the chain of Markov chain Monte Carlo samples. In Section 4.6.1, we discuss how to obtain a point estimate of a cluster matrix given a sample of cluster matrices. Followed by Section 4.6.2, which outlines the post processing of the gene and tissue parameters. Finally, in Section 4.6.3 we show how to re-instantiate the cluster parameters in order to obtain a point estimate after integrating these parameters out of the Markov chain.

4.6.1 Model consensus

Our target distribution is the collapsed posterior distribution of our model which includes a cluster matrix Z. Recall, each row of this cluster matrix is a one-hot encoding of an observation indicating its cluster assignment. Recall, the inference algorithms outlined in Section 4.2, update the cluster assignments in a Gibbs or Split-Merge fashion. After each iteration of these samplers one should think of the output as a sample of the posterior distribution of Z. Therefore, we obtain a list of Markov chain Monte Carlo samples of the cluster matrix Z. It is non-trivial to obtain a point estimate of a cluster matrix given a chain of cluster matrices obtained from Markov chain Monte Carlo inference algorithms. One method to obtain a point estimate is to construct a distance metric which can be naturally extended to a

dendrogram and then maximize a pre-specified criterion. Let

$$S_{ij}^k = \mathbb{1}\{\text{observation } i \text{ and } j \text{ are clustered together}\}$$

where k indicates the Markov chain Monte Carlo iteration and i and j indicate observation indices. In other words, S^k is an $N \times N$ matrix indicating whether two observations were clustered together in the k^{th} Markov Chain Monte Carlo iteration. This matrix is often referred to as a similarity matrix. We can construct a notion of distance between any two observation i and j by defining

$$d(i,j) = 1 - \frac{1}{N_{iter}} \sum_{k=1}^{N_{iter}} S_{ij}^k.$$

Given that S is a similarity matrix, the component-wise addition of two similarity matrices has the interpretation of the number of times the two observations have been clustered together in two Markov chain Monte Carlo iterations. Hence, the interpretation of distance measure $d(i,j)$ would be the proportion of times the two samples i, j were not clustered throughout the Markov chain. Indeed, this has the interpretation of a distance because as two samples are clustered less together then distance metric $d(i,j)$ increases and if two observations are frequently clustered together the distance metric decreases. Once a distance metric is obtained between any two observations we can extend this to perform hierarchical clustering between observations. Then, we can obtain a point estimate of cluster assignments using the resulting dendrogram to maximize a criteria called Maximum Posterior Expected Adjusted Rand or MPEAR (Fritsch and Ickstadt, 2009).

4.6.2 Tissue, gene, and α parameters

We include the tissue, gene, and alpha parameters in one Section as the same posterior inference method will be used. To obtain point estimates for

each of these parameters we simple take the sample average from the Markov chain Monte Carlo chain. More precisely, the point estimates for the follow parameters are calculated as follows:

$$\hat{\psi}_{tm} = \frac{1}{N_{iter}} \sum_{k=1}^{N_{iter}} \psi_{tm}^{k}$$
$$\hat{\nu}_{m} = \frac{1}{N_{iter}} \sum_{k=1}^{N_{iter}} \nu_{m}^{k}$$
$$\hat{\alpha} = \frac{1}{N_{iter}} \sum_{k=1}^{N_{iter}} \alpha^{k}.$$

4.6.3 Cluster parameters

To get a posterior point estimate of the cluster parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$, the maximum a posteriori was used from the conditional distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma^2}$ conditioned on the sample average of $\boldsymbol{\nu}$ and $\boldsymbol{\psi}$ and the MPEAR clustering of $\boldsymbol{Z}$. This was efficiently obtained using conjugate analysis of Gaussian distributions.

$$\begin{aligned} p(\mu_{km}, \sigma_{km}^2 | Z, \nu, \psi, X) &= p(\mu_{km}, \sigma_{km}^2 | Z = k, \nu_m, \psi_{\dot{m}}, X_{\dot{m}}) \\ &= p(X_{\dot{m}} | Z = k, \nu_m, \psi_{\dot{m}}, \mu_{km}, \sigma_{km}^2) p(\mu_{km}, \sigma_{km}^2) \\ &= p(\tilde{X}_{\dot{m}} | \mu_{km}, \sigma_{km}^2) p(\mu_{km}, \sigma_{km}^2) \\ &= \prod_{z_i = k} \mathcal{N}(\tilde{x}_{im} | \mu_{km}, \sigma_{km}^2) \mathcal{N}I\chi^2(\mu_{km}, \sigma_{km}^2). \end{aligned}$$

We have seen in Section 4.1.1 that this is also a Normal Inverse Chi-Squared distribution with posterior parameters. Therefore, we will use the Maximum A Priori estimate which correspond to the mean of the mean parameter and mean of the variance parameter as follows:

$$\hat{\mu}_{km} = \mu_{km}$$
$$\hat{\sigma^2}_{km} = \frac{\nu_0}{\nu_0 - 2}\sigma^2.$$

Chapter 5

Applications

This chapter applies the Gene Expression model and approximate inference techniques using two synthetic datasets and one real-world dataset. In Section 5.1, we test model performance and inference implementation by simulating multiple datasets from the generative process described in Section 3.2. Section 5.2, discusses experiments using semi-synthetic data informed by an open source dataset obtained from the Genome Tissue Expression portal (GTEx). For both synthetic datasets, we compare results from approximate inference to the known true parameters used in the synthetic data generation process. We conclude by testing our model on another open source dataset obtained from the International Cancer Genome Consortium (ICGC) in Section 5.3.

5.1 Fully-synthetic experiment

In order to test model performance and inference algorithms we conduct experiments using synthetic data. This synthetic data is forward simulated from the generative process described in Section 3.4. We set the dimensions of the synthetic data experiments to $N = 100$ samples and $M = 200$ genes. These dimensions were selected to follow the structure of bulk RNA-seq data,

specifically that each observation has many features or gene reads. We can increase the number of samples N to obtain more secondary clusters generated from the Dirichlet Process to further examine cluster performance. We can also increase the number of genes M to test the efficiency of our inference procedures.

An instance of a generated dataset and its parameters used for forward generation are given in Figures 5.1 and 5.3. Specifically, Figure 5.1 shows the forward simulated gene expression matrix $\boldsymbol{X}$ where x_{nm} denotes the gene expression for sample n gene m. The colour bar on the right most side of the gene expression matrix indicates the magnitude of a gene expression where a lighter colour implies a more positive gene expression and a darker colour implies a more negative gene expression. The Subfigures in Figure 5.3 are the forward generated tissue, gene, and cluster parameters. Therefore, the mean parameter shown in Figure 5.2a, should be understood as the element-wise sum of the Subfigures a, c, and b. These Subfigures denote the effect of each component of the model on a specific sample and gene. Subfigure a represents the tissue effect on each element of the gene expression matrix. For example, the tissue effect up-regulates sample 0 gene 0 significantly and down-regulates sample 99 gene 0 significantly. In Subfigure b each sample is effected in the same way by the gene effect matrix because each row is equivalent. Specifically, gene 3 is highly up-regulated by the gene effect. In Subfigure c, we observe cluster specific effects on the gene expression matrix. For example, the cluster assignment for observation 20 up-regulates gene 101 and down-regulates gene 123.

5.1.1 Forward generated data

For each of the simulated datasets we fix the data seed and sample parameters according to the following distributions:

$$\begin{aligned}
\alpha &\sim \text{Gamma}(1.5, 1) \\
Z|\alpha &\sim \text{DP}(\alpha) \\
(\phi_{km}, \sigma^2_{km}) &\sim \mathcal{N}I\chi^2(0, 10, 10, 2) \\
\nu_m &\sim \mathcal{N}(0, 1) \\
\psi_{0m} &\sim \mathcal{N}(2, 0.5) \\
\psi_{1m} &\sim \mathcal{N}(-2, 0.5)
\end{aligned}$$

where m denotes the gene, k denotes the cluster, and ψ has two pseudo tissues 0 and 1. We will further discuss the tissue structure used in the subsequent paragraph. We want the tissue effect to mask the secondary cluster effect and the magnitude of the tissue parameters to be larger than the magnitude of the cluster parameters. The Dirichlet process hyper-parameter is centred around 1.5 with a small variance to encourage an interesting clustering given $N = 100$ observations.

Given that the tissue assignments are known a priori, we need to impose a specific structure on them throughout the forward simulated experiments. The samples are divided up into two pseudo tissues: the first half of samples are assigned to tissue 0 and the second half of samples are assigned to tissue 1. This tissue structure is represented as a matrix in Equation 5.1 where rows indicate samples and columns indicate tissue assignment. For completeness, the other parameters used in forward simulation are shown along side the tissue assignments with the cluster assignment matrix show in Equation 5.2 having the same interpretation. To ensure that the tissue effect is the dominant signal and that there is a large enough variance between the primary and secondary clusters and the two pseudo tissues, we need to

impose a further structure on the tissue parameters. We impose a structure on the tissue parameters such that the tissue effects from both tissues are significantly different and they are larger in magnitude relative to the latent cluster effect.

$$\psi_{0m} \sim \mathcal{N}(2, 0.5)$$
$$\psi_{1m} \sim \mathcal{N}(-2, 0.5)$$

where tissue 0 tends to up-regulate all gene expressions and tissue 1 tends to down-regulate all gene expressions. This brings us to the mathematical formulation of the matrices given in Figure 5.3.

$$\boldsymbol{T} \cdot \boldsymbol{\psi} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \psi_{00} & \psi_{01} & \dots & \psi_{0M-1} \\ \psi_{10} & \psi_{11} & \dots & \psi_{1M-1} \end{bmatrix} \tag{5.1}$$

$$\boldsymbol{Z} \cdot \boldsymbol{\phi} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \phi_{00} & \phi_{01} & \dots & \phi_{0M-1} \\ \phi_{10} & \phi_{11} & \dots & \phi_{1M-1} \end{bmatrix} \tag{5.2}$$

$$\boldsymbol{\nu} = \begin{bmatrix} \nu_0 & \nu_1 & \dots & \phi_{M-1} \end{bmatrix} \tag{5.3}$$

$$\boldsymbol{\mu} = \mathbb{1}_{N\times 1} \cdot \boldsymbol{\nu} + \boldsymbol{T} \cdot \boldsymbol{\psi} + \boldsymbol{Z} \cdot \boldsymbol{\phi}. \tag{5.4}$$

Tissue 0 is regulating gene m by a value of ψ_{0m} and tissue 1 is regulating gene m by ψ_{1m}. Therefore, the dot product $\boldsymbol{T} \cdot \boldsymbol{\psi}$ has the interpretation of being the overall tissue effect on each observation. $\boldsymbol{Z} \cdot \boldsymbol{\phi}$ and $\mathbb{1}_{N\times 1} \cdot \boldsymbol{\nu}$ have the

same interpretation. As alluded to before Equation 5.4 is the element-wise summation.

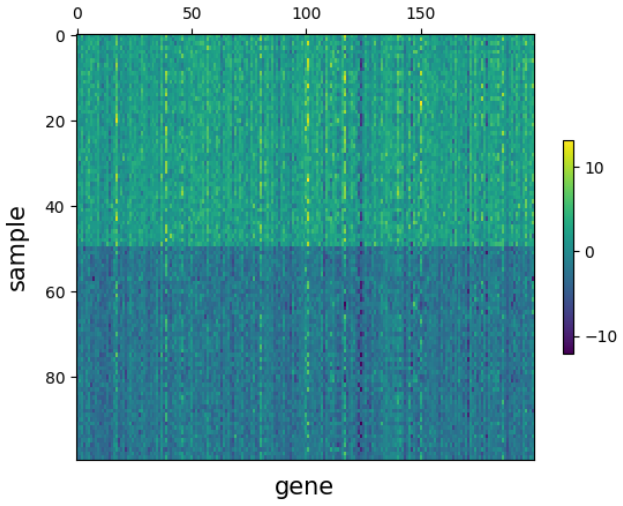
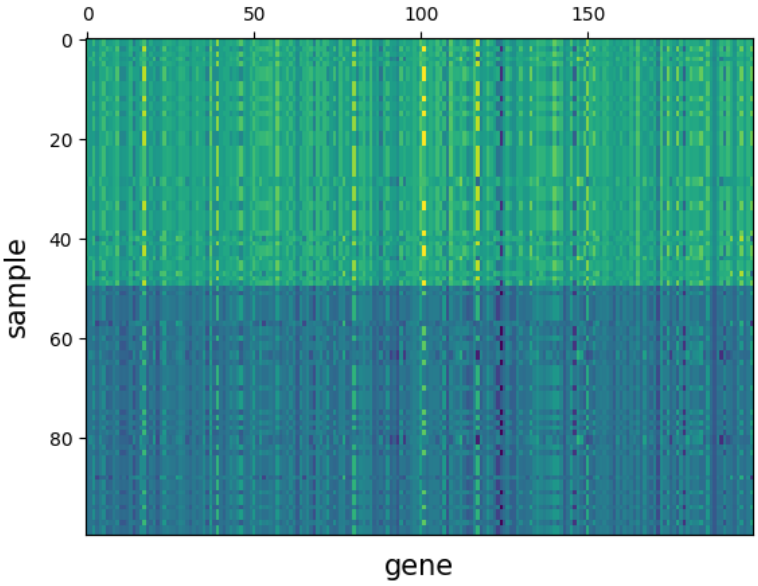


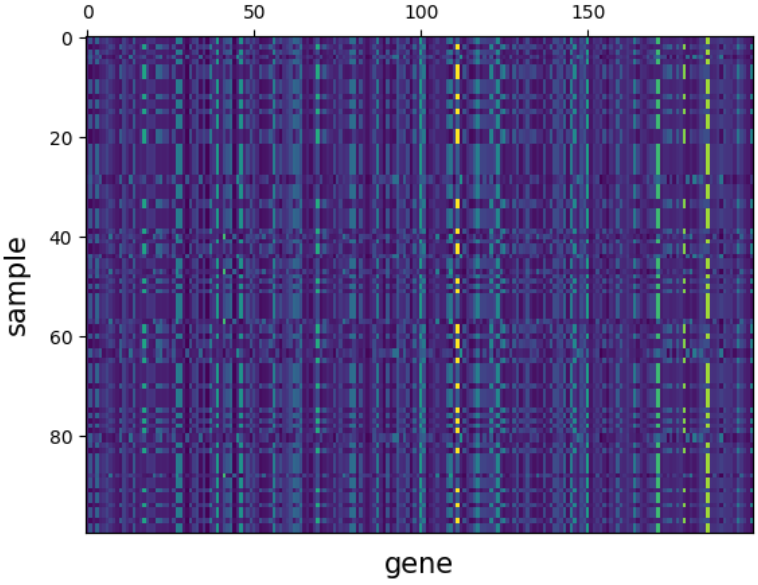
Figure 5.1: Forward generated data from gene expression model used in synthetic data experiments.

5.1.2 Simulation setup

Four datasets were generated using four different data seeds and four Markov chain Monte Carlo (MCMC) samplers were restarted using four different sampler seeds. Therefore, in total sixteen experiments were conducted using synthetic data. The parameters were initialized at $\boldsymbol{\nu} = 0$, $\boldsymbol{\psi} = 0$, $\alpha = 1$, and $\boldsymbol{Z}$ set to the singleton clusters. That is, each observation was initialized in their own cluster. In practice, the singleton clusters are often used as an initialization as this configuration is less likely to get stuck in a local mode. Throughout the following presentation we will focus on one specific synthetic experiment. Namely, the experiment conducted when setting data seed to 5 and sampler seed to 0. First, we will show model results when updating the



(a) Mean parameters



(b) Variance parameters

Figure 5.2: Forward generated mean and variance parameters for gene expression model used in synthetic data experiments.

tissue parameters $\boldsymbol{\psi}$, then we will show model results when $\boldsymbol{\psi}$ is not being updated. Since the tissue parameters are initialized at $\boldsymbol{\psi} = 0$, we will be modelling the gene expression data with the model introduced in Section 3.2 not controlling for the tissue effect. In this scenario, we will say the tissue parameters are turned *off*. Each Markov chain was run for 200 seconds of burnin and 1000 seconds of inference.

5.1.3 Posterior inference with tissue parameters

We will first show the posterior inference on the mean and variance parameters of the model $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Followed by, posterior inference on the constituent parameters $\boldsymbol{\psi}, \boldsymbol{\nu},$ and $\boldsymbol{\phi}$ that make up the mean parameter. Recall, the variance parameter of the model is solely comprised of a cluster specific variance parameter. Figure 5.5 shows the posterior inference on the mean and variance $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ where Subfigure a shows mean and Subfigure b shows the variance. We observe from both these subfigures that inference on the mean and variance parameters capture the structure of the forward generated data. More precisely, when calculating the component-wise difference between the forward generated parameters and inferred parameters we achieve a mean squared error of 0.25 and 0.50 for the mean and variance, respectively.

Figure 5.5 shows the posterior inference on the tissue, gene, and cluster parameters. More specifically, Subfigure a shows the inferred tissue parameters, Subfigure b shows the inferred gene parameters, and Subfigures c and d show the inferred mean and variance cluster parameters, respectively. Recall, the synthetic data for this experiment was generated such that tissue 0 tends to up-regulates gene expressions and tissue 1 tends to down-regulates gene expressions. We observe in Subfigure a that the inferred tissue parameters do capture this pattern. That is, the observations assigned to tissue 0, the first half of observations, have tissue parameters that tend to up-regulate gene expressions and the observations assigned to tissue 1, the latter half of observations, have tissue parameters that tend to down-regulate gene expres-

sions. Hence, the sharp contrast in parameter values at sample 50, the half way point, in Subfigure a. Furthermore, we observe in Figure 5.6 that the average of the tissue parameters in each tissue are centred around the mean value that was used to generate them. The average of tissue 0 parameters are centred around 2 and the average of sampled tissue 1 parameters are centred around -2. This is also observed for the gene parameters in Figure 5.7 as the average of the sampled parameters are centred around 0. Recall, this experiment focused on the tissue and cluster parameters therefore the forward generated gene parameters were centred around 0.

We will now discuss posterior inference on cluster assignments for the synthetic data experiments. Figure 5.8 shows the posterior clustering where Subfigure a shows the V-measure as a function of MCMC iteration and Subfigure b shows the similarity matrix for observations. Recall, each element of the similarity matrix can be interpreted as the proportion of MCMC iterations the corresponding observations were clustered together. One measurement to calculate the distance between two cluster configurations is called the V-measure. The V-measure is a scalar value between 0 and 1 that measures the similarity between two configurations where 1 implies equivalence and 0 implies a large difference. In other words, the closer the V-measure tends towards 1 the more similar two cluster configurations are. We will use the forward generated cluster configuration and current inferred cluster configuration to calculate the V-measure for each MCMC iteration. This is show in Subfigure a. We observe that posterior inference does capture the correct cluster configuration simulated by the Dirichlet Process prior. More precisely, after a burn-in period of 200 seconds the correct clustering is obtained for the full duration of MCMC inference. In Subfigure b, the left most side of the similarity matrix has tissue labels indicated by colour for each observation and the corresponding legend above. For example, observation 99, the first row of the similarity matrix, is assigned to tissue 1. From this, we observe that there are clusters forming with observations from each tissue

type such as cluster 0, the largest cluster, as it is comprised of observations from tissue 0 and tissue 1. Hence, we are not clustering by the primary cluster but a latent secondary clustering. Later we will show the more efficient Split-Merge sampler is required to obtain the correct cluster configuration.

We return to the forward generated gene expression data and show the inferred posterior clustering of observations. Figure 5.9 shows the inferred cluster configuration on plots of the simulated gene expressions data and Figure 5.10 shows its shifted counterpart. Both plots also contain tissue and cluster labels indicated by colour on the left most side of the data matrices. For example, in Figure 5.9, the first row shows that observation 63 is assigned to tissue 1 and cluster 2. We define the shifted gene expressions as the gene expressions after subtracting the inferred tissue and gene parameters. Intuitively, the shifted gene expression should reflect the effect of the latent secondary clustering we aim to cluster on. More precisely, the shifted data is defined as $y_{nm} = x_{nm} - \psi_{t_n m} - \nu_m$ for all n, m. In Figure 5.10, we observe the block structure in the shifted gene expression matrix that corresponds to the cluster effects where the blocking occurs at observations $42, 63$ and 47. This block structure correspond to the colouring of cluster labels as we are clustering on the gene expressions after controlling for the tissue and gene effect. Hence, the purpose of juxtaposing these two plots is to show that after controlling for the tissue and gene effects we cluster on the remaining signal in the data. Figure 5.10 provides an insight into what signal the model is using to cluster observations. This is also the reason why it is vital for the model the capture the tissue and gene effect. Otherwise, the model will cluster observations based on a combination of the remaining tissue and gene effect, whereas the objective is to cluster observations based on the secondary cluster effect. Using the tissue and cluster labels in Figure 5.10, we observe that there is mixing between tissue types within inferred clusters. For example, observations from tissue 0 and tissue 1 are present in multiple inferred clusters such as cluster 0. This reenforces that we have unmasked

the hidden or latent structure in the data after controlling for the other covariates.

5.1.4 Identifiability problems

There seems to be some identifiability problems for the gene, tissue, and cluster parameters. There are multiple combination of values of gene, tissue, and cluster parameters that produce the same value for the mean parameter. For example, suppose we are targeting a mean gene expression value of μ_{nm} for sample n and gene m. There are many values of $\psi_{tm}, \nu_m, \phi_{km}$ that satisfy $\mu_{nm} = \psi_{tm} + \nu_m + \phi_{km}$ where $T_n = t$ and $Z_n = k$. The model cannot differentiate between the correct values of $\psi_{tm}, \nu_m, \phi_{km}$ that originally generated the data as it only takes into account the sum of these parameters. More concretely, we observe the forward generated gene and tissue parameters both express gene 150 moderately in Figure 5.3. However, we observe the inferred tissue and gene parameters both express gene 150 with large magnitudes. The tissue parameters have a very large positive magnitude indicating a high gene expression and the inferred gene parameters has a very low negative magnitude indicating a low gene expressed. This can be seen in the column 150 of both matrices in Subfigures a and b. In other words, the drastic up-regulation of gene 150 resulting from the tissue parameters is offset by the drastic down-regulation from the gene parameters, ultimately allowing for a decent inference on the mean parameter μ_{nm}.

5.1.5 Posterior inference without tissue parameters

We will now use the same synthetic data to perform inference while setting the tissue parameters $\psi = 0$. We do so by initializing the tissue parameters to $\psi = 0$ and performing inference without the Metropolized-Gibbs steps to updated the tissue parameters. This effectively does not allow the model to learn the tissue effect in the data. Therefore, this reduces the model to the

first model introduced in Section 3.2 as it only controls for the gene effect. In this experiment, we will focus on plots that show different aspects of the inferred cluster configurations after keep the tissue parameters static. This is because we aim to show that if the primary clustering is not controlled, then the model will naturally cluster on it. That is, we will observe a direct correspondence between the tissue and cluster labels of the observations. Indeed, we observe this in Figure 5.15. Furthermore, we observe the residual signal in the data is is representative of the tissue effect in Figure 5.13. Hence, the model is clustering on the tissue effect in the data. In other words, the model now leaves the secondary clustering of interest masked by the primary clustering.

5.1.6 Posterior inference without Split-Merge sampler

In the following experiment we return to sampling the tissue parameters ψ and therefore use a model that controls for the tissue effect. We will now show the necessity of the Split-Merge sampler introduced in Section 4.2.2 using the same synthetic data as the previous experiment. We will do so by showing that the inferred clustering when using only the Gibbs sampler outlined in Section 4.2.1 results in the incorrect inferred cluster configuration. Gibbs samplers for cluster assignments often gets stuck in local modes as it is difficult to split or merge clusters when updating a single cluster assignment per iteration. For example, to perform a split of a cluster using only a Gibbs sampler a single observation would have to create their own cluster, then the other compatible observations would have to join it sequentially. This sequence of events has a low probability and therefore would require many iterations before arriving to the desired state. That is, the Gibbs sampler is not efficient enough on its own to sample the posterior of cluster assignments. Hence, we need a more sophisticated sampling techniques such as the Split-Merge sampler. This is observed in Subfigure a as the cluster configuration is stuck in a local mode for a number of iterations before it jumps to a better

configuration. Notice the sampler never reaches the correct clustering as the V-measure is not equal to 1.

5.2 Semi-synthetic experiment: genotype tissue expression data

5.2.1 Semi-synthetic data and posterior inference

The Genotype-Tissue Expression (GTEx) project is an ongoing open source project with the goal to help better understand tissue-specific gene expression and regulation (Lonsdale et al., 2013). One of the ways the projects achieves this goal is by providing the public with tissue-specific gene expression data. In a nutshell, a biopsy is performed on healthy tissue and bulk RNA-sequencing technology is used to obtain gene expression counts. The tissue-specific gene expression data used was obtained in a similar manner to the methods described in Section 1.

The goal is to generate a non-forward generate dataset in which to test model inference. This is because our model will be naturally biased towards forward simulated data and model results will be less representative of real data cases. We generate a dataset using the tissue gene expressions obtain from the GTEx portal. We select four tissues in which we designate two tissues to be the dominant effect and two tissue to be the latent effect similar to the scenario outline in 3.1. That is, the two dominant tissues will mask the latent effect from the two latent tissues. Therefore, we would obtain a generated dataset informed by real data which we know the dominant and latent cluster assignments but not model parameters. Our goal is to cluster by the two latent tissue effects. We select brain and ovary to be the dominant tissue effect and lung and thyroid to be the latent effect.

We outline the process in which we generate semi-synthetic data. Using the gene expression data obtain from the GTEx portal, we average the gene

expressions across each sample per tissue to create four gene expression vectors. For each tissue we take 50 samples and 500 genes. Then we take convex combinations of the four vectors to new create four alternative vectors. More precisely, we take a proportion p of each dominant vector add it to $1-p$ of each latent vector. Each vector will be used as a concentration parameter of length 500. These concentration parameters will be passed through a Dirichlet distribution to sample a simplex corresponding to each parameter. This proportion p is left an a parameter. We then use this simplex to sample a count vector using a multinomial distribution. We continually sample from a multinomial distribution until we get a desired number of pseudo gene expression values. Let $\chi_{\text{Brain}}, \chi_{\text{Colorectal}}, \chi_{\text{Lung}}$ and χ_{Breast} denote the gene expression vectors created using the GTEx data. The concentration parameters and sampling method are shown below:

$$
\begin{aligned}
\alpha_1 &= p\chi_{\text{Brain}} + (1-p)\chi_{\text{Colorectal}} \\
\alpha_2 &= p\chi_{\text{Brain}} + (1-p)\chi_{\text{Lung}} \\
\alpha_3 &= p\chi_{\text{Breast}} + (1-p)\chi_{\text{Colorectal}} \\
\alpha_4 &= p\chi_{\text{Breast}} + (1-p)\chi_{\text{Lung}} \\
p_1 &\sim \text{Dirichlet}(\alpha_1) \\
p_2 &\sim \text{Dirichlet}(\alpha_2) \\
p_3 &\sim \text{Dirichlet}(\alpha_3) \\
p_4 &\sim \text{Dirichlet}(\alpha_4) \\
X_i &\sim \text{Multinomial}(10^5, \alpha_{Z_i}).
\end{aligned}
$$

This was used to test the the models ability to captures varying levels of signal between the dominant and latent effects. We were able to recover a maximum of 0.8 proportion indicating the model has success to recover a latent when the dominant effect is large.

In Figure 5.17a we show an instance of the synthetic data generated using

the tissue gene expression data obtained from the GTEx portal. The way the semi-synthetic data was generated there is an obvious primary structure where the first half of observations are clustered together and the latter half of observations are clustered together. This primary structure is similar to the fully synthetic data generated in the previous experiments. We observe that there is a finer signal in the semi-synthetic data that is indicative of the secondary clustering. More precisely, the observations with even indices share a common structure and the observations with even indices share a common structure. The goal is to cluster observations based on this secondary structure. Therefore, the following plots will primarily focus on the posterior clustering of the observation.

5.2.2 Posterior inference without tissue parameters

We test the gene expression model without controlling for the tissue effect with the generated semi-synthetic data. Figure 5.22, shows the inferred mean parameters of the gene expression model. We observe the inferred gene expressions on the latter half of observations have the same value when the semi-synthetic data shows different values for even and odd observation indices. Subfigure b, shows the similarity matrix of observations on the semi-synthetic data without sampling the tissue parameters. The similarity matrix shows the model detected a difference between first half of samples since they are clustered into two sub-clusters corresponding to even and odd indices. The samples from the other tissue were clustered together as their tissue effect is dominant in the data. Therefore, we have that the sample still clusters by tissue with some smaller within-tissue clusters. Subfigure a shows the V-measure after a burnin phase never reaches a value of 1 which implies the model never obtains the true cluster configuration.

5.2.3 Posterior inference without Split-Merge sampler

We perform an experiment using semi-synthetic data in which we only use the Gibbs sampler to sample cluster assignments. Figure 5.25, shows the inferred clustering on the semi-synthetic data. We observe the Gibbs sampler is unable to sample the posterior distribution efficiently therefore the Split-Merge sampler is required to perform inference. More specifically, Subfigure b shows that the Gibbs sampler is able to split the primary clustering according to the finer signal corresponding to the secondary clustering, however the sampler is unable to merge the observations accordingly. Recall, the data was generated such that that the only true parameter known are the cluster assignments. Therefore we are able to calculate the V-measure between the current inferred cluster configuration and the true cluster configuration. Subfigure a shows the V-measure never reaches a value of 1 which implies the model never obtains the true cluster configuration used to generate the semi-synthetic data.

5.3 Real-world application: international cancer genome consortium data

5.3.1 ICGC data

We test our model on a real data set obtained from the International Cancer Genome Consortium (ICGC) (Zhang et al., 2019). We selected to model Bulk RNA-sequencing data from a cohort of cancer patients. Specifically, we focus on a cohort with the following cancer types: Breast, Brain, Liver, Colorectal, and Lung. Each tissue type is taken from one project so that potential batch effects are accounted for within the tissue parameters. For example, breast cancer data is from a project named BRCA-US and Lung cancer data is from a project named LUAD-US on the ICGC portal. We run our experiment on a dataset with $N = 125$ samples and $M = 766$ genes where each cancer

type has 25 samples. We used a publicly available list of genes obtained from NanoString Technologies (NanoString Technologies, 2015). This list of genes is created based on their relationship with gene pathways, a potential secondary clustering with hope to cluster on after controlling for the tissue effect. Figure 5.26 shows the ICGC dataset we perform inference on. We observe a staggered block structure every 25 rows. This is because the data is organized to stack each of the 25 samples from each cancer type. For example, the first two block correspond to breast and ovarian cancer where rows 0 to 24 correspond to the former and rows 25 to 49 correspond to latter.

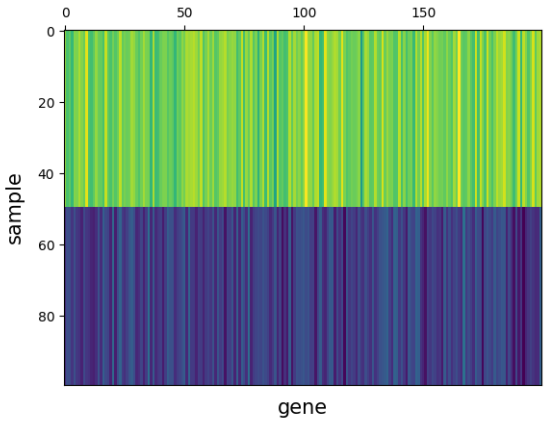
5.3.2 Posterior inference without tissue parameters

We first show the experiment performed using the model that does not control for the tissue effect in the data. The purpose of this is to ensure that the model does capture the tissue effect and therefore cluster on it. In other words, it is an ad-hoc sanity check to ensure the model captures the correct clustering. We initialize the tissue parameters $\psi = 0$ and do not update them using the Metropolized-Gibbs sampler. In Figure 5.27, we observe for the inferred mean parameter captures the general structure of the data. Moreover, in Subfigure a we observe that the cluster parameters do capture the tissue effect from each cancer type. For example, gene 5 seems to be heavily down-regulated by breast cancer and gene 95 is heavily up-regulated by colorectal cancer. Subfigure 5.28b, shows the inferred gene expression parameters. We observe that the trend of gene parameters values are tending from positive to negative. This was expected as the genes are ordered in descending order measured by the mean absolute deviation. In Figure 5.29, we observe that the clustering is by tissue type with some mixing between Colorectal and Liver cancer forming a small cluster. This was expected as the tissue effect is not controlled for in this model. The mixing between Colorectal and Liver cancer may be because both these cancers affect a particular gene pathway. Clustering by tissue type is observed again in Figure 5.30 where there is

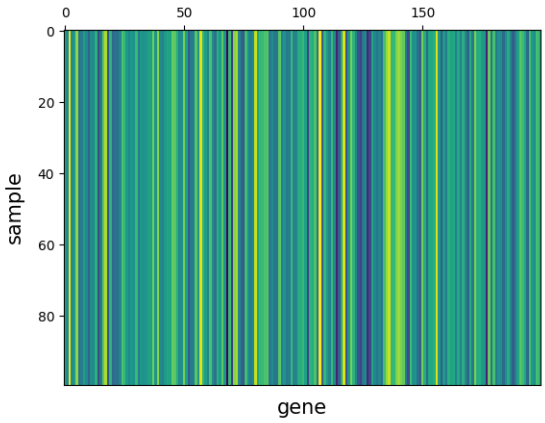
a direct correspondence between the tissue label and cluster label with the exception of the additional inferred cluster.

5.3.3 Posterior inference with tissue parameters

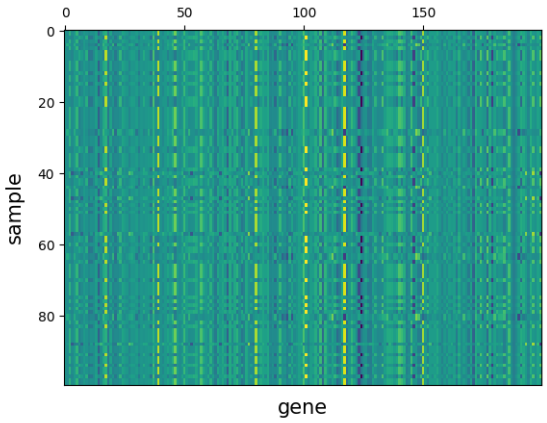
We now sample the tissue parameters ψ to control for the tissue effect in the data. This is to allow the model to cluster on a potential secondary structure. More specifically, we want the tissue parameters to capture the tissue effect that the previous model clustered on. Therefore, we conduct the same experiment as in the previous Section but now sample the tissue parameters ψ using Metropolized-Gibbs updates. We initialize the tissue parameters ψ using the average gene expression of each tissue. Often a different initialization for the tissue parameters would lead to our sampler getting stuck in some cluster configuration that is similar to the tissue configuration. This shows some evidence that the sampler for the tissue parameters may not be efficient enough to get out of a poor initialization. This shows some evidence that the Metropolized-Gibbs sampler used for the tissue parameters is not efficient enough to traverse a poor initialization. An alternative sampling techniques that uses gradient information such as Hamiltonian Monte Carlo (HMC) may be a possible solution to remedy this problem. Figure 5.34 shows the inference on the mean parameter. The mean seems to capture the general structure of the data. We observe the tissue effect seems to be captured in the tissue parameters in 5.33a. We also observe the gene specific effects are captured with the gene parameters in 5.33b. Now that the tissue parameters account for the dominant signal in the data the model is able to cluster on a finer hidden structure. Indeed, we observe in Figure 5.35, the clustering is not on the tissue type of the cancer as there is mixing of tissue types within inferred clusters. Expert knowledge is required to interpret this clustering. We also observe the in Figure 5.37, the signal from the tissue effect is no longer present and seems to be captured by the tissue parameters in Figure 5.33a.



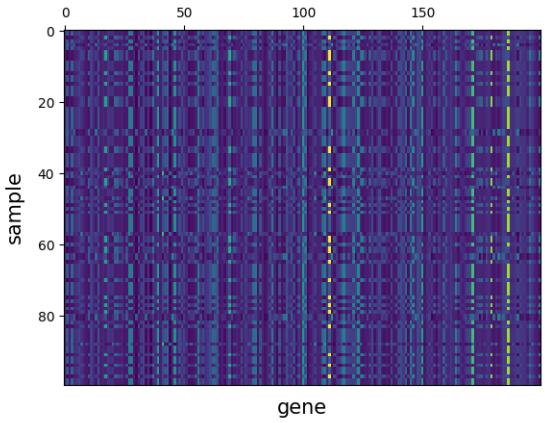
(a) Tissue parameters: $\boldsymbol{T} \cdot \boldsymbol{\psi}$.



(b) Gene parameters: $\mathbb{1}_{N\times 1} \cdot \boldsymbol{\nu}$.

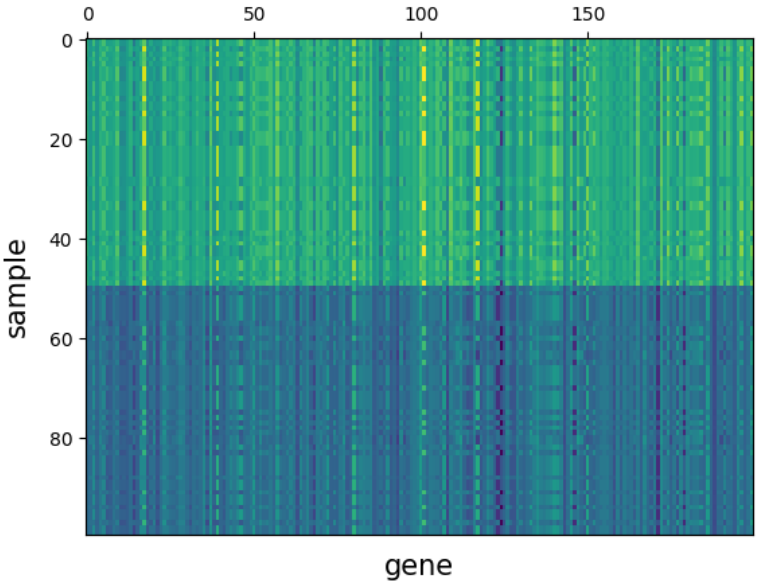


(c) Cluster Mean Parameters: $\boldsymbol{Z} \cdot \boldsymbol{\phi}$.

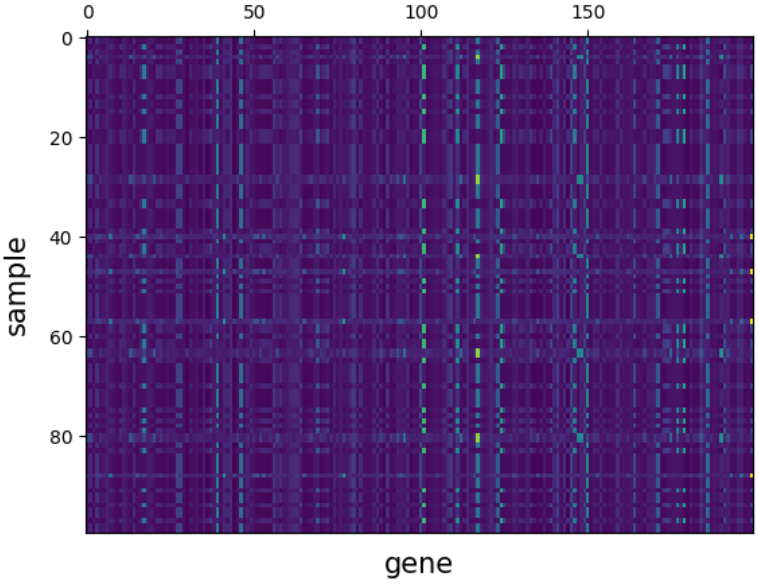


(d) Cluster Variance Parameters.

Figure 5.3: Forward generated parameters for gene expression model used in synthetic data experiments.

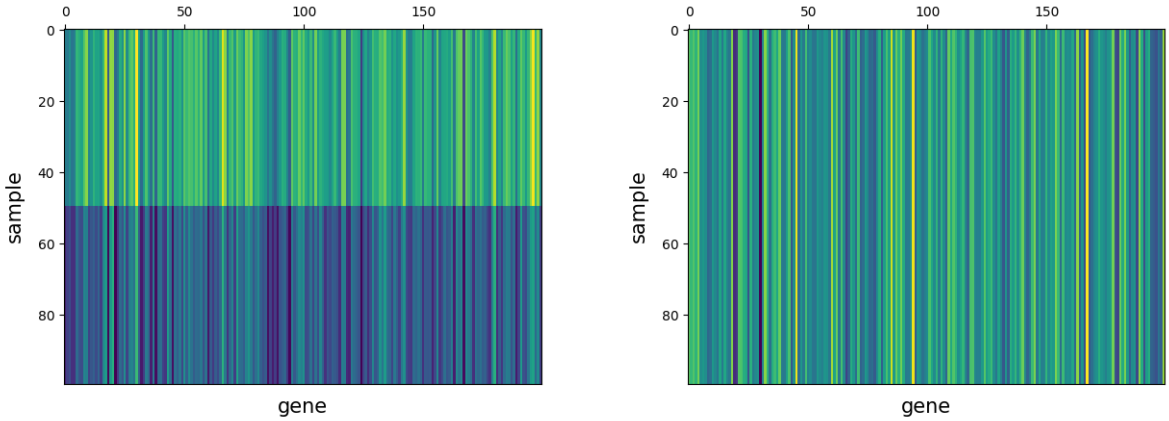


(a) Inferred mean parameters



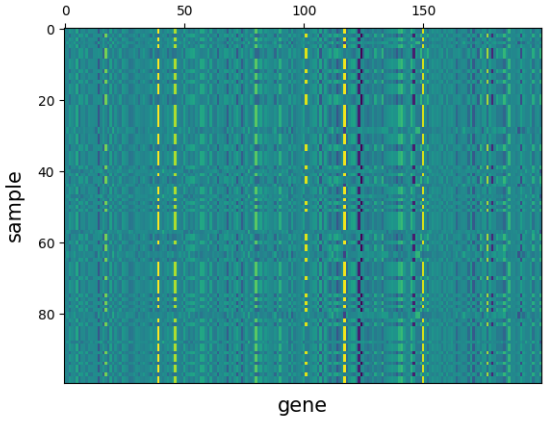
(b) Inferred variance parameters

Figure 5.4: Posterior inference of mean and variance parameters for gene expression model controlling for tissue effects using forward generated data.

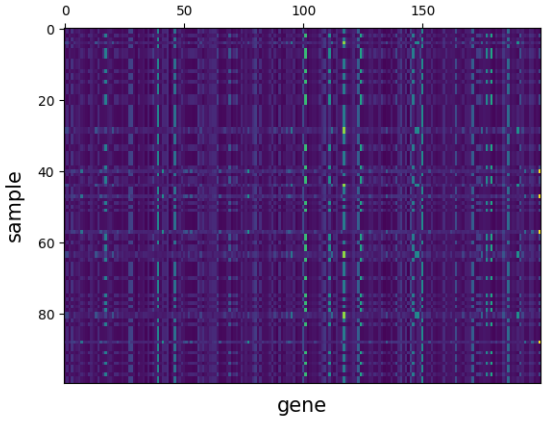


(a) Inferred tissue parameters.

(b) Inferred gene parameters.

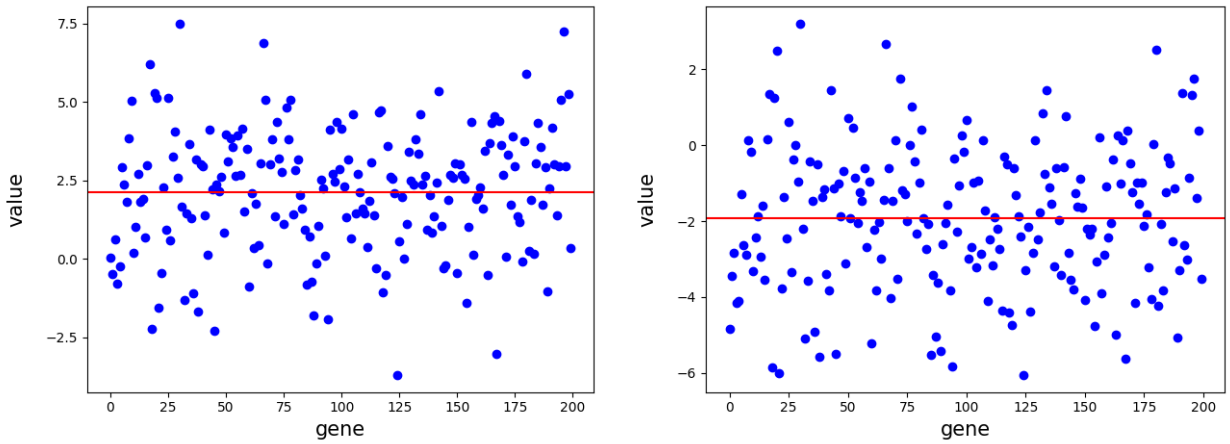


(c) Inferred mean cluster parameters.



(d) Inferred variance cluster parameters.

Figure 5.5: Posterior inference of tissue, gene, and cluster parameters for gene expression model controlling for tissue effects using forward generated data.



(a) Tissue 0 Parameters

(b) Tissue 1 Parameters

Figure 5.6: Sample average of tissue parameters in gene expression model controlling for tissue effects using forward generated data.

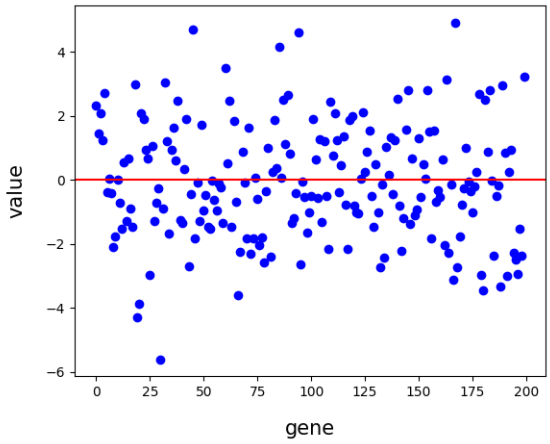
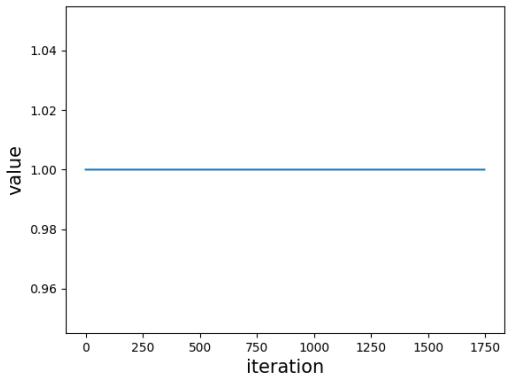
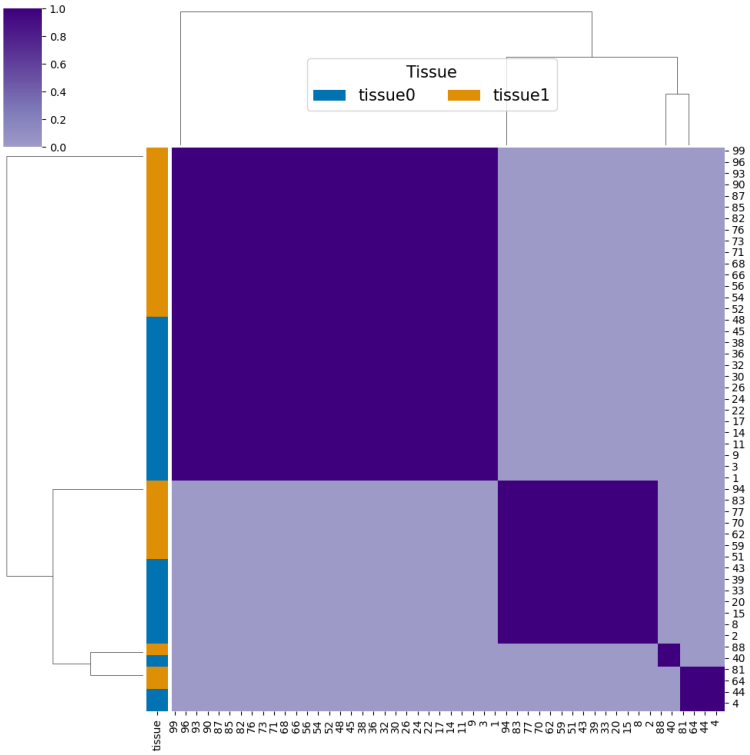


Figure 5.7: Sample average of gene parameters in gene expression model controlling for tissue effects using forward generated data.



(a) V-measure value of each MCMC iteration.



(b) Similarity matrix.

Figure 5.8: V-measure and similarity matrix of MCMC trace for gene expression model controlling for tissue effect using forward generated data. Tissue labels are indicated on the left most axis of similarity matrix.

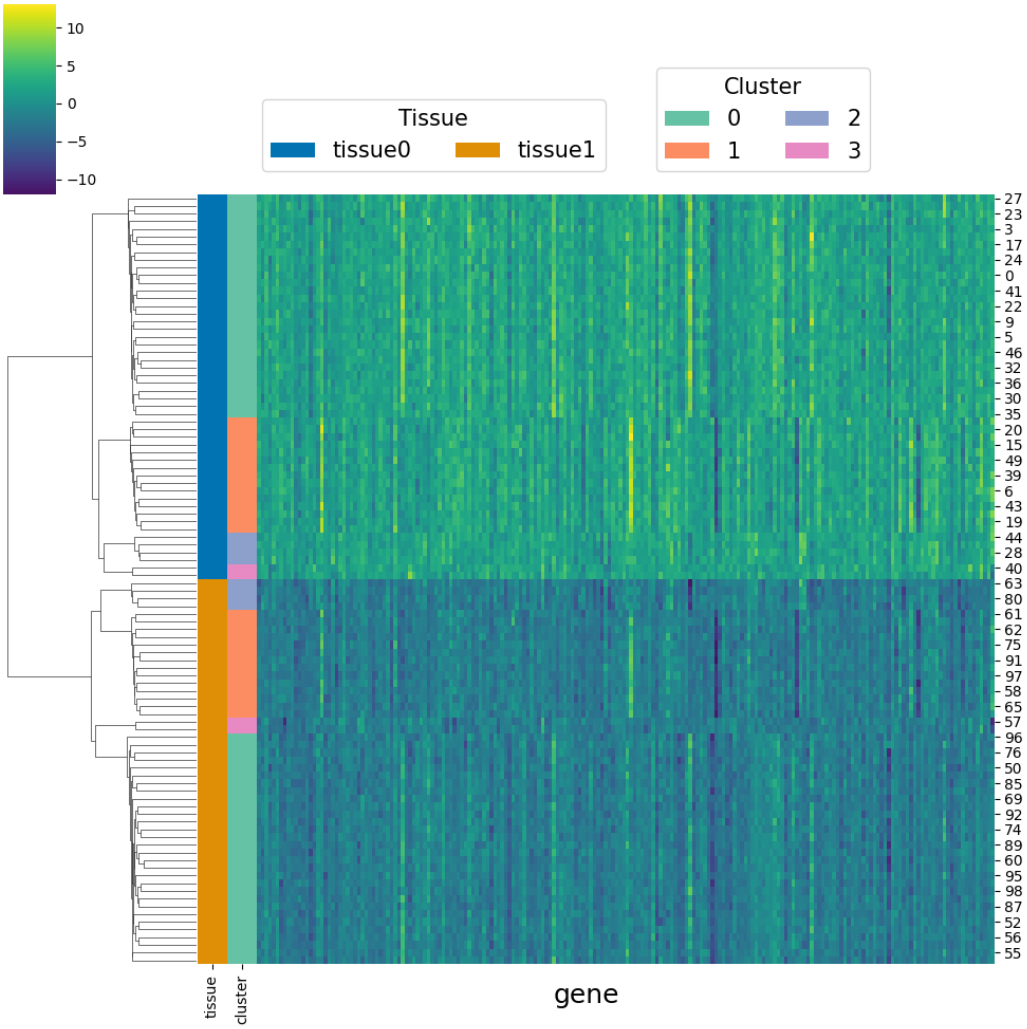


Figure 5.9: Gene expression matrix with tissue and inferred cluster labels. Inference was performed with gene expression model controlling for tissue effect using forward generated data.

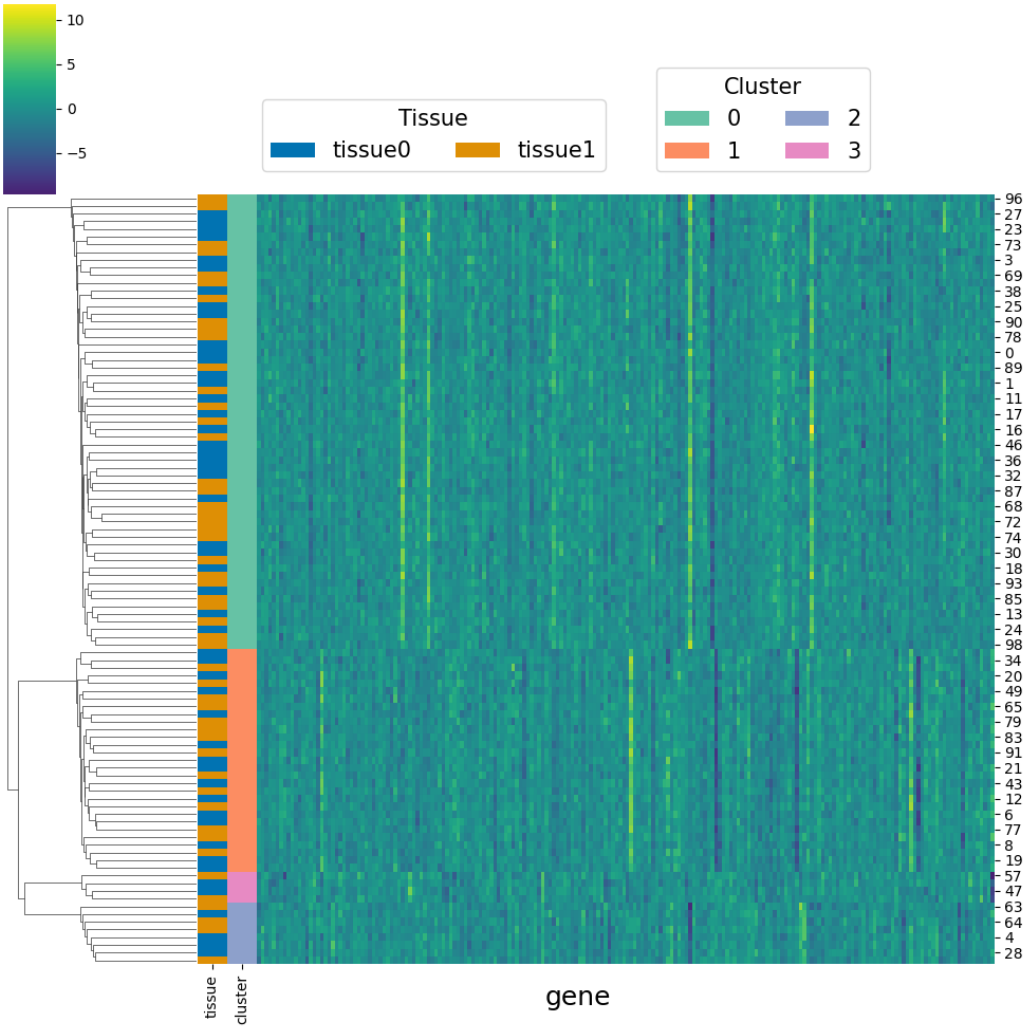


Figure 5.10: Shifted gene expression matrix with tissue and inferred cluster labels. Inference was performed with gene expression model controlling for tissue effect using forward generated data.

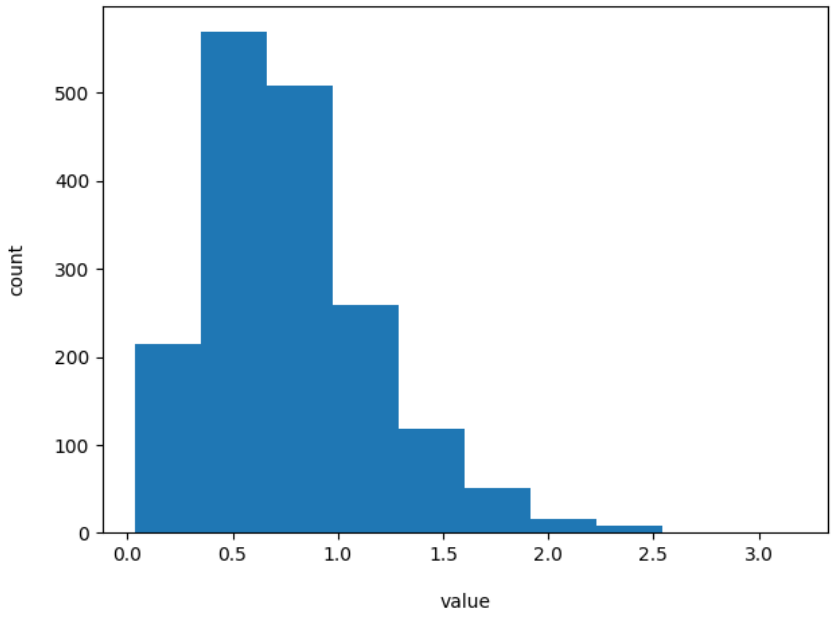


Figure 5.11: Posterior samples of Dirichlet Process hyper-parameter α from synthetic data experiment.

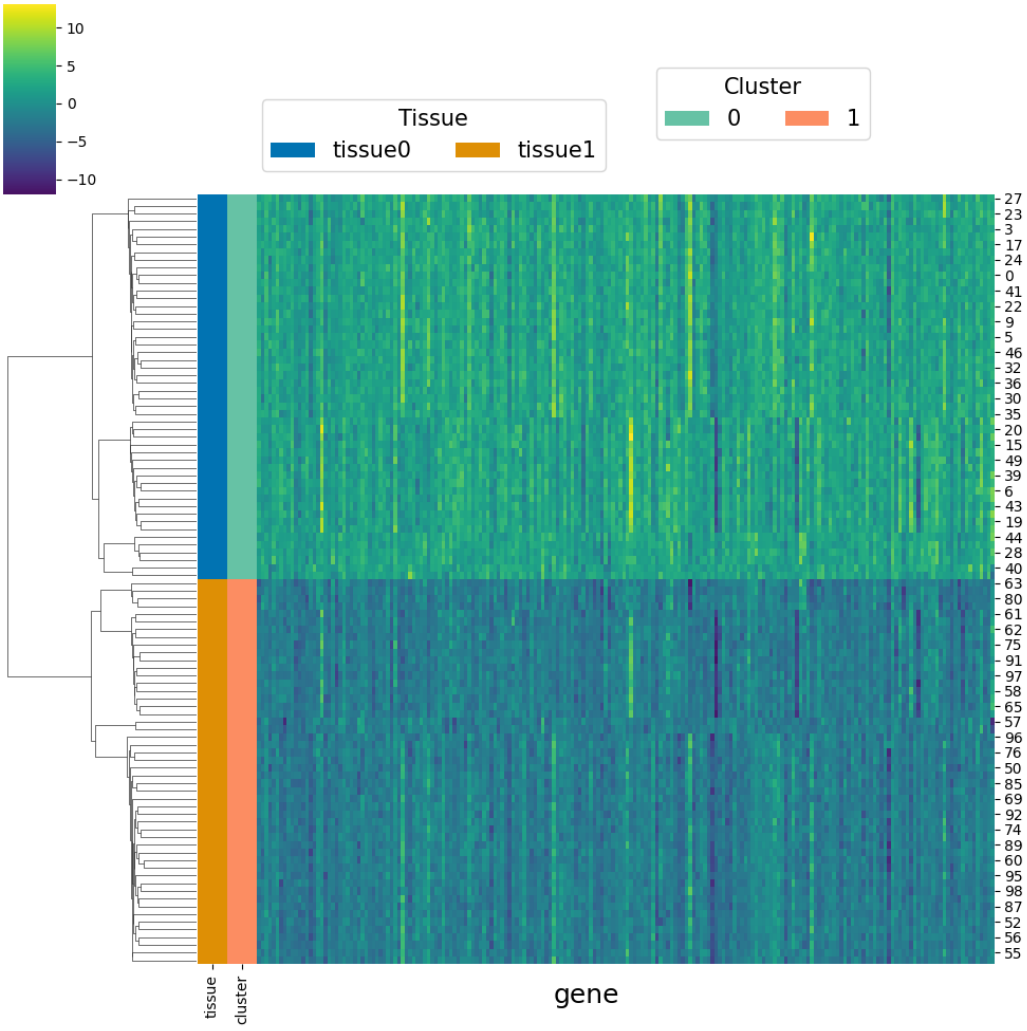


Figure 5.12: Gene expression matrix and shifted gene expression matrix with tissue and inferred cluster labels. Inference was performed with gene expression model not controlling for tissue effect using forward generated data.

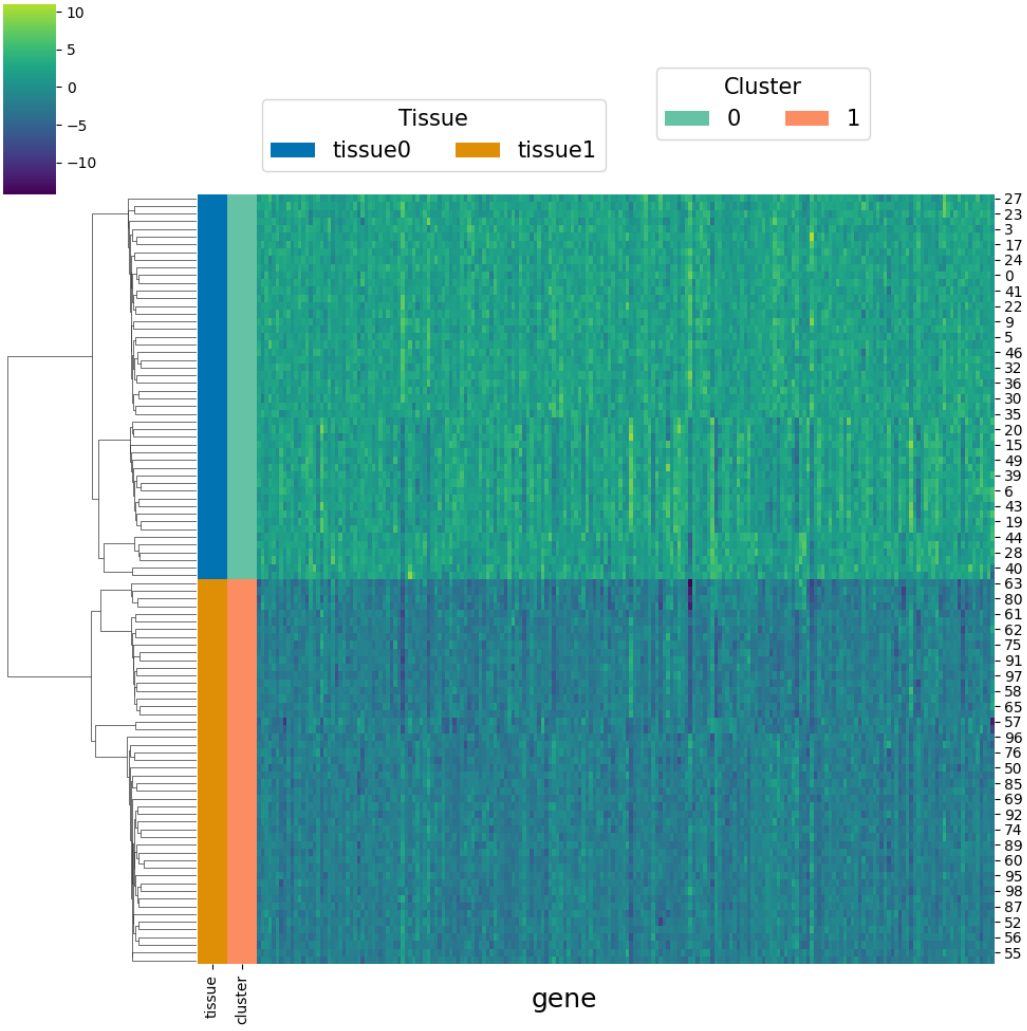
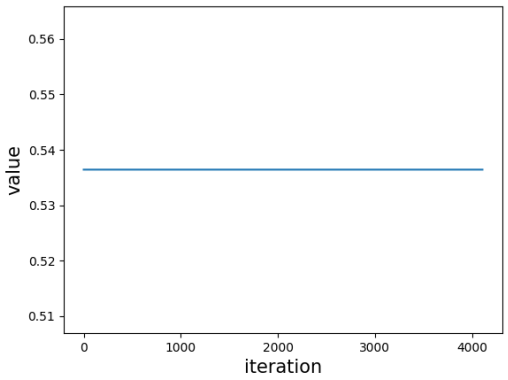
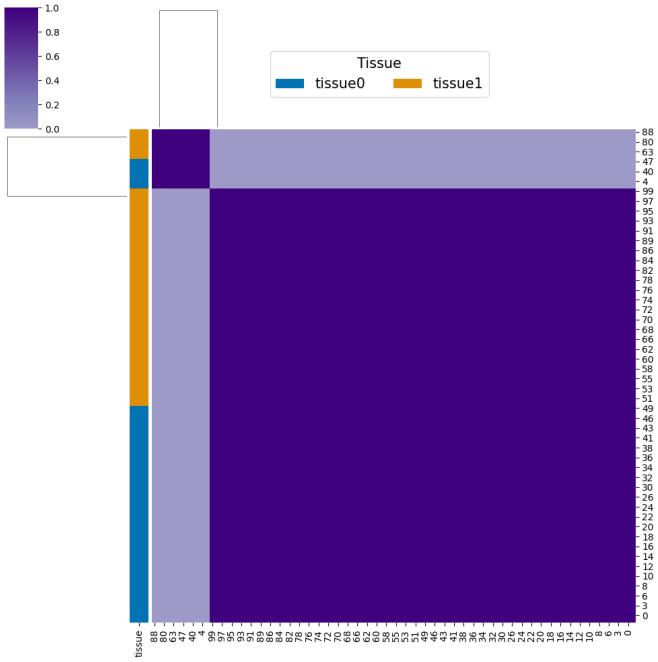


Figure 5.13: Shifted gene expression matrix and shifted gene expression matrix with tissue and inferred cluster labels. Inference was performed with gene expression model not controlling for tissue effect using forward generated data.



(a) V-measure value of each MCMC iteration.



(b) Similarity matrix.

Figure 5.14: V-measure and similarity matrix of MCMC trace for gene expression model controlling for tissue effect using forward generated data. Posterior inference was performed without Split-Merge sampler. Tissue labels are indicated on the left most axis of the similarity matrix.

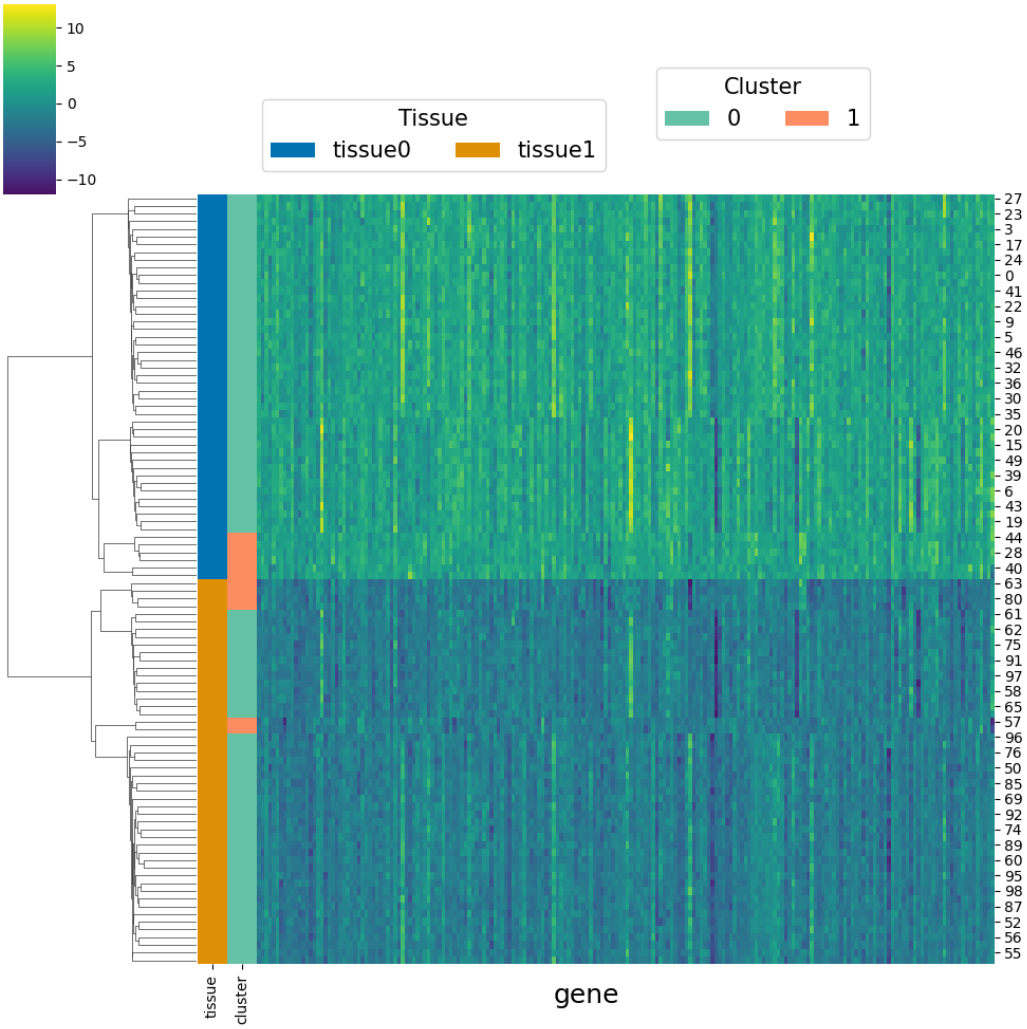


Figure 5.15: Gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model controlling for tissue effect using forward generated data without Split-Merge sampler.

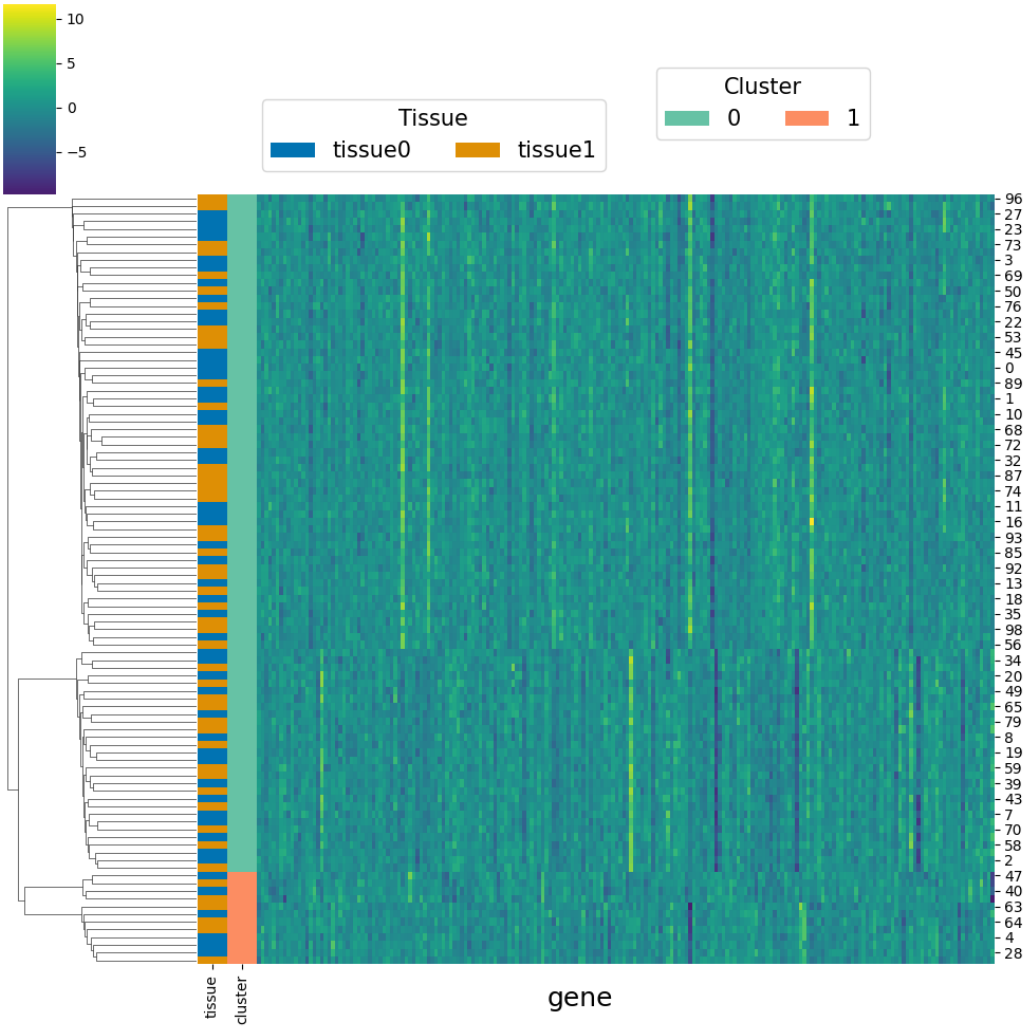
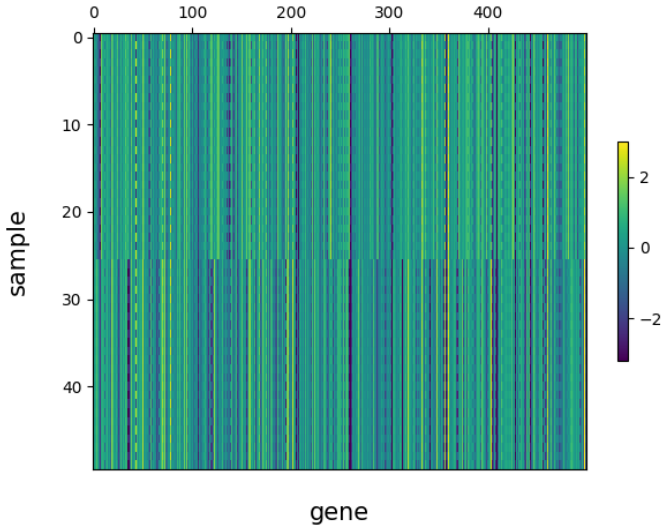
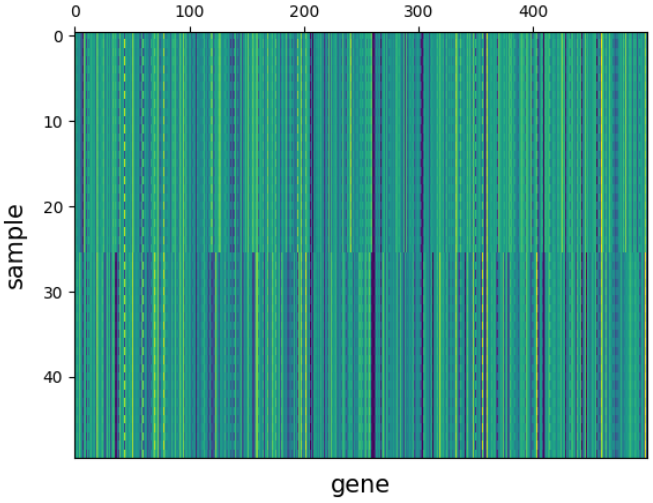


Figure 5.16: Shifted gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model controlling for tissue effect using forward generated data without Split-Merge sampler.

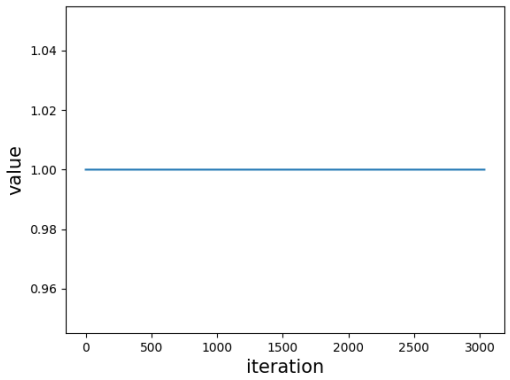


(a) Semi-synthetic data generated using GTEx portal.

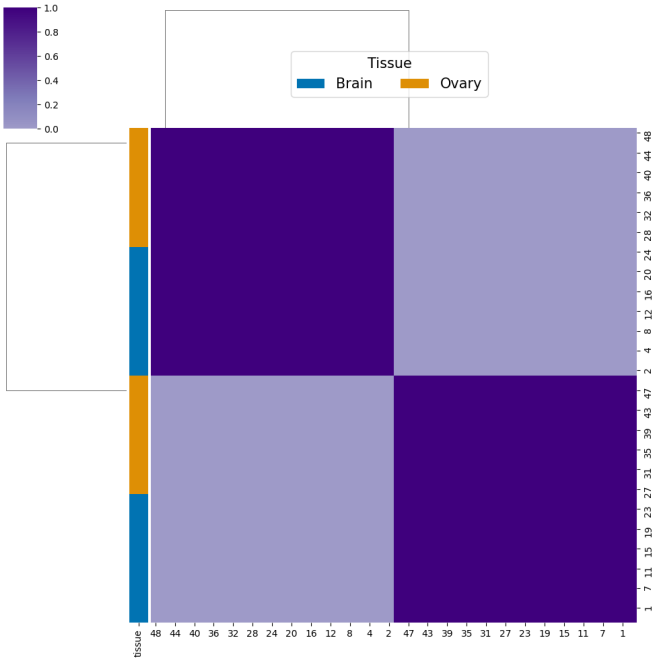


(b) Inferred mean parameter.

Figure 5.17: Semi-synthetic data and posterior inference of mean parameter for gene expression model controlling for tissue effect using data generated with GTEx portal.



(a) V-measure value of each MCMC iteration.



(b) Similarity matrix.

Figure 5.18: V-measure and similarity matrix of MCMC trace for gene expression model controlling for tissue effect using semi-synthetic data generated from GTEx portal. Tissue labels are indicated on the left most axis of the similarity matrix.

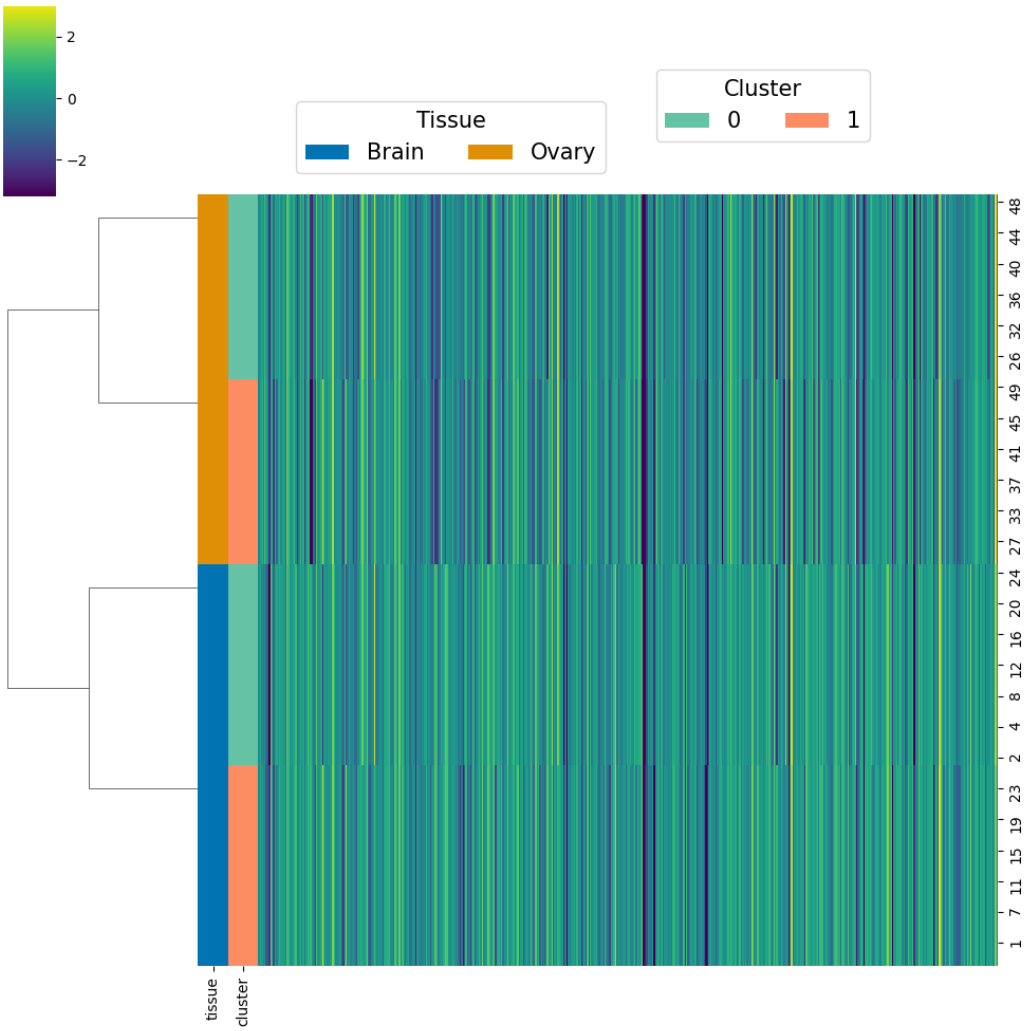


Figure 5.19: Gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model controlling for tissue effect using semi-synthetic data generated with GTEx portal.

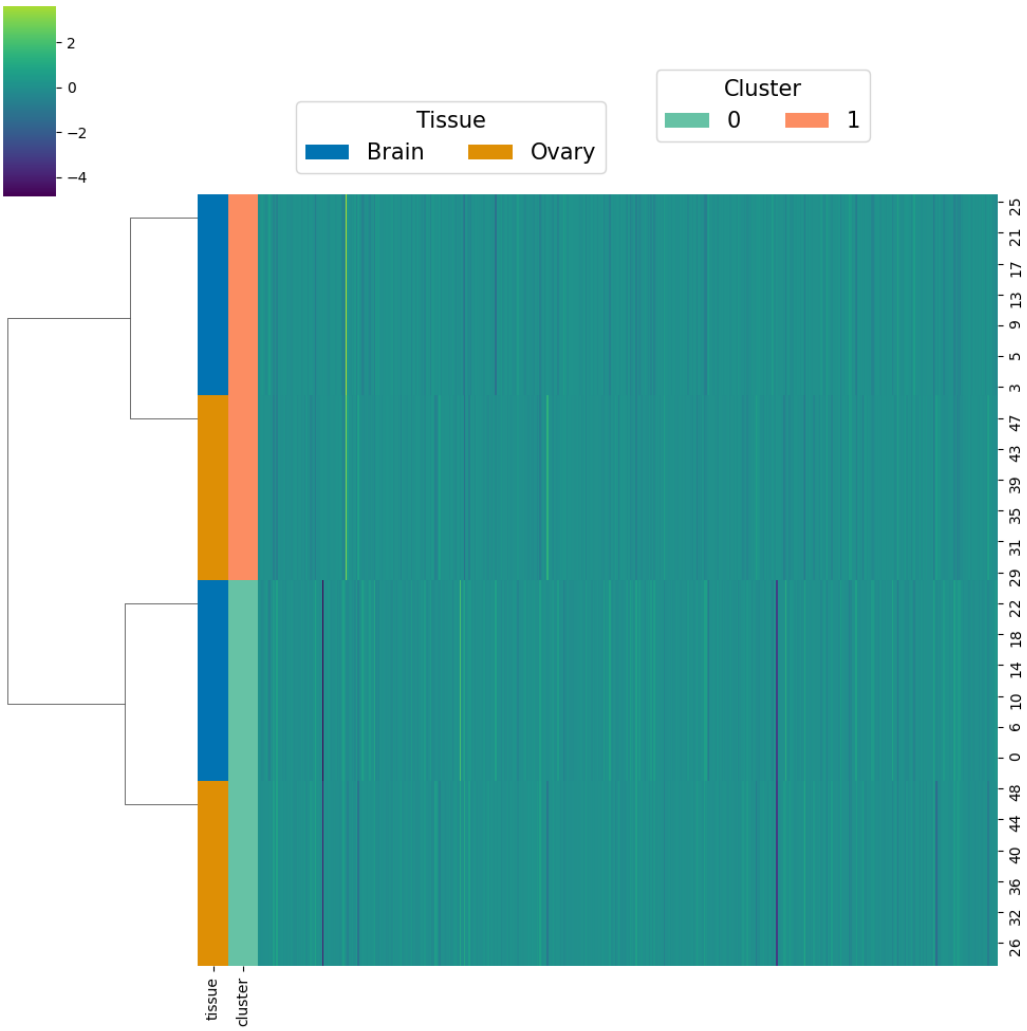


Figure 5.20: Shifted gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model controlling for tissue effect using semi-synthetic data generated with GTEx portal.

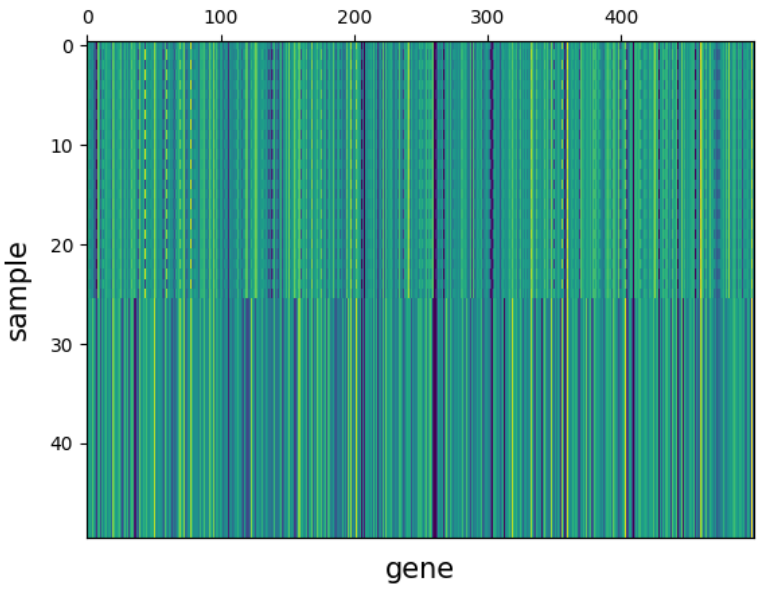
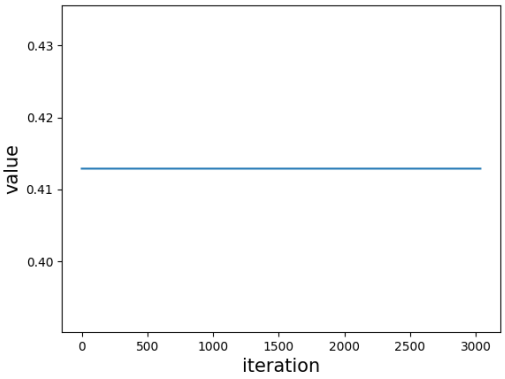
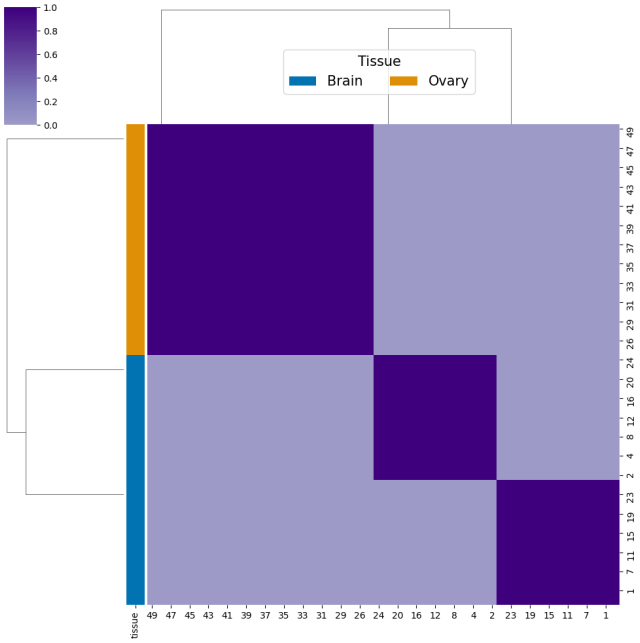


Figure 5.21: Posterior inference for mean parameter of gene expression model not controlling for tissue effect using semi-synthetic data generated from GTEx portal.



(a) V-measure value of each MCMC iteration.



(b) Similarity matrix.

Figure 5.22: V-measure and similarity matrix of MCMC trace for gene expression model not controlling for tissue effect using semi-synthetic data generated with GTEx portal. Tissue labels are indicated on the left most axis of the similarity matrix.

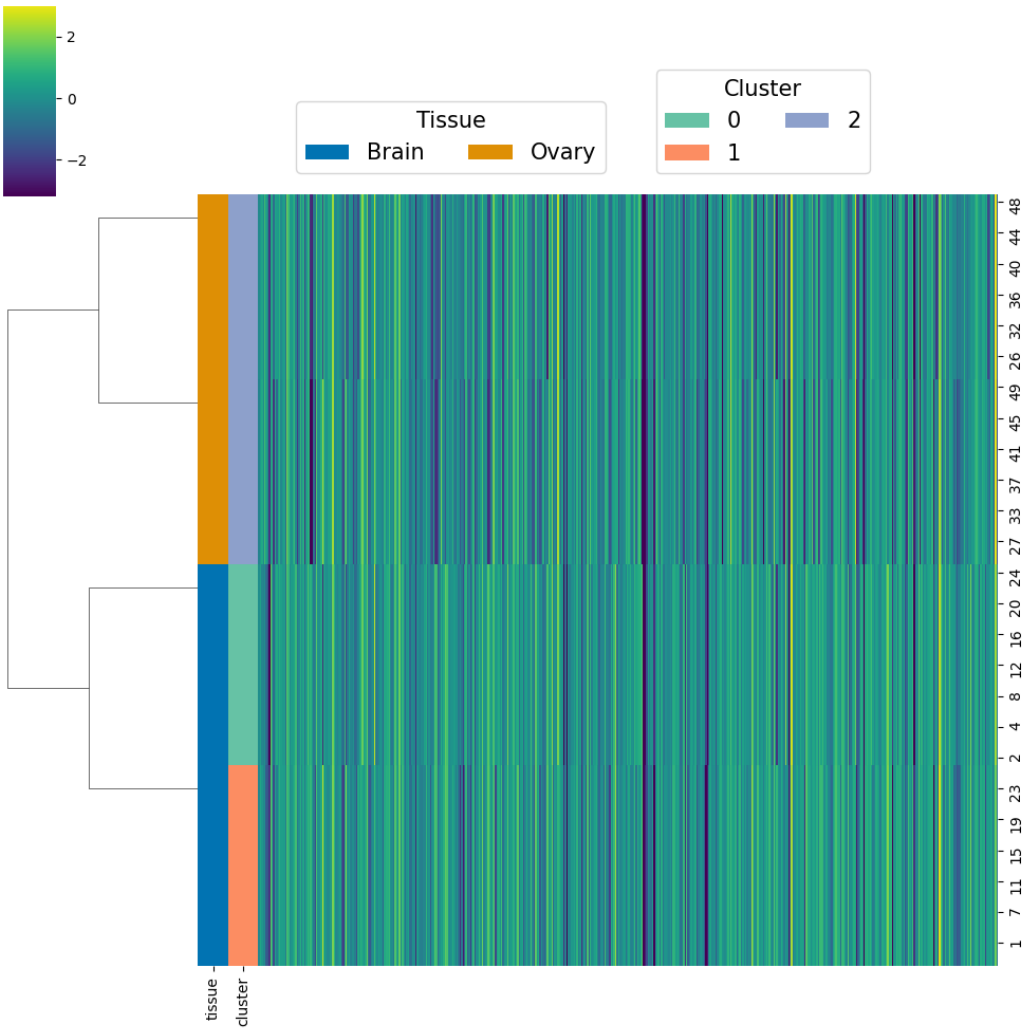


Figure 5.23: Gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model not controlling for tissue effect using semi-synthetic data generated with GTEx portal.

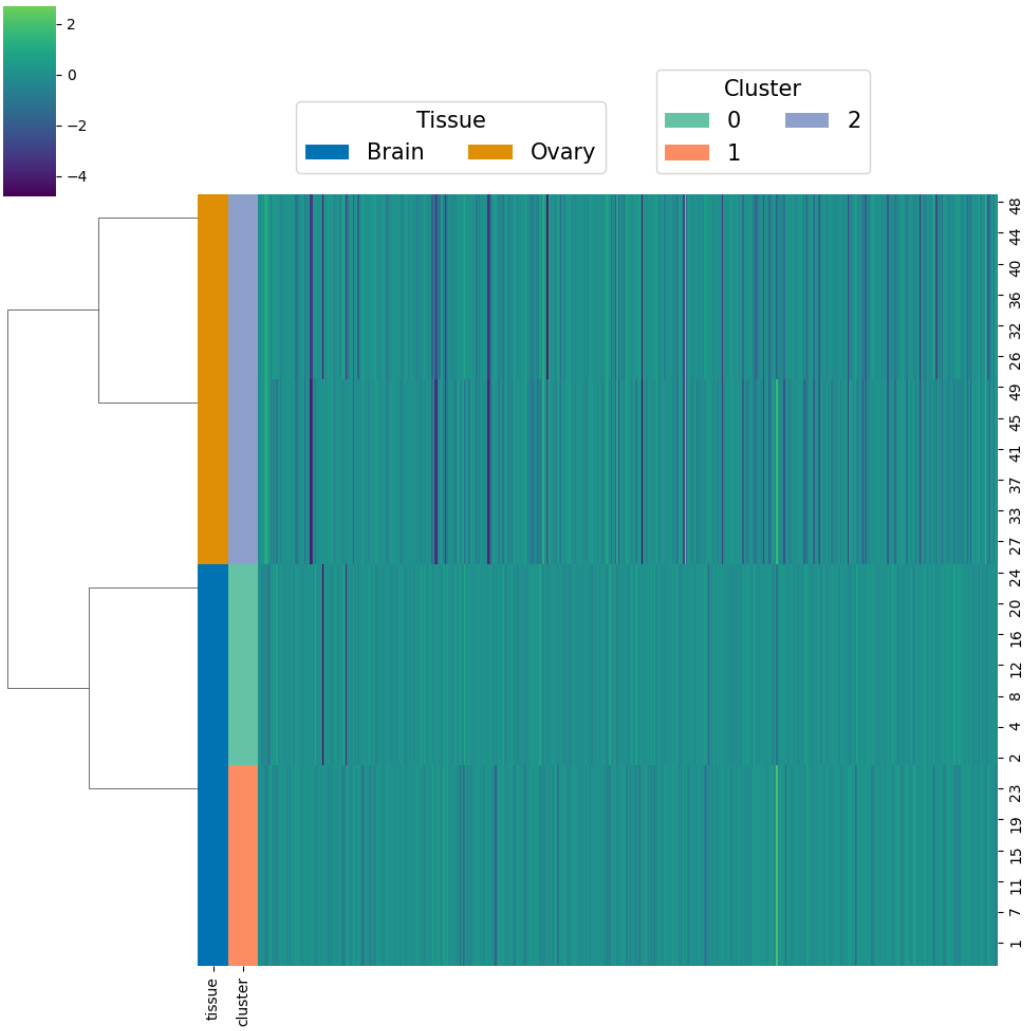
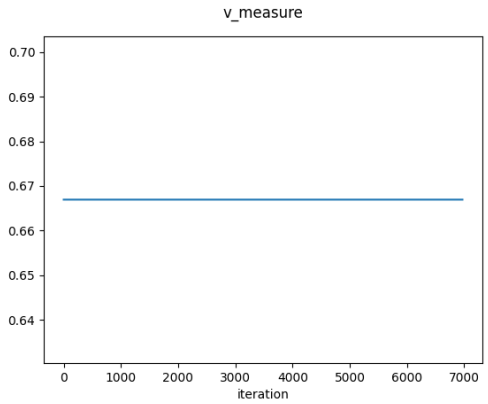
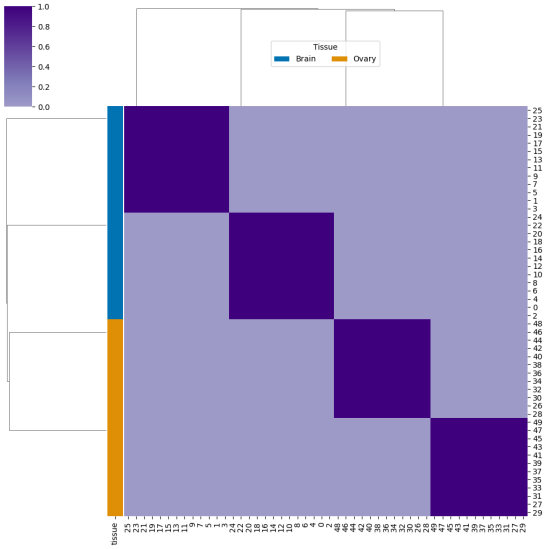


Figure 5.24: Shifted gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model not controlling for tissue effect using semi-synthetic data generated with GTEx portal.



(a) V-measure value of each MCMC iteration.



(b) Similarity matrix.

Figure 5.25: V-measure and similarity matrix of MCMC trace for gene expression model controlling for tissue effect using semi-synthetic data generated with GTEx portal. Posterior inference was performed without Split-Merge sampler. Tissue labels are indicated on the left most axis of the similarity matrix.

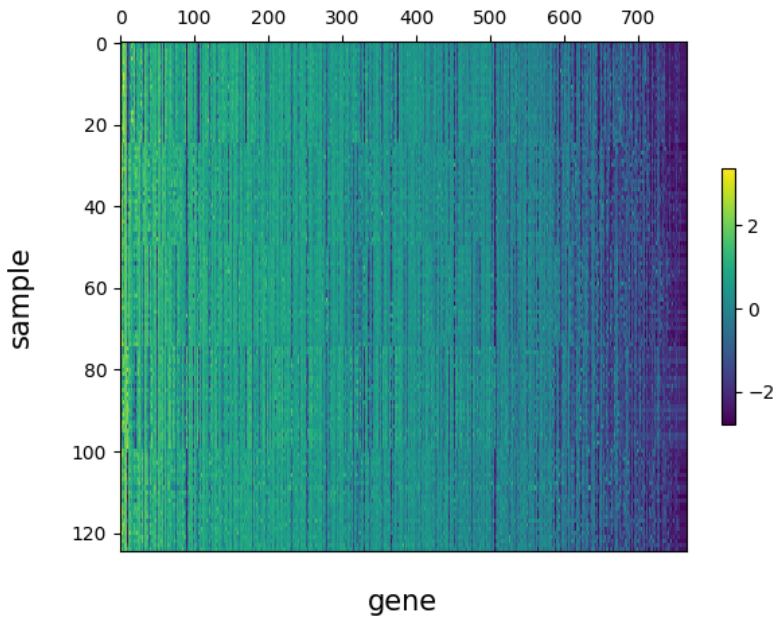


Figure 5.26: Bulk RNA-sequencing data from International Cancer Genome Consortium (ICGC) portal of 125 samples from breast, brain, liver, colorectal, and lung tissues and 766 genes selected from NanoString Technologies pan-cancer pathway panel.

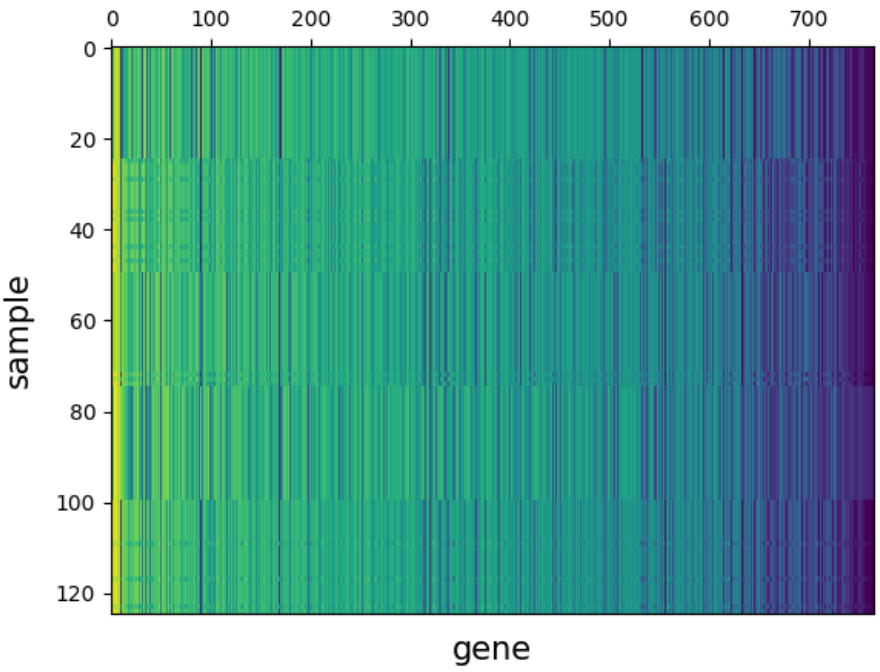
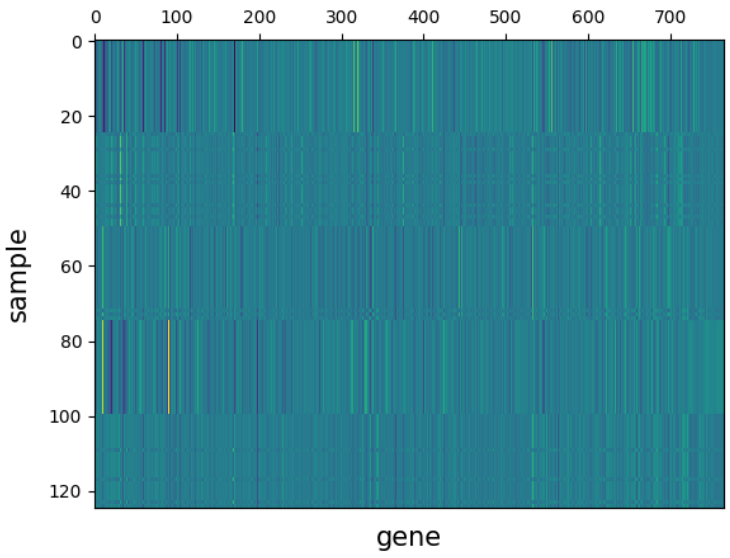
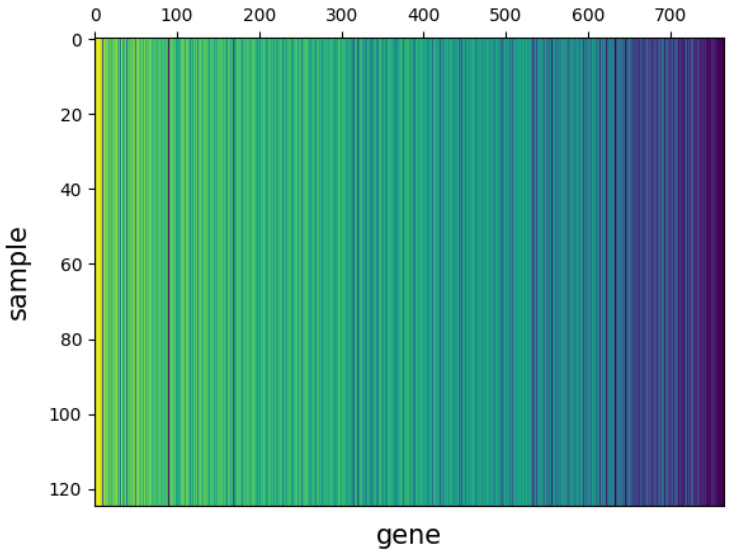


Figure 5.27: Posterior inference for mean parameter of gene expression model not controlling for tissue effect using ICGC data.



(a) Inferred cluster parameter.



(b) Inferred gene parameter.

Figure 5.28: Posterior inference for cluster and gene parameter of gene expression model not controlling for tissue effect using ICGC data.

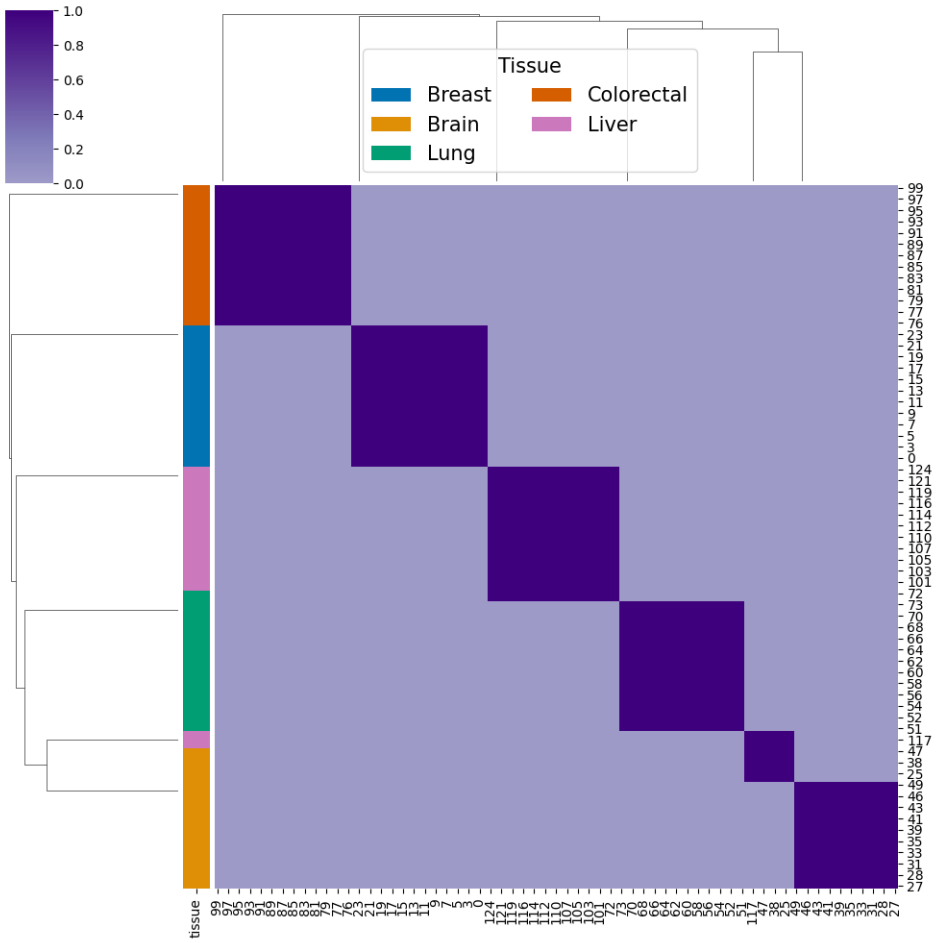


Figure 5.29: Similarity matrix of MCMC trace for gene expression model not controlling for tissue effect using ICGC data. Tissue labels are indicated on the left most axis of the similarity matrix.

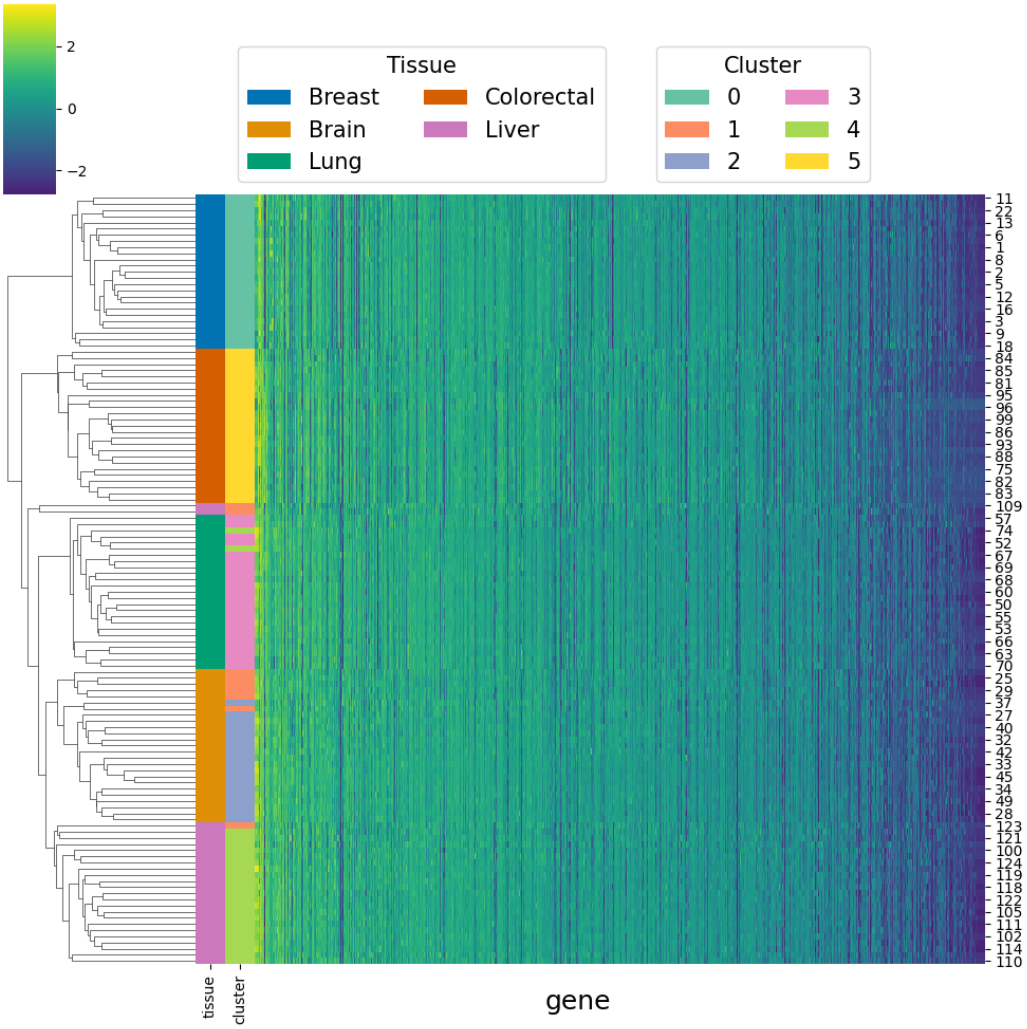


Figure 5.30: Gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model not controlling for tissue effect using ICGC data.

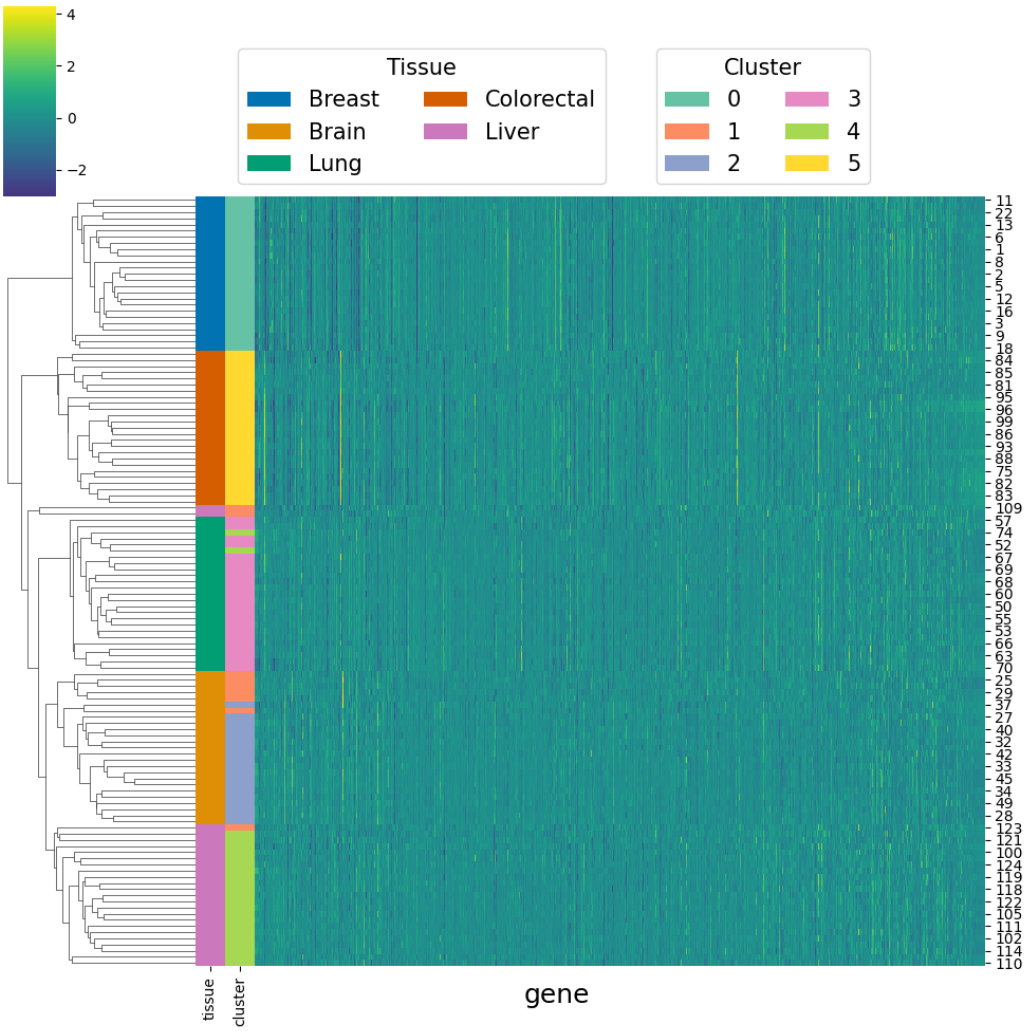


Figure 5.31: Shifted gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model not controlling for tissue effect using ICGC data.

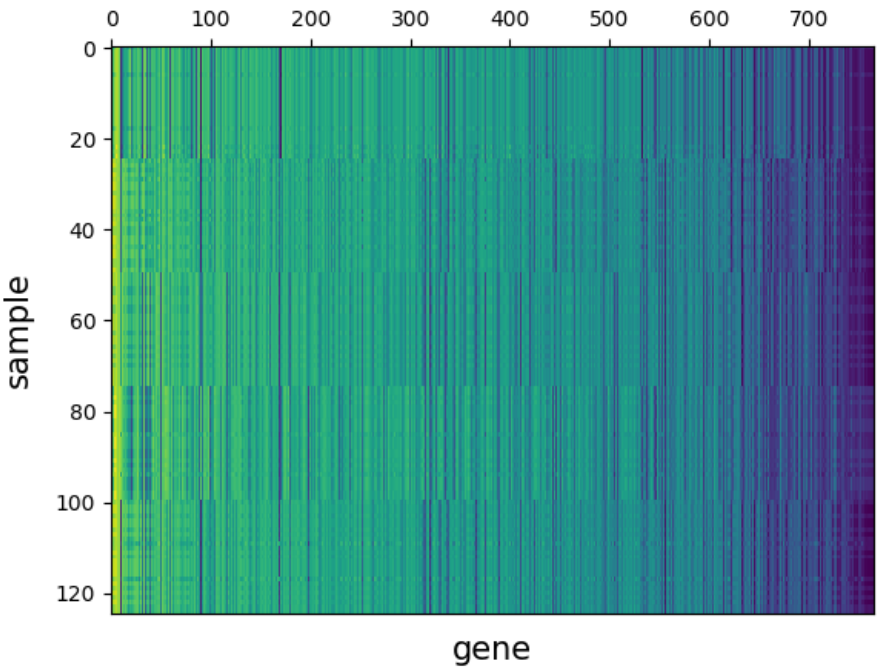
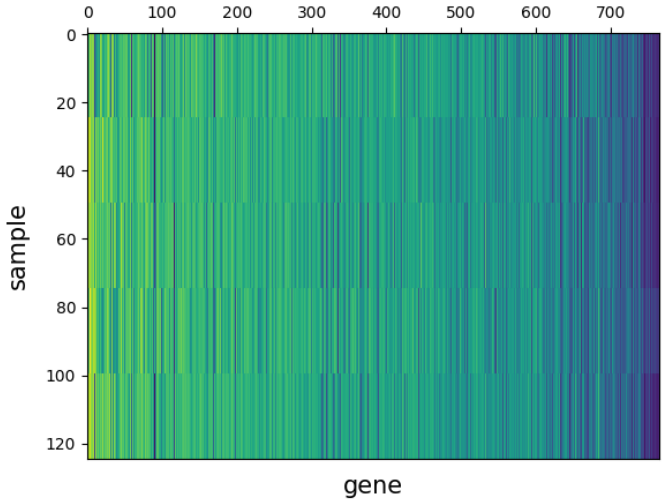
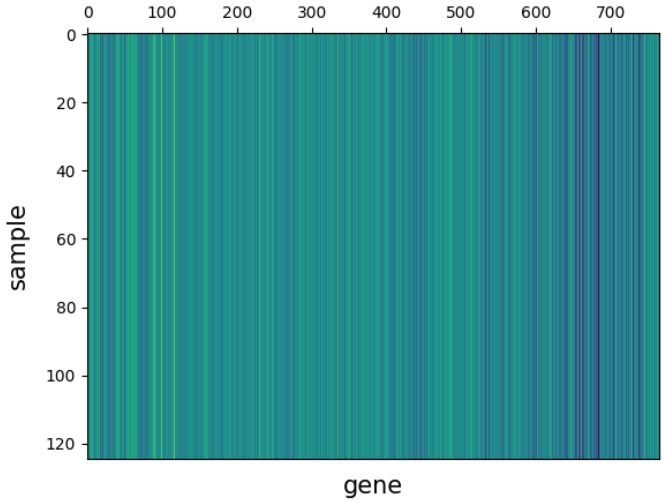


Figure 5.32: Posterior inference for mean parameter of gene expression model controlling for tissue effect using ICGC data.



(a) Inferred tissue parameter.



(b) Inferred gene parameter.

Figure 5.33: Posterior inference on tissue and gene parameters of the gene expression model controlling for tissue effects using ICGC data.

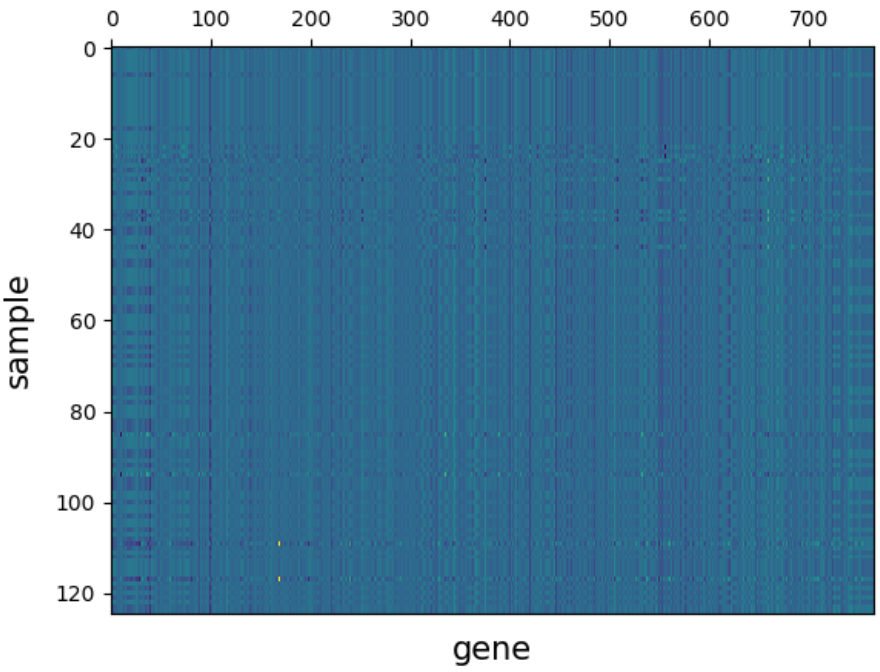


Figure 5.34: Posterior inference on cluster parameters of the gene expression model controlling for tissue effects using ICGC data.

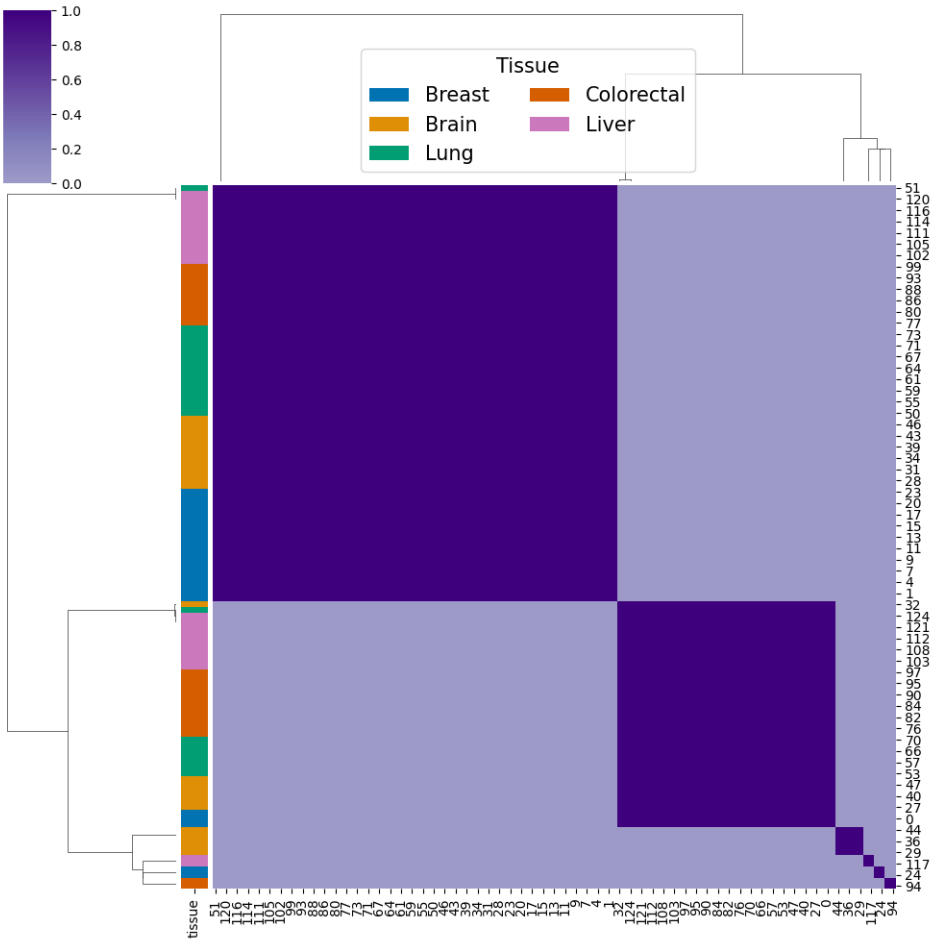


Figure 5.35: Similarity matrix of MCMC trace for gene expression model controlling for tissue effect using ICGC datas. Tissue labels are indicated on the left most axis of the similarity matrix.

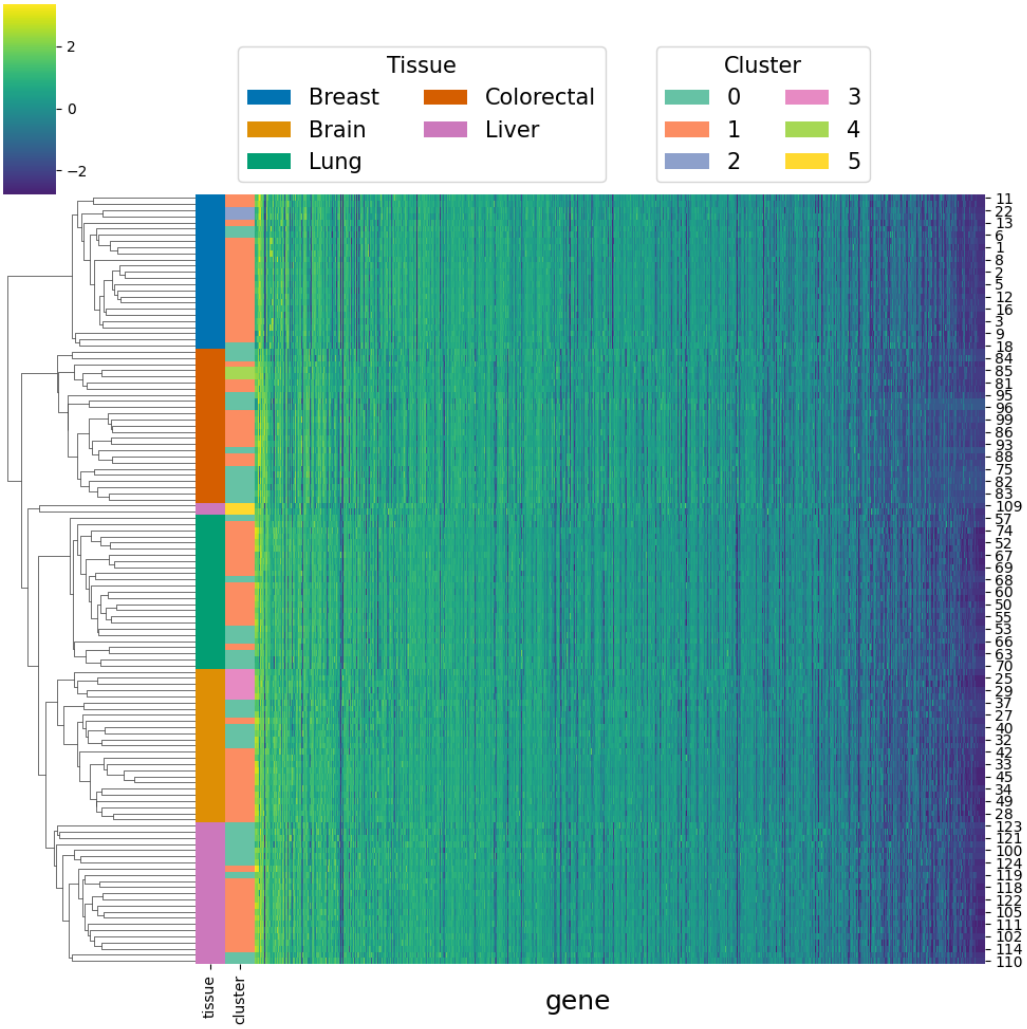


Figure 5.36: Gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model controlling for tissue effect using ICGC data.

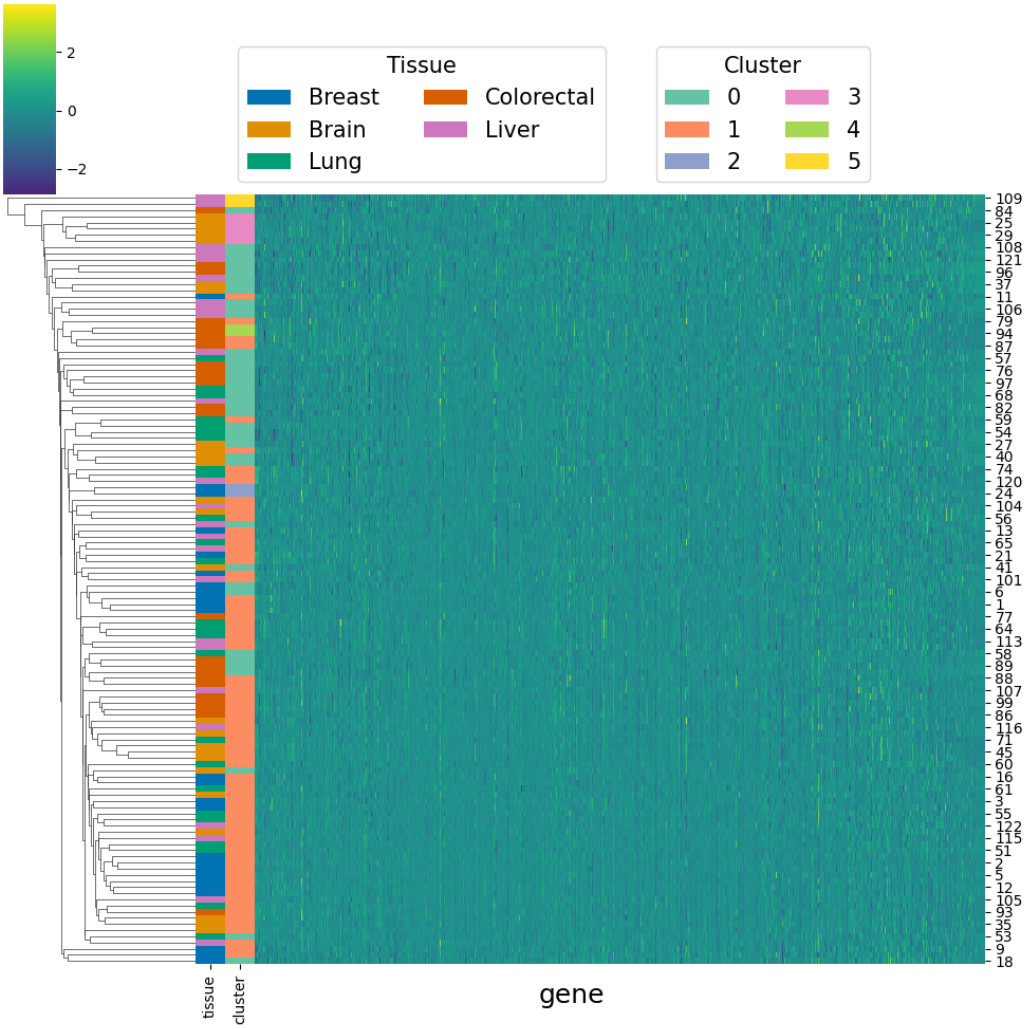


Figure 5.37: Shifted gene expression matrix with tissue and inferred cluster labels. Inference was performed on gene expression model controlling for tissue effect using ICGC data.

Chapter 6

Conclusions and Future Directions

We introduced a Bayesian nonparametric model that controls for the tissue effect and clusters based on a latent structure using a Dirichlet Process prior. This model learns the tissue effect by using tissue parameters in a supervised learning setting, while simultaneously learning the latent structure based on the resulting residuals in an unsupervised setting. We demonstrated our model by showing results on synthetic data, semi-synthetic data generated from the Genome-Tissue Expression portal (GTEx), and a publicly available dataset from the International Cancer Genome Consortium (ICGC). Results on synthetic data showed the model was able to capture the tissue and gene effects and consequently cluster on the latent secondary structure. In all experiments, the Split-Merge sampler was vital for inference as the Gibbs sampler would often get stuck in a cluster configuration related to the observations tissue assignments. In the real-world application using ICGC data, an informed initialization for the tissue parameters was important otherwise inference would also get stuck in a cluster configuration related to the tissue configuration.

Expert knowledge may be needed to interpret the inferred clustering on

the ICGC experiments. That is, a biological analysis is needed to determine if the inferred clustering is biologically meaningful. One possible method to start could include investigating which genes each inferred cluster up-regulate and down-regulate to determine if they fit the pattern of a specific gene pathway.

As alluded to in Section 5.3.3, a more efficient sampler for the tissue and gene parameters could help the model better control for these effects in the data. We cluster on the residuals of the data after controlling for the tissue and gene parameters therefore the inference on these parameters is vital for clustering. One possible method would be to leverage gradient inference through the use of Hamiltonian Monte Carlo introduced in Neal (2012). This may be an easy implementation as we model the data, tissue parameters, and gene parameter with Gaussian distributions.

One possible method to model bulk RNA-seq data is to use a Negative Binomial distribution. This method would allow us to model the raw data instead of the normalized data where we could include another parameter to account for the scale of each observation. This would involved the use of non-conjugate samplers for the cluster assignments. Non-conjugate Gibbs samplers do exist, however there exists a gap in the scholarly knowledge to non-conjugate split-merge samplers. One may be able to leverage SMC and PMCMC methods to develop a non-conjugate split-merge sampler similar to the PGSM sampler.

Bibliography

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W. T., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., et al. (2020). The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Bouchard-Côté, A., Doucet, A., and Roth, A. (2017). Particle Gibbs split-merge sampling for Bayesian inference in mixture models. *Journal of Machine Learning Research*, 18(28):1–39.

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H. K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C.-L., Faquin, W. C., et al. (2020). Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I.,

Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

Chakravarthy, A., Furness, A., Joshi, K., Ghorani, E., Ford, K., Ward, M. J., King, E. V., Lechner, M., Marafioti, T., Quezada, S. A., et al. (2018). Pan-cancer deconvolution of tumour composition using DNA methylation. *Nature Communications*, 9(1):1–13.

Chen, F., Zhang, Y., Gibbons, D. L., Deneen, B., Kwiatkowski, D. J., Ittmann, M., and Creighton, C. J. (2018). Pan-cancer molecular classes transcending tumor lineage across 32 cancer types, multiple data platforms, and over 10,000 cases. *Clinical Cancer Research*, 24(9):2182–2193.

Chopin, N. and Papaspiliopoulos, O. (2020). *An introduction to Sequential Monte Carlo*. Springer, Cham, Switzerland.

Colombo, P.-E., Milanezi, F., Weigelt, B., and Reis-Filho, J. S. (2011). Microarrays in the 2010s : The contribution of microarray-based gene expression profiling to breast cancer classification, prognostication and prediction. *Breast Cancer Research*, 13(3):1–15.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.

Das, R. (2014). Collapsed Gibbs sampler for Dirichlet process Gaussian mixture models. `http://rajarshd.github.io/talks/DPGMM_tutorial.pdf`.

Desmedt, C., Voet, T., Sotiriou, C., and Campbell, P. J. (2012). Next generation sequencing in breast cancer : First take home messages. *Current Opinion in Oncology*, 24(6):597.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Fritsch, A. and Ickstadt, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367 – 391.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110 – 120.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2005). Interpolating between types and tokens by estimating power-law generators. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, page 459–466, Cambridge, MA, USA. MIT Press.

Görür, D. and Edward Rasmussen, C. (2010). Dirichlet process Gaussian mixture models : Choice of the base distribution. *Journal of Computer Science and Technology*, 25(4):653–664.

Grosse, R. (2014). Lecture notes on mixture models. `https://www.cs.toronto.edu/~rgrosse/csc321/mixture_models.pdf`.

Gu, S., Ghahramani, Z., and Turner, R. E. (2015). Neural adaptive Sequential Monte Carlo. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, page 2629–2637, Cambridge, MA, USA. MIT Press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Lapointe, J., Li, C., Higgins, J. P., Van De Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., et al. (2004).

Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences*, 101(3):811–816.

Liu, X., Li, N., Liu, S., Wang, J., Zhang, N., Zheng, X., Leung, K.-S., and Cheng, L. (2019). Normalization methods for the analysis of unbalanced transcriptome data : A review. *Frontiers in Bioengineering and Biotechnology*, 7:358.

Lonsdale, J., Thomas, J., and Salvatore, M. (2013). The genotype-tissue expression project. *Nature Genetics*, 45(6):580–585.

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., et al. (2013). Gene expression classification of colon cancer into molecular subtypes : Characterization, validation, and prognostic value. *PLoS Medicine*, 10(5):e1001453.

Moral, P. D., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society. Series B*, 68(3):411–436.

Murphy, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution. `https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf`.

NanoString Technologies (2015). nCounter pancancer pathways panel. `https://www.nanostring.com/products/ncounter-assays-panels/oncology/ncounter-pancancer-pathways-panel/`.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Neal, R. M. (2012). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*.

Orbanz, P. (2015). Lecture notes on Bayesian nonparametrics. `http://www.gatsby.ucl.ac.uk/~porbanz/papers/porbanz_BNP_draft.pdf`.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160.

Pedregosa, F., Varoquaux, G., and Gramfort, A. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Rasmussen, C. E. (1999). The infinite Gaussian mixture model. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 554–560, Cambridge, MA, USA. MIT Press.

Reynolds, D. (2009). Gaussian mixture models. *Encyclopedia of Biometrics*, pages 659–663.

Robert, C. P. and Roberts, G. O. (2021). Rao-Blackwellization in the MCMC era. arXiv:2101.01011.

Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M., Groß, M., Backofen, R., and Diederichs, S. (2019). A pan-cancer analysis of synonymous mutations. *Nature Communications*, 10(1):1–14.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1):1–18.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

Teh, Y. W. (2010). Dirichlet process. *Encyclopedia of Machine Learning*, pages 280–287.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Yang, T.-H. O., Porta-Pardo, E., Gao, G. F., Plaisier, C. L., Eddy, J. A., et al. (2018). The immune landscape of cancer. *Immunity*, 48(4):812–830.

Tinker, A. V., Boussioutas, A., and Bowtell, D. D. (2006). The challenges of gene expression microarrays for the study of human cancer. *Cancer cell*, 9(5):333–339.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq : A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models.

Wood, F., Goldwater, S., and Black, M. J. (2006). A nonparametric Bayesian approach to spike sorting. *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1165–1168.

Xu, Y., Müller, P., and Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes. *Biometrics*, 72(3):955–964.

Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L. D., and Ferretti, V. (2019). The international cancer genome consortium data portal. *Nature Biotechnology*, 37(4):367–369.