ONLINE BAYESIAN NONPARAMETRIC MIXTURE MODELS VIA REGRESSION

A THESIS SUBMITTED TO

THE UNIVERSITY OF KENT AT CANTERBURY

IN THE SUBJECT OF STATISTICS

FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY BY RESEARCH

By

An Kang

March 2018

Abstract

Sensors are widely used in modern industrial systems as well as consumer devices, such as food production, energy transportation and nuclear power plant. The sensors of interest in this project from an engineering company are associated with industrial control systems, where high precision is the dominant concern. Due to manufacturing variation, sensors manufactured from the same production line are non-identical. Therefore, each sensor needs to be characterised via parameter estimation to achieve a high precision or accuracy before sending to the end users. The classical linear regression model has been adopted in current industry procedure, which requires a certain number of measurements per device to achieve the required level of accuracy. The aim of the project is, under guarantee of the required level of accuracy, to use the available information and advanced statistical models to reduce the number of measurements needed per sensor, and hence reduce both costs and time for the characterisation process.

To achieve this, a Bayesian linear model with Dirichlet process mixture prior (BL-DPMP) is proposed, where the Bayesian linear regression presents the relationship between the response variable and the covariates demonstrated to be appropriate by the company, and the regression coefficients are modelled by a Dirichlet process mixture (DPM) model. The idea here is to apply the DPM model to the historical information from similar sensors to provide adequate prior information to the linear regression model in order to compensate the current characterising sensor with block missing measurements, at the same time to maintain the required level of accuracy. The slice sampling scheme based on the full conditional posteriors of hyperparameters is used to update the parameters in

the DPM model. Also, a generalised Dirichlet process mixture regression model is proposed with a data-driven prediction procedure to deal with the considered situation.

By reducing the number of measurements required per sensor, we could drastically reduce the characterisation period. However, two proposed approaches are quite computationally intensive, which counteract the time saved from collecting a fewer number of measurements. Hence, there is a clearly pressing need for dramatically faster alternatives. A hybrid Variational Bayes (HVB) procedure following a greedy searching scheme is proposed, which can dramatically reduce the computational time, at the same time provide highly accurate approximations of the exact posterior distributions.

The ultimate goal of this project is to implement the proposed advanced statistical model in the production line, where the model can be executed within seconds (online). An optimal permutation sequential (OPS) algorithm for the DPM model is proposed, which differs from MCMC algorithms. The idea is to draw approximate independent and identically distributed samples from the posterior distribution of the latent allocations, and to draw samples from the weights and locations conditional on the allocations. Hence, independent draws are taken from the posterior distribution, which allow us to take independent samples from the predictive distribution. The OPS algorithm requires only a single run which the computational costs of a few seconds. We present examples to show model performance on simulated and real datasets.

It is worth noting that the proposed Bayesian linear model with Dirichlet process mixture prior together with the OPS algorithm is under the testing stage of being implemented in our industrial partner's production line. This research acts as the underpinning work and contributes to a potential impact case for the Research Excellence Framework (REF) 2021.

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor, Dr. Xue Wang, for her invaluable guidance and endless patience during the four years of my study. You helped me not only grow as an academic throughout the journey but also as a person.

I am also very grateful for the continuous support that I received from Professor Jim E. Griffin. His insightful ideas are of great inspiration to me, and his patient teaching has helped me appreciate the subject of Bayesian statistic.

Moreover, I would like to thank Dr Alfred Kume, for the encouragement that he has given me throughout my studies. Thanks to Dr Cristiano Villa, Dr Fabrizio Leisen, Professor Jian Zhang and Dr Clare Dunning for their support and advice.

I would like to thank and acknowledge the School of Mathematics, Statistics and Actuarial Science at the University of Kent for funding my Ph.D. research.

Next, I would like to thank all my friends in Canterbury. A special thank you to Su Wang for always being there for me, especially when I was in difficult times. Your friendship is one of the most treasured things in my life. I also would like to thank Sam Cramer for all the happy moments we shared together. I am grateful to Bill Sakaria for taking care of me since the first day we met. Thanks to Andrea Cremaschi for your help and the tasty Italian food you made. Thanks to James Smith for all the fun and the English words you have taught me.

I also would like to thank Katherine Horan for her kindness. Thanks to Theodoros Gkolias for all the games we played. Thanks to Lucy Barnes and Brendan Nolan for making our times so enjoyable. Also, my thanks goes to

Georgios Papamikos, Mark Roelands and Gianluca Frasca Caccia for all the fun we had together. I also would like to thank Mingjie Hao, Tita Vanichbuncha and Elaheh Oftadeh for their friendship. Many thanks to Claire Carter for her help and constant encouragement, you are simply the best and the most irreplaceable person in the department.

Finally, I would like to thank to my parents Deju Kang and Wenling Xie, who have always supported and trusted me throughout all these years. No matter how far I am away, I know there is always a home for me with you.

Thank you to all the wonderful people that I have met during this journey, I would never be able to complete this thesis without you.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

A local engineering company manufacturing high precision sensors had accumulated huge historical databases of information on sensors which have been tested. The aim of this project is to use this historical data to reduce the number of measurements needed per sensor and maintain the required level of accuracy. In this thesis, we show how this can be achieved under a Bayesian framework, and introduce the new ideas for linear regression models which demonstrate how the reduction in individual sensor measurements can be achieved and how this procedure can be carried out much faster online.

Sensors are widely used in industrial systems and in our daily lives, such as monitoring ingredients in food during the production processes, controlling different chemical materials in power stations and recording the energy consumption in households. In this thesis, the sensors under study are mainly used to industrial control systems, where high precision is the primary requirement. During the manufacturing process, there are uncontrollable factors that could cause variations in the manufactured sensors. For example, these factor can be slight variations in the supplied material, changes in the environmental conditions during production, and human errors. These can lead to products with varying quality. Therefore, each sensor needs to be characterized or calibrated via parameter estimation to achieve a high precision before sending the sensors to the end users. The classical linear regression model has been adopted in current industry procedure, which requires a certain number of measurements per device to achieve the required level of accuracy.

Historical data provides the results of the characterisation process, in the form of parameter estimates. Eves et al. [2015] propose a Bayesian approach and use the historical data, which in Bayesian terminology is known as the prior information. This replaces information needed to be collected on the current sensor without sacrificing the accuracy of the estimate. The proposed method works well in certain cases. However, when we tried to apply this model to a different sensor dataset, the proposed Bayesian method failed completely as the parameter estimates exhibit the characteristics of multi-modality.

Recently, there has been a significant amount of research on Bayesian non-parametric regression models. This has primarily focused on developing models of the form

$$f(y|X) = \sum_{j=1}^{\infty} \omega_j(X) f(y|X, \theta_j(X)),$$

where y represents the response variable, X represents the covariate, $f(y|X, \theta_j(X))$ is a parametric density function, and $\omega_j(X)$ are the mixture weights that sum to 1 at every value of the covariate vector X. An intuitive approach to Bayesian nonparametric regression has been proposed by Müller et al. [1996]. However, we are interested in keeping the clear linear relationship of y and X, which has been demonstrated to work well by the models used at the company and should form the basis of our model. Specific proposals of Bayesian nonparametric mixture models are discussed in this thesis.

1.1 Outline of the Thesis

The aim of the project is to use the available historical information and advanced statistical models to reduce the number of measurements needed per sensor, at the same time maintain the high level of accuracy. Moreover, the proposed procedure for the characterisation is expected to be executed quickly.

In Chapter 2, the essential concepts of Bayesian nonparametrics are reviewed. We begin with the commonly used Dirichlet process (DP), and we give two useful DP alternative representations, which are the Pólya urn scheme and the stick-breaking process. We introduce the Dirichlet process mixture (DPM) model, which is constructed by assigning the DP as a mixing measure. The flexibility of the DPM model makes it suitable for density estimation problems. We introduce the Markov Chain Monte Carlo (MCMC) sampling techniques to compute the posterior distributions of the parameters. Two commonly used methods, the Metropolis-Hastings sampler and the Gibbs sampler are discussed. In the last section of Chapter 2, we provide details of the characterisation process in current use where this actual industrial problem comes from. In order to characterise the sensor to the required level of accuracy, a certain number of measurements are collected under different conditions. We describe the measurement collection process and demonstrate how we can achieve equivalent estimation accuracy while reducing the number of measurements.

In Chapter 3, we focus on this problem type of sensors, where the multi-modality is shown in the historical parameter estimates. In particular, we propose two approaches to deal with this problem. The first proposed approach is a Bayesian linear model with Dirichlet process mixture prior, where the Bayesian linear regression presents the relationship between the response variable and the covariates demonstrated to be appropriate by the company. In the mean time, the regression coefficients are modelled by a Dirichlet process mixture (DPM) model, which will provide adequate prior information to the linear regression model in order to compensate for the information loss, at the same time capturing the multi-modality shown in the data. The second proposed approach uses simple local linear regression models as building blocks to represent the relationships between the response variable and the covariates, where a data-driven prediction procedure is added to enhance the prediction accuracy for characterising the current sensor with reduced measurements. The slice sampler is applied for both approaches. The performances of the two approaches are assessed using the real industrial dataset.

In Chapter 4, we propose a Variational Bayes (VB) type of approach, as a fast alternative to MCMC, to estimate the mixture densities. In general, the VB framework proposes a family of variational distributions to approximate the exact posterior distributions instead of direct sampling from the true posterior distribution. Thus, VB uses less computational effort to provide deterministic approximations of the exact posterior distributions. In order to minimise the possibility of trapping into a local maximum density estimate, we propose a hybrid Variational Bayes (HVB) to be more likely to find a global maximum of the mixture density estimations. As a result, the unknown structure can be thoroughly described and the risk of becoming trapped in a local maximum is reduced. The advantages of applying HVB to estimate the mixture density is shown through both simulation study and the real dataset.

In Chapter 5, an optimal permutation sequential (OPS) algorithm is proposed in order to accelerate the processing speed of the model proposed in Chapter 3. In contrast to the *batch* learning algorithms such as MCMC approaches, where all observations are available at the same time, the sequential approach is often referred to as an *online* learning procedure, where each observation is proceeded as it arrives. A maximisation strategy is proposed to find an optimal permutation of data, which provides a way of dealing with the order-dependence of the sequential updating procedure. The proposed OPS algorithm needs only a single run and can be executed within a few seconds.

Finally, Chapter 6 shows a summary of the contributions made by this project and outlines a few topics for future work which will be investigated later.

Chapter 2

Literature Review

2.1 Dirichlet Process

As one of the most commonly used prior distributions in modern Bayesian statistics, the Dirichlet Process (DP) introduced by Ferguson [1973] plays a crucial role in the development of the Bayesian nonparametrics framework. He mentioned two desirable properties of the nonparametric prior distributions: one is that such a prior must have large enough support; the other one is that the posterior distribution given this prior should be in a closed form, which means that it should be conjugate. The introduced Dirichlet process opened a new era of Bayesian nonparametrics by fulfilling both desired properties.

In real world applications, the true data structure is always hidden to us. By fitting data in a parametric Bayesian model, we usually assume the prior distribution to follow a family of specific distributions with known density. For example, we can estimate the mean and variance parameters by fitting a univariate Normal distribution $\mathcal{N}(\mu, \sigma^2)$ to the data. However, the risk also exists of assuming the wrong or inappropriate distribution family. On the other hand, Bayesian nonparametric models do not require any assumptions of the prior distributions, which means the model structure is shaped completely by the data itself. Therefore, the risk of picking an inappropriate family of parametric priors can be reduced.

The main hurdle of building up a nonparametric prior is that the support of all possible prior distributions is too large to define a suitable topology directly. In the parametric case, suppose all Normal distributions are collected in a set $\mathscr{A}$, we know that every pair of mean and variance variables can uniquely determine a Normal distribution. Hence, every element of a $\mathcal{N}(\mu, \sigma)$ in $\mathscr{A}$ is mapped with its corresponding location specified by (μ, σ) in the 2-dimensional Euclidean space. Whereas in the nonparametric setting, $\mathscr{A}$ must include all probability measures, which are defined as continuous, right monotonic functions on the interval $(0, 1]$. In this section, we introduce the key concept called the Dirichlet process (DP) prior which has large enough support, and it is also a conjugate prior for the likelihood that is a multinomial distribution. A DP can be viewed as an infinite dimensional generalization of the Dirichlet distribution. First, we review some properties of the Dirichlet distribution which are needed for constructing a Dirichlet process.

2.1.1 Dirichlet Distribution

To construct a Dirichlet distribution, we start with a Gamma distribution denoted by $\mathcal{G}(y|\alpha, \beta)$ with a probability density function:

$$f(y|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta y} y^{\alpha - 1},$$

where $\alpha > 0$ represents the shape parameter, $\beta > 0$ represents the scale parameter and $\Gamma(\cdot)$ represents the gamma function, $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} \mathrm{d}t$. Let $y_1, y_2, \ldots, y_n$ be independent random variables generated by Gamma distribution with the same scale parameter $\beta = 1$, thus $y_i \sim \mathcal{G}(\alpha_i, 1)$, where $\alpha_i \geqslant 0$ for all $i = 1, \ldots, n$ and $\alpha_i > 0$ for some i.

The joint distribution of $p(y_1, y_2, \ldots, y_n)$ can be written as

$$\begin{aligned} p(y_1, y_2, \ldots, y_n) &= \prod_{i=1}^{n} p(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_i)} e^{-y_i} y_i^{\alpha_i - 1}. \end{aligned}$$

Then, we define the random variables x_i as,

$$x_i = \frac{y_i}{\sum_{i=1}^{n} y_i},$$

for $i = 1, \ldots, n$ and denote $\mathbf{x} = \{x_i\}_{i=1}^{n}$. Applying the change of variables formula to the jointing distribution, we obtain

$$p(\dot{y}, \mathbf{x}) = p\left(T(\dot{y}, \mathbf{x})\right) \ |J(T)|.$$

First we need to define $\dot{y} = \sum_{i=1}^{n} y_i$, hence $y_i = \dot{y} x_i$ for $i = 1, \ldots, n-1$ and $y_n = \dot{x}(1 - \sum_{i=1}^{n-1} x_i)$. The transformation T is given as

$$(y_1, \ldots, y_n) = T(\dot{y}, x_1, \ldots, x_{n-1}) = \left(\dot{y} x_1, \ldots, \dot{y} x_{n-1}, \dot{y}(1 - \sum_{i=1}^{n-1} x_i) \right).$$

The corresponding Jacobian matrix of this transformation is stated as

$$J(T) = \begin{pmatrix} x_1 & \dot{y} & 0 & 0 & \cdots & 0 \\ x_2 & 0 & \dot{y} & 0 & \cdots & 0 \\ x_3 & 0 & 0 & \dot{y} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ x_{n-1} & 0 & 0 & 0 & \cdots & \dot{y} \\ 1 - \sum_{i=1}^{n-1} x_i & -\dot{y} & -\dot{y} & -\dot{y} & \cdots & -\dot{y} \end{pmatrix},$$

with determinant $|J(T)| = \dot{y}^{n-1}$. Therefore, we can equally write the joint distribution as

$$\begin{aligned}
&p(\dot{y}, x_1, \ldots, x_{n-1})\\
=&p\left(\dot{y}x_1, \ldots, \dot{y}x_{n-1}, \dot{y}(1-\sum_{i=1}^{n-1} x_i)\right)\\
=&\left(\prod_{i=1}^{n-1} \frac{1}{\Gamma(\alpha_i)} e^{-\dot{y}x_i} (\dot{y}x_i)^{\alpha_i-1}\right)\left[\frac{1}{\Gamma(\alpha_n)}\left(\dot{y}(1-\sum_{i=1}^{n-1} x_i)\right)^{\alpha_n-1} e^{-\dot{y}(1-\sum_{i=1}^{n-1} x_i)}\right]\dot{y}^{n-1}\\
=&\frac{1}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\left(\prod_{i=1}^{n-1} x_i^{\alpha_i-1}\right)\left(1-\sum_{i=1}^{n-1} x_i\right)^{\alpha_n-1} \dot{y}^{(\sum_{i=1}^{n}\alpha_i)-1} e^{-\dot{y}}
\end{aligned}$$

The re-arranged expression of the joint distribution clearly shows that the part $\dot{y}^{(\sum_{i=1}^{n}\alpha_i)-1} e^{-\dot{y}}$ is the kernel of a Gamma distribution $\mathcal{G}(\sum_{i=1}^{n}\alpha_i, 1)$. Hence, we can derive the $(n-1)$-dimensional distribution of $(x_1, \ldots, x_{n-1})$ by integrating $\dot{y}$ out,

$$\begin{aligned}
p(x_1, \ldots, x_{n-1}|\alpha_1, \ldots, \alpha_n) =& \int_{\dot{y}} p(\dot{y}, x_1, \ldots, x_{n-1})\, \mathrm{d}\dot{y}\\
=&\frac{\Gamma(\sum_{i=1}^{n}\alpha_i)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\left(\prod_{i=1}^{n-1} x_i^{\alpha_i-1}\right)\left(1-\sum_{i=1}^{n-1} x_i\right)^{\alpha_n-1}.
\end{aligned}$$

Given parameters $(\alpha_1, \ldots, \alpha_n)$, this expression is known as the *Dirichlet distribution* denoted by $\mathcal{D}(\alpha_1, \ldots, \alpha_n)$. When $n = 2$, the Dirichlet distribution can be simplified to the Beta distribution. Hence, a Dirichlet distribution can be also treated as an infinite-dimensional extension of a Beta distribution.

One of the well-known property of Gamma distributions is the additive property. If $y_1 \sim \mathcal{G}(\alpha_1, 1)$ and $y_2 \sim \mathcal{G}(\alpha_2, 1)$ are independent, then $y_1 + y_2 \sim \mathcal{G}(\alpha_1 + \alpha_2, 1)$. It is trivial to show that Dirichlet distributions also have the *additive property*, which is stated as

If $(x_1, \ldots, x_n) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_n)$ *and* $r_1, \ldots, r_n$ *are integers such that* $0 < r_1 < r_2 < \cdots < r_l = n$*, then* $(\sum_1^{r_1} x_i, \ldots, \sum_{r_{l-1}+1}^{r_l} x_i) \sim \mathcal{D}(\sum_1^{r_1}\alpha_i, \ldots, \sum_{r_{l-1}+1}^{r_l}\alpha_i)$.

Since a Dirichlet distribution is a conjugate prior to a multinomial distribution, the posterior of a Dirichlet distribution can be derived as

Given prior distribution of $(x_1, \ldots, x_n) \sim \mathcal{D}(\alpha_1, \ldots, \alpha_n)$*, with observations* z *follows a multinomial distribution* $\mathbb{P}(z = j|x_1, \ldots, x_n) = x_i$ *for* $j = 1, \ldots, n$*. Then the posterior distribution of* $(x_1, \ldots, x_n)$ *given* $z = j$ *is* $\mathcal{D}(\alpha_1^{(j)}, \ldots, \alpha_n^{(j)})$*, where* $\alpha_i^{(j)} = \alpha_i$ *if* $i \neq j$ *and* $\alpha_i^{(j)} = \alpha_i + 1$ *if* $i = j$*.*

2.1.2 Dirichlet Process

As mentioned in the beginning of this chapter, an appropriate nonparametric prior distribution is desired to have support on the spaces of all probability distributions, denoted by $\mathscr{A}$. In this section, we introduce the Dirichlet process (DP) prior distributions which is commonly applied in Bayesian nonparametric. By assuming $\mathscr{A}$ to be separable, there exists small subsets A' which contain countable infinity number of elements. We can construct such a dense subset A' in $\mathscr{A}$, which means that for any element in $\mathscr{A}$ we can always find the element in A' or arbitrarily close to an element in A'. Therefore, the basis of A' can be the basis of $\mathscr{A}$, where every element in $\mathscr{A}$ can be linearly spanned by the basis in A'. Since the elements in A' is countable infinity, the summation of basis set in $\mathscr{A}$ is countable infinity and can be viewed as a infinite-dimensional linear Euclidean space. Then, we are allowed to use the definitions and operations on the Euclidean space.

Suppose a space can be partitioned into

$$\mathscr{A} = A_1 \cup A_2 \cup \cdots \cup A_k, \quad \text{and} \quad A_i \cap A_j = \emptyset, \text{ for } \; i \neq j.$$

We can calculate the probabilities of unknown distribution on each partition, then use this probability distribution to approximate the unknown distribution. In this sense, each possible distribution in $\mathscr{A}$ can be mapped to a k-dimensional vector.

$$G \to (G(A_1), G(A_2), \ldots, G(A_k)).$$

Since the probability of each partition $G(A_i)$ must be positive and the summation of them must be one. The Dirichlet distribution is an obvious choice for the prior distribution.

$$(G(A_1), G(A_2), \dots, G(A_k)) \sim \mathcal{D}(\alpha(A_1), \dots, \alpha(A_k))),$$

where $\alpha(\cdot)$ denotes the distribution which is specified on the corresponding partition.

In order to provide an arbitrary accurate approximation to any probability distribution in the space $\mathscr{A}$, we would like the number of partition k to tend to infinity. Moreover, for a given value of k, there are an infinite number of possible partitions and each partition has its corresponding distribution. Hence, the Dirichlet process is defined as the infinite-dimensional generalisation of the Dirichlet distribution, and it was firstly introduced by Ferguson [1973].

Definition 2.1. *Let α be a non-null finite measure on the Borel space $(\Theta, \mathscr{A})$. We say G is a Dirichlet Process on $(\Theta, \mathscr{A})$ with parameter α and base distribution G_0, written as $G \sim DP(\alpha, G_0)$, if for every $k = 1, 2, \dots,$ and measurable partition $(A_1, \dots, A_k)$ of Θ, the distribution of $(G(A_1), G(A_2), \dots, G(A_k))$ is $DP(\alpha G_0(A_1), \dots, \alpha G_0(A_k))$.*

The definition of DP provides a construction of each and every partition of $\mathscr{A}$, instead of directly defining a prior distribution on $\mathscr{A}$ with large support. A DP is usually described as a distribution over distributions. Ferguson [1973] made use of Kolmogorov's consistency theorem to show that a distribution over functions from the measurable subsets of Θ to $[0, 1]$ exists satisfying **Definition 2.1** for all finite measurable partitions of Θ.

Let $G \sim \mathrm{DP}(\alpha, G_0)$, since G itself is a randomly drawn distribution from DP, we can independently draw samples $\theta_1, \dots, \theta_n$ from G. Let $A_1, \dots, A_k$ be a finite measurable partition of probability space Θ, and let $n_j = \sum_{i=1}^{n} \delta_{\theta_i}(A_j)$ be the number of observed values in A_j for $j = 1, \dots, k$, where δ_x gives mass one to the point x. Based on the conjugacy between the Dirichlet and the multinomial distribution, we have

$$(G(A_1), \ldots, G(A_k)|\theta_1, \ldots, \theta_n) \sim \mathcal{D}(\alpha G_0(A_1) + n_1, \ldots, \alpha G_0(A_k) + n_k). \quad (2.1.1)$$

This shows that the concentration parameter α can be updated to $\alpha + n_j$ and the base distribution G_0 is updated to $(\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i}(A_j))/(\alpha + n)$ for each partition $\{A_j\}_{j=1}^{k}$, where δ is the Dirac delta measure,

$$\delta_\theta(A) = \begin{cases} 1, & \text{if } \theta \in A \\ 0, & \text{if } \theta \notin A. \end{cases}$$

Since the expression in (2.1.1) is true for all finite measurable partitions, the posterior distribution of G must be again a DP, and it can be written as

$$G|\theta_1, \ldots, \theta_n \ \sim \mathrm{DP}\left(\alpha + n,\ \frac{\alpha}{\alpha + n} G_0 + \frac{\sum_{i=1}^{n} \delta_{\theta_i}(A_j)}{\alpha + n}\right).$$

The probability distribution space $\mathscr{A}$ is partitioned by an infinite number of A_k as $k = 1, 2, \ldots$. Hence, as $k \to \infty$ δ_{A_i} we have

$$\mathrm{DP}(\theta) = p_1 \delta_{\theta_1}(\theta) + p_2 \delta_{\theta_2}(x) + \cdots + p_i \delta_{\theta_i}(x) + \cdots,$$

where θ_i represents a point randomly drawn from the sample space. For such an infinite-dimensional vector, the Dirichlet distribution is no longer valid, thus it is difficult to describe the distribution for $p_1, p_2, \ldots, p_i, \ldots$. Ferguson [1973] obtained the distribution of $p_1, p_2, \ldots, p_i, \ldots$, by constructing a Gamma process. Based on this representation, one of the important property of DP is that the probability distribution sampled from DP are discrete with probability one. The discreteness is important since it implies that the potential draws from DP can take the same values. During the sampling process, DP actually induces a clustering process. For continuous distributions, the probability of two samples sharing the same value is zero. Also we notice that DP has no limit for the base measure

G_0. However, choosing the base measure G_0 to be continuous, the DP can draw discrete samples almost surely.

2.1.3 Blackwell-MacQueen Urn Scheme

From the infinite sum construction of DP above, we conclude that samples drawn from DP are discrete almost surely. As the number of samples $n \to \infty$, the sample can be arbitrarily close to any continuous distribution. Following the notations used in previous sections, we suppose $G \sim \mathrm{DP}(\alpha, G_0)$, and i.i.d. sequences $\theta_1, \theta_2, \ldots$ are drawn from G. For a measurable $A \subset \Theta$, the predictive distribution for θ_{n+1} given previous n observations $\theta_1, \theta_2, \ldots, \theta_n$ is given by

$$P(\theta_{n+1} \in A | \theta_1, \ldots, \theta_n) = \frac{1}{\alpha + n} \left(\alpha G_0(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A) \right) \tag{2.1.2}$$

which is called a Pólya representation with parameter α. Notice that expression (2.1.2) is identical to the expectation $\mathbb{E}[G(A)|\theta_1, \ldots, \theta_n]$ of the posterior distribution $G(A)$ given the observations $\theta_1, \ldots, \theta_n$. Hence, the predictive distribution of θ_{n+1} is derived by integrating out G:

$$\theta_{n+1} | \theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n} \left(\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i} \right). \tag{2.1.3}$$

Therefore, the posterior base distribution given $\theta_1, \ldots, \theta_n$ is also shown to be the predictive distribution of θ_{n+1}.

Blackwell and MacQueen [1973] provided an alternative approach by exploiting the connection between DP and the generalized Pólya urn schemes. They visualised the conceptional DP by describing a ball-drawn process. For finite Θ, the sequence $\{\theta_i\}_{i=1}^{n}$ represents the results of successive draws from an urn where initially the urn has $\alpha(\theta)$ number of balls with colour θ. In each step of draw, a ball is randomly picked from the urn. Then, we identify the colour $\alpha(\theta_i)$, put the drawn ball back and add a ball with the same colour to the urn. Blackwell and MacQueen [1973] proved such a generalised Pólya urn scheme is equivalent to the

DP. Therefore the distribution of θ_{n+1} given $\{\theta_i\}_{i=1}^n$ can be described as duplicating any previous θ_i with probability $1/(\alpha + n - 1)$ and will obtain a fresh draw from G_0 with probability $\alpha/(\alpha+n-1)$. Since there exists balls which are expected to share the same colour θ_j^*, the conditional draw can be characterized by setting it equals to θ_j^* with probability $n_j/(\alpha + n - 1)$, where θ_j^* represents the distinct values found in $\{x_i\}_{i=1}^n$ with frequencies $n_j = \sum_{i=1}^n \delta_{\theta_j^*}(\theta_i)$ for $j = 1, \ldots, k$. Thus, the predictive distribution (2.1.3) can be expressed as

$$\theta_{n+1}|\theta_1, \ldots, \theta_n = \begin{cases} \delta_{\theta_{j^*}}, & \text{for } j = 1, \ldots, k \\ G_0, & \text{for } j = k+1, \end{cases} \tag{2.1.4}$$

where k is the number of distinct values.

The predictive probability function in Equation (2.1.3) implies a probability model of the joint density for $(k, \mathbf{n})$,

$$p(k, n_1, \ldots, n_k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^k \prod_{j=1}^{k} \Gamma(n_j),$$

where $\mathbf{n} = (n_1, \ldots, n_k)'$ represents the cluster sizes of n observations. According to Antoniak [1974], we can obtain the probability mass function for the number of component k_n,

$$p(k_n) = S_{n,k} n! \alpha^{k_n} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)},$$

where $S_{n,k}$ is the absolute value of the Stirling number of the first kind. Therefore, the expectation of k is given by

$$\mathbb{E}[k|\alpha, n] = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \approx \alpha \log\left(\frac{\alpha + n}{\alpha}\right),$$

which grows with n roughly at the rate of $\mathcal{O}(\log(n))$.

The validity of the Pólya urn scheme can be shown by using the proof of the existence of DP through the de Finetti's theorem. The theorem is applied to exchangeable observations, such as the samples $\theta_1, \ldots, \theta_n$ from the Pólya urn scheme.

Definition 2.2. (Exchangeability of random sequences). *A random process* $(\theta_1, \theta_2, \dots)$ *is called infinitely exchangeable if for any* $n \in \mathbb{N}$ *and any permutation* η *on* $1, \dots, n$*, the probability of generating* $(\theta_1, \dots, \theta_n)$ *is equal to the probability of drawing them in a different order* $(\theta_{\eta(1)}, \dots, \theta_{\eta(n)})$:

$$p(\theta_1, \dots, \theta_n) = p(\theta_{\eta(1)}, \dots, \theta_{\eta(n)}).$$

Therefore, we can construct a distribution over the sequence $\theta_1, \dots, \theta_n$ by iteratively drawing θ_i given $\boldsymbol{\theta}_{-i} = \{\theta_l\}_{\{l \neq i\}}$. For $n \geq 1$ the joint distribution over the first n observations can be written as

$$p(\theta_1, \dots, \theta_n) = \prod_{i=1}^{n} p(\theta_i | \{\theta_l\}_{\{l \neq i\}}),$$

where each conditional distributions $p(\theta_i | \cdots)$ follows the predictive distribution described by Equation (2.1.4). It is straightforward to verify that this random sequence is infinitely exchangeable.

If there are k colours, which implies the action of picking a new colour from G_0 occurred k times, and n_j balls have been drawn for each colour θ_j^*, $j = 1, \dots, k$, then, the statistics $\{n_j\}_{j=1}^k$ do not change with different permutations of $(\theta_1, \dots, \theta_n)$. The joint distribution $p(\theta_1, \dots, \theta_n)$ depends only on $\{n_j\}_{j=1}^k$ by

$$p(\theta_1, \dots, \theta_n) = \frac{\alpha^k \prod_{j=1}^k G_0(\theta_j^*)(n_j - 1)}{(\alpha + n - 1)(\alpha + n - 2) \cdots \alpha},$$

where $n = \sum_{j=1}^k n_j$. The de Finetti's theorem states that there exists a prior over the random distribution $P(G)$. This $P(G)$ matches exactly the Dirichlet process $\mathrm{DP}(\alpha, G_0)$.

This exchangeable property is turned out to play a crucial role in the development of Markov chain Monte Carlo (MCMC) algorithms for latent variables sampled from a Dirichlet process, which will be discussed in Section 2.2 and

Section 2.3. We emphasise the discreteness property here. For a long enough sequence of draws from G, the value of any draw will be repeated by another draw. It means that G is composed only of a weighted sum of point masses, therefore G is a discrete distribution.

2.1.4 Stick-breaking Process

There is another remarkable presentation of the $\mathrm{DP}(\alpha, G_0)$ known as the stick-breaking representation. Proposed by Sethuraman [1994], the constructive definition of the DP is given as

$$G = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j},$$

where

$$\begin{aligned} \omega_1 =& v_1, \\ \omega_j =& v_j \prod_{l<j} (1 - v_j) \quad \text{for} \quad j > 1, \\ v_j \sim& \mathrm{Beta}(v_j | 1, \alpha), \end{aligned} \tag{2.1.5}$$

and $\theta_j \sim G_0$ is independent of v_j.

The construction of the weights ω_j can be viewed as the result of breaking a stick of unit length randomly in infinite fragments. For $j = 1$, we first break the stick at $v_1 \sim \mathrm{Beta}(v_1|1, \alpha)$ and assign the mass v_1 to a random point $\theta_1 \sim G_0$. Then, the remaining mass $(1 - \nu_1)$ is split in a proportion $v_2 \sim \mathrm{Beta}(v_2|1, \alpha)$ and the mass $(1 - \nu_1)\nu_2$ is assigned to a random point $\theta_2 \sim G_0$. Keep breaking the stick until the assignment of the whole mass is completed by countably infinite number of points.

Sethuraman [1994] proves that the resulting process is actually a $\mathrm{DP}(\alpha, G_0)$. Such an infinite sequence of weights $\boldsymbol{\omega} = \{\omega_j\}_{j=1}^{\infty}$ is referred to be distributed according to a GEM process, $\boldsymbol{\omega} \sim \mathrm{GEM}(\alpha)$ with concentration parameter α, where the name stands for Griffiths, Engen and McCloskey. It is clear that the weights tend to zero as $j \to +\infty$. The concentration parameter α determines the distribution of the weights $\boldsymbol{\omega}$. For a small value of α only the first couple of segments have significant impacts, while the remaining sticks have little influence. On the other hand, the stick lengths will tend to be more uniform as the value of α gets large.

One of the main differences between the stick-breaking construction and the Pólya urn representation is that, for each group j, there is a proportion ω_j of points falling into group j. Therefore, instead of running the Pólya urn scheme to figure out these proportions, the stick-breaking process provides a way to sample them directly. Ishwaran and James [2001] extended the stick-breaking process to a more general setting by allowing $v_j \sim \text{Beta}(v_j|M_1, M_2)$ with constants M_1 and M_2. They showed that the sum of weights equals to 1 almost surely, when

$$\sum_{j=1}^{\infty} \log(1 + M_1/M_2) = \infty.$$

The stick-breaking representation of the DP reveals a significant way of truncating the infinite number of elements $\{\omega_j\}_{j=1}^{\infty}$ to a finite set. As mentioned earlier, the sampled weights are decreasing to zero as the number of breaks tends to infinity. As the process continues to a large number of breaks, the remaining length of the stick will be too short. Thus, the weights generated from that length are negligible. In this sense, the stick-breaking process allows us to approximate a Dirichlet process with an appropriate finite truncation. This process is especially useful when applying MCMC techniques on sampling from intractable posterior distributions.

2.1.5 Dirichlet Process Mixtures

For continuous density estimation problems, DP is not suitable since it assigns probability one to the space of discrete probability measures. Lo [1984] proposed a class of priors that chooses Dirichlet process as the mixing distribution. Related to the infinity sum representation of Dirichlet process, the continuity of density is translated to mixture models with a countably infinite number of components. This scenario is called a *Dirichlet process mixtures* (DPM), which can be treated as a nonparametrically mixture model with parametric components.

Suppose observations $\mathbf{x} = \{x_1, \ldots, x_n\}$ are generated from a mixture of kernel densities, the DPM model is stated as

$$\begin{aligned} x_i|\theta_i &\sim f(\cdot|\theta_i), \quad \text{for} \quad i = 1, \ldots, n \\ \theta_1, \ldots, \theta_n|G &\sim G \\ G|\alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0). \end{aligned} \tag{2.1.6}$$

Since the discreteness of the distribution G, the i.i.d. drawn latent parameter θ_i can share the same value. Observations x_i with the same value θ_i can be clustered in the same component.

This DPM model can also be constructed by a stick-breaking process mentioned in Section 2.1.4. Define variable $\gamma_i = j$ to indicate the cluster assignment for observation x_i to the j^{th} component for $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Then DPM model stated in (2.1.6) can be equally written as

$$\begin{aligned} \omega|\alpha &\sim \mathrm{GEM}(\alpha) \\ \gamma_i|\boldsymbol{\omega} &\sim \mathrm{Multinomial}(\boldsymbol{\omega}) \\ \theta_1, \ldots, \theta_k|G_0 &\sim G_0 \\ x_i|\gamma_i, \theta_1, \ldots, \theta_k &\sim F(\theta_{\gamma_i}), \quad \text{for} \quad i = 1, \ldots, n. \end{aligned}$$

In this model setting, the sampled distribution $G = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j}$, with ω_j representing the mixing proportion of components, θ_j are cluster-specific parameters, $f(x_i|\theta_j)$ the density of observations, and G_0 the prior distribution of the cluster-specific parameters.

The Dirichlet Process Mixture model is a mixture with a countably infinite number of clusters, which gives great flexibility to model fitting. From the definition of stick-breaking process, the weight parameter ω decreases exponentially quickly. Hence, only a relative small number of clusters will be involved in model fitting. In fact, the expected number of components is proportional to the logarithmic of the number of observations. Such a way of deciding the number of component in DPM is extremely useful in Bayesian model for density estimation. Since the number of component is completely inferred from the data, the risk of mis-specification present in parametric model is eliminated.

The DPM models have largely dominated the Bayesian nonparametric literature recently as a consequence of the realization that full posterior computation is feasible by using simulation methods. We apply the DPM models to capture the unknown structure of data in Chapter 3.

2.1.6 Mixture of Dirichlet Process

The widely applied DP prior has large enough support and closed form under posterior distribution. Meanwhile DP also has many other properties such as discreteness, tail-freeness and so on, which are proved to be very useful in Bayesian analysis. However, there are situations where the Dirichlet Process is inadequate. For example, given DP as prior distribution on the bio-assay problem mentioned in Antoniak [1974], the posterior distribution is not the Dirichlet Process but a mixture of Dirichlet processes (MDP) instead.

In practice, it is natural to propose a parametric family with unknown parameters. A mixture of Dirichlet processes is a Dirichlet process where the parameter α is itself treated as random variable having a certain distribution. Hence the MDP is a parametric mixture of nonparametric priors. The most important property is that the posterior distribution based on a MDP prior is again a MDP. The MDP also has similar properties to the DP such as samples from an MDP are discrete almost surely, and so on. On the other hand, the self-similarity and tail-freeness properties no longer hold for the MDP.

2.1.7 Hierarchical Dirichlet Processes

The Dirichlet process mixtures (DPM), as one of the most commonly applied nonparametric prior, can also be viewed as a simple version of the hierarchical Dirichlet processes (HDP) proposed by Teh et al. [2006]. Assume observations x_{ij} comes from unknown distributions F_i, and the distributions $F_1, \ldots, F_k$ are obtained by randomly sampled i.i.d. from a Dirichlet process whose centre measure G is itself again randomly sampled from a Dirichlet process. Since Dirichlet samples are discrete with probability one, the discreteness of G forces $F_1, \ldots, F_k$ to share their atoms. Hence dependence will be observed in the values of x across different components. This feature is often desirable in machine learning applications. For example, in multi-task settings each task might be associated with a probability measure with dependence across the tasks implemented using a hierarchical Bayesian model.

The adaptation to the nonparametric model with infinite-dimensional parameters was undertaken in Teh et al. [2006], where they discuss details of Dirichlet processes and indicate their extensions to other priors. In this model, data is distinguished into different group and each group is associated with a mixture model.

2.2 Markov chain Monte Carlo

In early sections, we have described a Bayesian nonparametric model with large support. From the Bayesian perspective, the inference about unknowns can be updated through Bayes' theorem by observing new evidence. Let x denotes the observed data, and θ denotes the model parameters. In general, the inference requires setting up a joint probability distribution $p(x, \theta) = \mathrm{L}(x|\theta) f_0(\theta)$, which includes a prior distribution $f_0(\theta)$ and a likelihood function $\mathrm{L}(x|\theta)$. Prior distributions represent the belief about the unknown parameters θ before observing any evidence, which could be based on subjective facts, previous experiences or personal opinions. Meanwhile, the information contained in the observations is reflected by the likelihood function. Therefore, Bayes' theorem provides a way of calibrating personal beliefs by updating evidence through probabilities

$$p(\theta|X) = \frac{\mathrm{L}(X|\theta)\pi_0(\theta)}{\int \mathrm{L}(X|\theta)\pi_0(\theta)\ \mathrm{d}\theta}.$$

Having observed x, Bayes' theorem is used to determine the distribution of θ conditionally on x, which is known as the *posterior* distribution of θ. Then, the expectation of the parameter θ can be derived as

$$\mathbb{E}[f(\theta)|x] = \frac{\int f(\theta)\mathrm{L}(x|\theta)f_0(\theta)\ \mathrm{d}\theta}{\int \mathrm{L}(x|\theta)f_0(\theta)\ \mathrm{d}\theta}.$$

The integrations involved in the expectation usually cause practical difficulties in Bayesian inference. In most applications, the analytic evaluation of $\mathbb{E}[f(\theta)|x)]$ is not available. Alternative approaches including numerical evaluation or analytic approximation can provide good results for low dimensional models. Beyond that, these methods can be very complex when dealing with high dimensional models. In this section, we introduce the Monte Carlo techniques to draw inference from models with large dimensions.

2.2.1 Monte Carlo Integration

To evaluate the integration of the posterior distribution in Bayesian inference, we introduce the Monte Carlo integration which evaluates $\mathbb{E}[f(\theta)]$ by drawing samples $\{\theta_i\}_{i=1}^n$ from posterior distribution $p(\theta|x)$, then approximating

$$\mathbb{E}\left[f(\theta)\right] \approx \hat{I}_1 = \frac{1}{n}\sum_{i=1}^{n} f(\theta_i).$$

The population mean of $f(\theta)$ is estimated by a sample mean, which is also called the simple *Monte Carlo* (MC) estimator. A simple MC estimator of this density is obtained by sampling θ_i from $f(\theta_i|\boldsymbol{\theta}_{-i})$ given $\{\boldsymbol{\theta}_{-l}\}_{\{l\neq i\}}$, and it is computationally expensive in general. Simple Monte Carlo methods must be extended by the use of draws from auxiliary (importance) densities. Notice that the estimator $\hat{I}_1$ is an unbiased estimators since $\mathbb{E}_q[\hat{I}_1] = I$. More importantly, the central limit theorems reveals that

$$\sqrt{n}\,\frac{\hat{I}_k - I}{\sigma_k} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as} \quad n \to \infty, \tag{2.2.1}$$

where σ_k denotes standard deviation of $\hat{I}_i$. Meanwhile, following the laws of large numbers these two estimators are strongly consistent estimators of I since

$$\hat{I}_k \xrightarrow{a.s.} I \quad \text{as} \quad n \to \infty,$$

for both $k = 1, 2$.

In general, drawing samples $\{\theta_i\}_{i=1}^n$ independently from $f(\cdot)$ is not feasible when $f(\cdot)$ appears to be non-standard. However, the $\{\theta_i\}_{i=1}^n$ do not necessarily need to be independent. Observations $\{x_i\}_{i=1}^n$ can be generated by any process which draws samples through the support of $f(\cdot)$ in the correct proportions. One way of doing so is through a Markov chain having $f(\cdot)$ as its stationary distribution. This is then the well-known *Markov chain Monte Carlo* (MCMC) approach.

2.2.2 Markov Chain Monte Carlo

A Markov chain is a discrete time stochastic process with the property that the distribution of θ_t given all previous values of the process, $\theta_0, \theta_1, \ldots, \theta_{t-1}$ only depends upon θ_{t-1}. For any set A, the *memoryless* property can be expressed as

$$\mathbb{P}\left(\theta_t \in A | \theta_0, \theta_1, \ldots, \theta_{t-1}\right) = \mathbb{P}\left(\theta_t \in A | \theta_{t-1}\right),$$

where $\mathbb{P}(\cdot|\cdot)$ denotes a conditional probability. For a discrete state-space, define the transition probabilities $\mathbb{P}_{ij}(t) = \mathbb{P}(\theta_t = j | \theta_0 = i)$ as the probability of moving from state i to state j in t steps.

The asymptotic behaviour of the Markov chain is one of our priory interests in the context of simulation when the number of iterations goes to infinity $n \to \infty$. A key concept called *stationary* distribution π is defined as

$$\sum_{\theta \in S} \pi(\theta) P(\theta|\psi) = \pi(\psi), \ \forall \psi \in S, \tag{2.2.2}$$

where S denotes the state space. Equation (2.2.2) can also be written in matrix notation as $\pi = \pi P$. The stationary distribution plays an important role in the convergence of the Markov chain. If the marginal distribution at any given step n is π, then the distribution at the next step is $\pi P = \pi$. Therefore, once the chain reaches a stage with stationary distribution π, the chain retains this distribution for all subsequent stages. This distribution is also known as the *limiting* distribution. If the stationary distribution π exists and $\lim_{n \to \infty} P^n(\theta, \psi) = \pi(\psi)$, then $\pi^{(n)}$ will approach π as $n \to \infty$ independently of the initial distribution of the chain.

For the distribution of θ_t to converge to a stationary distribution, the Markov chains are desired to satisfy three properties at least. First, it has to be *irreducible*, which says the Markov chain can reach any non-null set with positive probability from all possible starting points after some countable number of iterations. Second, the Markov chain needs to be *aperiodic*. We consider a Markov chains which contains only two states $\{0, 1\}$ and let the transient probability to be $P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. It is clear that the chain will just oscillate between two states as t tends to infinity, but will never converge. Such a chain is obviously irreducible, hence we have to prevent this phenomenon by constructing the chain to be *aperiodic*.

The last property of Markov chains is that the chain must be *positive recurrent*. When started at state i, the positive recurrent assures that the time of the first return to state i, denoted by $\tau_{ii} = \min\{t > 0 : \theta_t = i | \theta_0 = i\}$, is finite almost surely and the expectation of this τ_{ii} is finite. The last property can be expressed in terms of the existence of a stationary distribution $\pi(\cdot)$. If the initial value θ_0 is sampled from a stationary distribution $\pi(\cdot)$, then all subsequent iterates will also be distributed according to $\pi(\cdot)$. Here, mathematical definitions for these properties are presented as follows,

Definition 2.3.

(i). *X is called irreducible of for all i, j, there exists a $t > 0$ such that $P_{i,j} > 0$.*

(ii). *An irreducible chain X is recurrent if $\mathbb{P}(\tau_{ii} < \infty) = 1$ for some (and hence for all) i. Otherwise, X is transient. Another equivalent condition for recurrence is*

$$\sum_i P_{ij}(t) = \infty$$

for all i, j.

(iii). *An irreducible recurrent chain X is called positive recurrent if $\mathbb{E}[\tau_{ii}] < \infty$ for some (and hence for all) i. Otherwise, it is called null-recurrent. Another equivalent condition for positive recurrence is the existence of a stationary probability distribution for X, that is there exists $\pi(\cdot)$ such that*

$$\sum_i \pi(\theta_i) P_{ij}(t) = \pi(\theta_j)$$

for all j and $t > 0$.

(iv). *An irreducible recurrent chain X is called aperiodic if for some (and hence for all) i,*

$$\gcd\{t > 0 : P_{ii}(t) > 0\} = 1.$$

A state is called *ergodic* if it is aperiodic and positive recurrent. Then a chain is ergodic if all its states are ergodic. As we have mentioned earlier, the stationary distribution from an aperiodic positive recurrent Markov chain is also the *limiting* distribution of successive iterates from the chain. Then, we define the *ergodic averages* as

$$\bar{f}_N = \frac{\sum_{t=1}^{N} f(\theta_t)}{N},$$

which converges to their expectations under the stationary distribution.

Theorem 2.1. *If X is positive recurrent and aperiodic then its stationary distribution $\pi(\cdot)$ is the unique probability distribution satisfying Equation (2.2.2). If X is ergodic, the following consequences hold:*

(i). $\mathbb{P}_{ij}(t) \to \pi(\theta_j)$ *as* $t \to \infty$ *for all* i, j.

(ii). (**The Ergodic theorem**) *If* $\mathbb{E}_\pi[|f(\theta)|] < \infty$, *then*

$$\mathbb{P}\left(\bar{f}_N \to \mathbb{E}_\pi[f(\theta)]\right) = 1,$$

where $\mathbb{E}_\pi[f(\theta)] = \sum_i f(\theta_i)\pi(\theta_i)$, *the expectation of* $f(\theta)$ *with respect to* $\pi(\cdot)$.

The ergodic theorem is the Markov chain analogue to the strong law of large numbers (SLLN) in part (ii) of Theorem 2.1. It states that averaged chain values can also provide strongly consistent estimates of parameters from the limiting distribution π despite their dependence. For discrete state spaces, the ergodic averages are simply counting the relative frequency of x values in the realisations of the chain. Then, this relative frequency converges almost surely to the average frequency of visiting the state x. It allows one to ignore the dependence between draws of the Markov chain when we calculate quantities of interest from the draws. The central limit theorem mentioned in (2.2.1) for Markov chains also holds.

Therefore the MCMC is a class of methods in which we can simulate draws that are slight dependent from a posterior distribution. We then take those draws and calculate quantities of interest for the posterior distribution. In next sections, we introduce the two commonly used MCMC algorithms: the Metropolis-Hastings sampler and the Gibbs sampler.

2.2.3 Metropolis-Hastings Sampler

As the properties of MCMC are introduced, a natural step forward is to construct a Markov chain such that its stationary distribution $\pi(\cdot)$ is precisely our distribution of interest $f(\cdot)$. The name of Metropolis-Hastings stems from papers by Metropolis et al. [1953] and Hastings [1970]. The original paper by Metropolis et al. [1953] deals with the calculation of properties of chemical substances. Hastings [1970] made a significant contribution in terms of inference by providing the generalisation of the Metropolis samplers. As a result, the Metropolis-Hastings algorithm has become one of the most widely applied MCMC samplers.

At each time t in a Markov chain, the Metropolis-Hastings algorithm chooses the next state $\theta^{(t+1)}$ by first sampling a *candidate* point θ^* from a *proposal* distribution $q(\cdot|\theta^{(t)})$. The candidate point θ^* is then accepted with probability $\alpha(\theta^{(t)}, \theta^*)$ where

$$\alpha(\theta^{(t)}, \theta^*) = \min\left\{1, \; \frac{\pi(\theta^*)q(\theta^{(t)}|\theta^*)}{\pi(\theta^{(t)})q(\theta^*|\theta^{(t)})}\right\}.$$

Compare the value of α to an independent randomly generated quantity u from a Uniform distribution in a closed interval $[0, 1]$. If $u \leq \alpha$, the candidate point is accepted, the next state becomes $\theta^{(t+1)} = \theta^*$. Otherwise, if $u > \alpha$, the candidate is rejected, the chain does not move, then $\theta^{(t+1)} = \theta^{(t)}$.

Remarkably, the proposal distribution $q(\cdot)$ can have any form and the stationary distribution of the chain will be $\pi(\cdot)$. When the candidate is accepted $\theta^{(t+1)} = \theta^*$, the transition $P(\theta^{(t+1)}|\theta^{(t)})$ can be expressed as

$$P(\theta^{(t+1)} = \theta^*|\theta^{(t)}) = q(\theta^{(t+1)}|\theta^{(t)})\alpha(\theta^{(t)}, \theta^{(t+1)}).$$

A density $P(\cdot|\theta^{(t)})$ is defined for every possible value of the parameter different from $\theta^{(t)}$. Consequently, there is a positive probability left for the chain to remain at $\theta^{(t)}$ given by

$$P(\theta^{(t+1)} = \theta^{(t)}|\theta^{(t)}) = \mathbb{1}_{\{\theta^{(t+1)}=\theta^{(t)}\}}\left(1 - \int q(\theta^*|\theta^{(t)})\alpha(\theta^{(t)},\theta^*)\mathrm{d}\theta^*\right).$$

where $\mathbb{1}$ denotes the indicator function (taking the value 1 when its argument is true, and 0 otherwise). Thus, the transition kernel for the Metropolis-Hastings algorithm is

$$\begin{aligned}&P(\theta^{(t+1)}|\theta^{(t)})\\ &\quad=q(\theta^{(t+1)}|\theta^{(t)})\alpha(\theta^{(t)},\theta^{(t+1)}) + \mathbb{1}_{\{\theta^{(t+1)}=\theta^{(t)}\}}\left(1 - \int q(\theta^*|\theta^{(t)})\alpha(\theta^{(t)},\theta^*)\mathrm{d}\theta^*\right).\end{aligned}$$

Using the fact that

$$\pi(\theta^{(t)})q(\theta^*|\theta^{(t)})\alpha(\theta^{(t)},\theta^*) = \pi(\theta^*)q(\theta^{(t)}|\theta^*)\alpha(\theta^*,\theta^{(t)}),$$

the detailed balance equation is obtained as

$$\pi(\theta^{(t)})P(\theta^{(t+1)}|\theta^{(t)}) = \pi(\theta^{(t+1)})P(\theta^{(t)}|\theta^{(t+1)}). \tag{2.2.3}$$

By integrating both sides of Equation (2.2.3) with respect to $\theta^{(t)}$, we have

$$\int \pi(\theta^{(t)})p(\theta^{(t+1)}|\theta^{(t)})\mathrm{d}\theta^{(t)} = \pi(\theta^{(t+1)}). \tag{2.2.4}$$

The left-hand side gives the marginal distribution of $\theta^{(t+1)}$ under the assumption that $\theta^{(t)}$ is sampled from the stationary distribution $\pi(\cdot)$. Hence Equation (2.2.4) represents the fact that if $\theta^{(t)}$ is following the stationary distribution $\pi(\cdot)$, then $\theta^{(t+1)}$ will be as well, which means that once a sample from the stationary distribution has been obtained, all subsequent samples will be from that distribution. The Metropolis-Hastings algorithm is such that $P^{(t)}(\theta^{(t)}|\theta_0)$ converges to the stationary distribution, see Gilks et al. [1996] for further details.

2.2.4 Gibbs Sampler

In this section, we introduce one of the most commonly applied MCMC algorithm, called the Gibbs sampler, which is commonly recognised as the first significant impact of MCMC made in statistics. The original work is presented by Geman and Geman [1984], which deals with sampling the posterior of interest from a Gibbs distribution in the context of image processing. Their sampling scheme explored the conditional structure implied by the local specification. Even though it was a well known and influential paper in the area, their work was introduced to the main statistical community via the paper written by Gelfand and Smith [1990]. Nowadays, most statistical applications of MCMC have used the Gibbs sampling.

The Gibbs sampler is a MCMC scheme where the transition kernel is formed by the full conditional distributions. Assume as before the posterior distribution of interest is $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^{n}$. Let $\boldsymbol{\theta}_{-i}$ comprises all of $\boldsymbol{\theta}$ except θ_i, hence $\boldsymbol{\theta}_{-i} = \{\theta_l\}_{\{l \neq i\}}$. Define the full conditional distribution $\pi(\theta_i|\boldsymbol{\theta}_{-i})$ as the distribution of the i^{th} component of θ conditioning on all the remaining components,

$$\pi(\theta_i|\boldsymbol{\theta}_{-i}) = \frac{\pi(\boldsymbol{\theta})}{\int_{\theta_i} \pi(\boldsymbol{\theta})\mathrm{d}\theta_i},$$

where $\boldsymbol{\theta}$ has distribution $\pi(\cdot)$. Consider the full conditional distributions are available, which means that they are completely known and can be sampled from. Gibbs sampling provides a scheme based on successive generations from the full conditional distributions as follows.

---

1. Initialise the iteration counter of the chain $j = 1$ and set initial values $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)})'$;

2. Obtain a new value $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_n^{(j)})'$ from $\theta^{(j-1)}$ through successive gen-

eration of values

$$
\begin{aligned}
\theta_1^{(j)} &\sim \pi(\theta_1 |\, \theta_2^{(j-1)}, \ldots, \theta_n^{(j-1)}) \\
\theta_2^{(j)} &\sim \pi(\theta_2 |\, \theta_1^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_n^{(j-1)}) \\
&\vdots \\
\theta_n^{(j)} &\sim \pi(\theta_n |\, \theta_1^{(j-1)}, \theta_3^{(j-1)}, \ldots, \theta_{n-1}^{(j-1)})
\end{aligned}
$$

3. Change counter j to $j+1$ and return to step 2 until convergence is reached.

When convergence is reached, the resulting value $\theta^{(j)}$ is a draw from π. As the number of iteration increases, the chain approaches its equilibrium condition. Convergence is then assumed to hold approximately.

2.3 MCMC for Dirichlet Process Mixture of Normals

As introduced in previous Section 3.1 and Section 3.2, a construction of Bayesian nonparametric prior called the Dirichlet process mixtures (DPM) is illustrated and two fundamental MCMC strategies, Metropolis-Hastings and Gibbs samplers, are described. In this section, we introduce several MCMC algorithms for Bayesian inference on DPM of Normal distributions.

Recall the Dirichlet process mixtures model introduced by Lo [1984] in Section 2.1.5, we are interested in a special case when Normal density is used for the Dirichlet base measure.

$$
\begin{aligned}
x_i|\theta_i &\sim \mathcal{N}(x_i|\theta_i), \quad \text{for} \quad i = 1, \ldots, n \\
\theta_1, \ldots, \theta_n|G &\sim G \\
G|\alpha, G_0 &\sim \text{DP}(\alpha, G_0).
\end{aligned}
\tag{2.3.1}
$$

The first level of hierarchy represents the observed data x_i following a continuous distribution with Normal density $\mathcal{N}(x_i|\theta_i)$ with component-specific parameters θ_i. At the second hierarchical level, θ_i are independently drawn from some random distribution G. Such an unknown distribution G is generated from the Dirichlet process prior with precision parameter $\alpha > 0$ and base measure G_0. The precision parameter α is often assumed to be fixed and it plays a crucial role. The larger the value of α, the less the generated distribution G will concentrate on the base G_0.

The first attempt to tackle the intractable Dirichlet process mixture of Normal distributions by applying MCMC methodology is proposed by Escobar [1994]. He formalised his ideas as part of his PhD thesis and subsequent published on the development of a Bayesian nonparametric estimate of a vector of Normal means. In the next year, Escobar and West [1995] applied the direct extension of the method on density estimation using Normal mixtures of Dirichlet processes. In a more general framework, they provided a computational method that allows evaluation of posterior distributions of all parameters and the predictive distributions.

The Normal mixture model is one of the commonly applied parametric models in Bayesian density estimation. Following the DPM model, we can extend a finite Normal mixture models to an infinite mixture model. Suppose the data $\mathbf{x} = \{x_i\}_{i=1}^n$ are conditionally independent and normally distributed $x_i|\theta_i \sim \mathcal{N}(x_i|\theta_i)$, with mean and variance parameters $\theta_i = \{(\mu_i, \sigma_i^2)\}_{i=1}^n$. In particular, suppose that $G \sim \mathrm{DP}(\alpha, G_0)$ and G_0 is a specified bivariate distribution function over $\mathbb{R} \times \mathbb{R}^+$. Then G_0 is the prior expectation of G such that $\mathbb{E}[G(\theta)] = G_0(\theta)$.

The major problem of a hierarchical model is to move from prior to posterior through the randomly generated distribution G from the Dirichlet process. This is not considered to be easy as G has an infinite number of atoms. Escobar and West [1995] solved this issue by involving the Pólya urn representation. As introduced by Blackwell and MacQueen [1973], the Pólya urn scheme circumvents the problem of generating G from DP and provides an expression with G integrated out. Then, we can draw θ based on its conditional distribution iteratively as follows,

$$\begin{aligned}
&\text{for} \quad i = 1, \quad \theta_1 \sim G_0, \\
&\text{for} \quad i = 2, \quad \theta_2 | \theta_1 \sim \frac{\alpha G_0 + \delta_{\theta_1}}{\alpha + 1}, \\
&\ \vdots \\
&\text{for} \quad i = n, \quad \theta_n | \theta_1, \dots, \theta_{n-1} \sim \frac{\alpha G_0 + \sum_{i=1}^{n-1} \delta_{\theta_i}}{\alpha + n - 1},
\end{aligned}$$

where $\delta_\theta(\cdot)$ is the distribution concentrated at location θ. Hence, the general form of these successive conditional distributions is given by

$$\theta_i | \theta_1, \dots \theta_{i-1} \sim \frac{\alpha}{\alpha + i - 1} G_0(\theta_i) + \frac{1}{\alpha + i - 1} \sum_{j=1,\ j \neq i} \delta_{\theta_j}(\theta_i). \tag{2.3.2}$$

The predictive distribution of θ_{n+1} conditional to $\theta_1, \dots, \theta_n$ is given in Equation (2.1.6) with similar expression.

According to Equation (2.3.2), we notice that both G_0 and $\sum_{j=1}^{i-1} \delta_{\theta_j}(\cdot)$ have the same form. Thus, sampling from them will be straightforward. By simulating a Markov chain which has this posterior as its stationary distribution, we can sample from the posterior distribution of $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^n$. Conditioned on exchangeable random variables $\{\theta_i\}_{i=1}^n$, the full conditional distribution for θ_i for each i can be derived as follows

$$\begin{aligned}
p(\theta_i | \boldsymbol{\theta}_{-i}, \mathbf{x}) \sim&\ \ q_0 G_i(\theta_i) + \sum_{j=1,\ j \neq i}^{n} q_j \delta_{\theta_j}(\theta_i) \\
&\propto \mathrm{K}(y_i|\theta_i) \left\{ \alpha f_0(\theta_i) + \sum_{j \neq i} \delta_{\theta_j}(\theta_i) \right\},
\end{aligned}$$

where $\boldsymbol{\theta}_{-i} = \{\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_n\}$, $f_0(\cdot)$ denotes the prior density and $\mathrm{K}(\cdot)$ represents the kernel function. The full conditional is proportional to

$$p(\theta_i | \boldsymbol{\theta}_{-i}, \mathbf{x}) \propto \alpha \int f_0(\theta_i) \mathrm{K}(x_i|\theta_i) \mathrm{d}\theta_i \left(\frac{f_0(\theta_i)\mathrm{K}(x_i|\theta_i)}{\int f_0(\theta_i)\mathrm{K}(x_i|\theta_i)\mathrm{d}\theta_i} \right) + \sum_{j \neq i} \mathrm{K}(x_i|\theta_i) \delta_{\theta_j}(\theta_i). \tag{2.3.3}$$

The normalizing constant is given as

$$c = \alpha \int f_0(\theta_i)\mathrm{K}(x_i|\theta_i)\mathrm{d}\theta_i + \sum_{j \neq i} \mathrm{K}(x_i|\theta_j).$$

The first part on the right hand side of the equation (2.3.3) is the density of a new drawn θ_i with observation x_i assigned into it, thus

$$f_i(\theta_i) = \left(\frac{f_0(\theta_i)\mathrm{K}(x_i|\theta_i)}{\int f_0(\theta_i)\mathrm{K}(x_i|\theta_i)\mathrm{d}\theta_i} \right).$$

The second term is a summation of point masses at the previously drawn components. Therefore, given other $\boldsymbol{\theta}_{-i}$, the θ_i can be drawn from

$$\theta_i \begin{cases} = \theta_j, & \text{with probability } \frac{\mathrm{K}(x_i|\theta_j)}{c} \\ \sim g_0(\theta_i), & \text{with probability } \frac{\alpha \int f_0(\theta_i)\mathrm{K}(x_i|\theta_i)\mathrm{d}\theta_i}{c}. \end{cases}$$

The most direct approach of sampling from model (2.3.1) is to repeatedly draw values for each θ_i from its conditional distribution conditioned on the data $\mathbf{x}$ and $\boldsymbol{\theta}_{-i}$. The conditional distribution is obtained by combining the prior distribution with the likelihood for θ_i that results from x_i having a Normal distribution $K(x_i|\theta_i)$. One benefit of the Normal DPM model is that if $f_0(\cdot)$ and $\mathrm{K}(\cdot)$ are conjugate, the posterior distribution of θ is available in closed form. Therefore, given the sampled values of $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_n^{(t)})$ at the t^{th} iteration, the predictive distribution for a new observation x_{n+1} can be estimated as

$$p(x_{n+1}|x_1, \ldots, x_n) \approx \frac{1}{T} \sum_{t=1}^{T} \mathrm{K}(x_{n+1}|\theta_{n+1}^{(t)}) \quad \text{for} \quad t = 1, \ldots, T,$$

where T is the total number of iterations and $\theta_i^{(t)}$ is drawn from

$$\theta_i^{(t)} \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{n + \alpha} \sum_{i=1}^{n} \delta_{\theta_j^{(t)}}.$$

Escobar and West [1995] proposed an algorithm which repeatedly draws new samples from $\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{x}$ as defined by Equation (2.3.3), for $i = 1, \ldots, n$. Such a Gibbs sampler can be extended by involving hyper-parameters and convergence properties. However, this sampler is not very efficient since the convergence to the posterior distribution is rather slow. The clustering property of DP implies there exists positive probabilities that observations are associated with the same component parameter denoted by θ_j^*. The proposed algorithm by Escobar and West [1995] can not change the allocation of θ for more than one observation simultaneously, and the frequency of such a change is rare. This is due to that the required moving through a low probability intermediate state in which observations in the group do not have the same θ value.

According to the discreteness of $G(\cdot)$ under the Dirichlet process, there exist k distinct values among the n elements $\boldsymbol{\theta}$ as $\theta_j^* = (\mu_j^*, \sigma_j^{2*})$. Then, the conditional posterior distribution in Equation (2.3.2) can be equally written as

$$\theta_i|\theta_1, \ldots \theta_{i-1} \sim \frac{\alpha}{\alpha + i - 1} G_0(\theta_i) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{k} n_j \delta_{\theta_j^*}(\theta_i). \tag{2.3.4}$$

In this way, the efficiency can be improved by taking a finite number of mixture models. Therefore, an additional hierarchy level is required to identifying the clusters classes in the Bayesian hierarchical structure mentioned in the model (2.3.1).

Marginal Conjugate Algorithms with Latent Class.

One of the algorithms that deals with such hierarchical model was introduced by MacEachern [1994]. Similar to the model (2.3.1), MacEachern [1994] constructs a hierarchical model by inserting a latent class associating with observations under Normal conjugacy,

$$
\begin{aligned}
x_i|\gamma_i, \boldsymbol{\theta} &\sim \mathcal{N}(x_i|\theta_{\gamma_i}), \quad i = 1, \ldots, n \\
\theta_{\gamma_1}, \ldots, \theta_{\gamma_n} &\sim G \\
G &\sim \mathrm{DP}(\alpha, G_0) \\
\gamma_1, \ldots, \gamma_n|\boldsymbol{\omega} &\sim \mathrm{Discrete}(\omega_1, \ldots, \omega_k) \\
\boldsymbol{\omega} &\sim \mathrm{DP}(\alpha/k, \ldots, \alpha/k).
\end{aligned}
\tag{2.3.5}
$$

For each latent class in the middle hierarchy level γ_i, the parameter θ_{γ_i} determines the distribution of observations from that class. The collection of all such distinct θ_{γ_i} is denoted by $\boldsymbol{\theta}$. The mixing proportions for the classes, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k)$, are given a symmetric Dirichlet prior with concentration parameter α/k. As k goes to infinity, the concentration parameter approaches zero.

By integrating over the mixing proportions $\boldsymbol{\omega}$, the prior for γ_i can be expressed as the product of conditional probabilities as,

$$
\begin{aligned}
P(\gamma_i = j|\gamma_1, \ldots, \gamma_{i-1}) &= \frac{P(\gamma_1, \ldots, \gamma_{i-1}, \gamma_i = j)}{P(\gamma_1, \ldots, \gamma_{i-1})} \\
&= \frac{\int \omega_{\gamma_1} \cdots \omega_{\gamma_{i-1}} \omega_{\gamma_i} \Gamma(\alpha)\Gamma(\alpha/k)^{-k} \omega_1^{(\alpha/k)-1} \cdots \omega_k^{(\alpha/k)-1} \mathrm{d}\boldsymbol{\omega}}{\int \omega_{\gamma_1} \cdots \omega_{\gamma_{i-1}} \Gamma(\alpha)\Gamma(\alpha/k)^{-k} \omega_1^{(\alpha/k)-1} \cdots \omega_k^{(\alpha/k)-1} \mathrm{d}\boldsymbol{\omega}} \\
&= \frac{n_j + \alpha/k}{i - 1 + \alpha}
\end{aligned}
\tag{2.3.6}
$$

where n_j is the number of $\gamma_i = j$ for $j = 1, \ldots, k$. Notice that as k goes to infinity, the conditional probabilities in Equation (2.3.6) reach the following limits,

$$
\begin{aligned}
\mathbb{P}(\gamma_i = j|\gamma_1, \ldots, \gamma_{i-1}) &\to \frac{n_j}{i - 1 + \alpha} \\
\mathbb{P}(\gamma_i = k + 1 \text{ for } j = 1, \ldots, k|\gamma_1, \ldots, \gamma_{i-1}) &\to \frac{\alpha}{i - 1 + \alpha},
\end{aligned}
\tag{2.3.7}
$$

see Neal [2000] for more details. Notice that when k tends to infinity, the model (2.3.5) is equivalent to the Dirichlet process mixture model in (2.3.1) due to the correspondence between the conditional probabilities of θ_i in Equation (2.3.4) and those implied by Equations (2.3.7).

The infinite number of $\boldsymbol{\theta}^* = \{\theta_j^*\}_{j=1}^k$ cannot be explicitly represented when k goes to infinity. Instead, we only represent those θ_j^* that are currently associated with some observations. The Gibbs sampling for allocation γ_i is based on the following conditional probabilities

$$\begin{aligned} P(\gamma_i = j|\gamma_{i-1}, x_i, \boldsymbol{\theta}^*) &= c\frac{n_j}{n-1+\alpha} f(x_i|\theta_j^*) \\ P(\gamma_i = k+1 \text{ for } j = 1, \ldots, k|\gamma_{i-1}, x_i, \boldsymbol{\theta}^*) &= c\frac{\alpha}{n-1+\alpha}\int f(x_i|\theta)\mathrm{d}f_0(\theta), \end{aligned} \tag{2.3.8}$$

where c is the normalising constant that makes the above probabilities sum to one. In a Normal conjugate context, we can eliminate θ_j^* by integrating it out. The state of the Markov chain then consists only by γ_i, which can be updated by Gibbs sampling using the following conditional probabilities:

$$\theta_i|\boldsymbol{\theta}_{-i}, \gamma_{-i}, x_i \sim c\frac{\alpha}{\alpha+n-1}\int f(x_i|\theta)\, \mathrm{d}f_0(\theta) + c\frac{n_j}{\alpha+n-1}\int f(x_i|\theta_j^*)\mathrm{d}f_{-i,\gamma_i=j}(\theta_j^*)$$

where $f_{-i,\gamma_i=j}(\theta)$ denotes the posterior distribution of θ based on the prior G_0 and all the observations except x_i and $\gamma_i = j$. Eventually, the algorithm proposed by MacEachern [1994] repeatly samples a new value from $\gamma_i|\boldsymbol{\gamma}_{-i}, x_i$ for $i = 1, \ldots, n$.

According to Equations (2.3.8), the numerical value of the γ_i is arbitrary. The γ_i determines what has been called the configuration in which the data items are grouped in accordance with shared values for θ_j^*. The value of γ_i may therefore be chosen to facilitate the display of mixture components in some desired order. When Gibbs sampling for γ_i chooses a value not equal to any other $gamma_j$, a value for θ_{γ_i} is chosen from G_0, the posterior distribution based on the prior G_0 and the single observation x_i. The algorithm proposed by Bush and MacEachern [1996] is stated as follows.

Let the state of the Markov chain consist of $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_n\}$ and $\boldsymbol{\theta} = (\theta_\gamma : \gamma \in$

$\{\gamma_1, \ldots, \gamma_n\})$. Repeatedly sample as follows:

• For $i = 1, \ldots, n$: If the present value of γ_i is associated with no other observation (i.e., $n_{-i,\gamma_i} = 0$), remove θ_{di} from the state. Draw a new value for γ_i from $\gamma_i | \boldsymbol{\gamma}_{-i}, x_i, \boldsymbol{\theta}$ as defined by Equation (2.3.8). If the new γ_i is not associated with any other observation, draw a value for θ_{γ_i} from G_0 and add it to the state.

• For all $\gamma_i \in \{\gamma_1, \ldots, \gamma_n\}$: Draw a new value from $\theta_{\gamma_i} | x_1, \ldots, x_n$ for which $\gamma_i = j$ that is, from the posterior distribution based on the prior G_0 and all the data points currently associated with latent class $\gamma_i = j$.

As was the case for the first Gibbs sampling method, this approach is feasible if we can compute $\int f(y_i, \theta) \mathrm{d}G_0(\theta)$ and sample from G_0, as will generally be the case when G_0 is the conjugate prior.

Slice sampler

For a DPM model, the sampling issue is the countably infinitness of the discrete masses from the random distribution functions generated from the DP prior. Walker [2007] proposed a slice sampling scheme to deal with the infiniteness. An extra latent variable U is introduced to truncate the infinite number of mixing components to a finite number. Moreover, all the conditional distributions can be sampled directed given conjugate priors.

As we have introduced in Section 2.1.2, a random probability measure G can be generated from a $\mathrm{DP}(\alpha, G_0)$. The infinite expression of G is then given by

$$G = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j}, \tag{2.3.9}$$

where $\{\omega_j\}_{j=1}^{\infty}$ are mixing weights, δ_θ denotes the measure with a point mass of 1 at θ. The weights are generated from the stick-breaking process introduced in Section 2.1.4,

$$\omega_1 = v_1 \quad \text{and} \quad \omega_j = v_j \prod_{l<j} (1 - v_j) \quad \text{for} \quad j > 1.$$

Meanwhile, we assume $\{\theta_j\}_{i=1}^{\infty}$ are independent and identically distributed from the base distribution G_0. Consider the random density function

$$f_G(y) = \int \mathcal{N}(y|\theta)\mathrm{d}G(\theta),$$

where $\mathcal{N}(y|\theta)$ denotes the Normal distribution with parameters $\theta = (\mu, \sigma^2)$. Using the form for G in Equation (2.3.9), we have

$$f_{\omega,\theta}(y) = \sum_{j=1}^{\infty} \omega_j \mathcal{N}(y|\theta_j).$$

The proposed slice sampler attempts to estimate the model by introducing a latent variable U, such that the joint density of (y, u) is given by

$$f_{\omega,\theta}(y, u) = \sum_{j=1}^{\infty} \mathbb{1}(u < \omega_j)\mathcal{N}(y|\theta_j). \tag{2.3.10}$$

It is obvious that the desired density $y_{\omega,\theta}(y)$ can be returned by integrating over u. Equivalently, Equation (2.3.10) can be written as

$$f_{\omega,\theta}(y, u) = \sum_{j=1}^{\infty} \omega_j \mathrm{Unif}(u|0, \omega_j)\mathcal{N}(y|\theta_j),$$

where $\mathrm{Unif}(\cdot)$ denotes the Uniform distribution. Then, the marginal density for u is given by

$$f_{\omega}(u) = \sum_{j=1}^{\infty} \omega_j \mathrm{Unif}(u|0, \omega_j).$$

Therefore, we are able to define a finite collection, such as,

$$A_{\omega}(u) = \{j : \omega_j > u\}.$$

Hence, the joint density of (y, u) in Equation (2.3.10) can also be stated as

$$f_{\omega,\theta}(y, u) = \sum_{j \in A_{\omega}(u)} \mathcal{N}(y|\theta_j).$$

The conditional density of y given u is given by

$$f_{\omega,\theta}(y|u) = \frac{1}{f_\omega(u)} \sum_{j \in A_\omega(u)} \mathcal{N}(y|\theta_j),$$

where $f_\omega(u)$ denotes the marginal density for u, and it is defined on $(0, \max(\omega_j)]$.

At this stage, the infinite mixture model in Equation (2.3.9) is truncated to a finite mixture density. Next, an indicator latent variable $\boldsymbol{\gamma}$ can be introduced to identify the component of the mixture from which y is to be taken, as mentioned earlier in Section 2.3. Therefore, the full likelihood function of observations $\{y_i\}_{i=1}^n$ can be expressed as

$$l_{\omega,\theta}(\{y_i, u_i, \gamma_i = j\}_{i=1}^n) = \prod_{i=1}^n \mathcal{N}(y_i|\theta_{\gamma_i}) \mathbb{1}(u_i < \omega_{\gamma_i}).$$

Given conjugated priors of parameters v_j and θ_j, we can follow a Gibbs sampling scheme to estimate the mixture densities. Further details are given in Section 3.3.

2.4 Industrial Application

In Chapter 1 we mentioned that sensors play an important role for a wide range of scientific experiments and industrial applications. The accuracy of output variables from sensing devices is one of the primary concerns. Due to inevitable issues taking place on the production line, no sensor can be guaranteed to be identical to others. Thus, the sensor's output variable could be deviated from the true value more or less. To minimise the inaccuracy caused by the manufacturing process, individual characterisation is crucial for each sensor.

In general, characterisation is implemented by comparing sensor's readings with the true values under different conditions. A sensor generates outputs of the liquid flowing through, each output is computed as a combination of the collected several elementary physical signals with pre-determined weights. Meanwhile, the corresponding true value are considered as known under a specified environmental condition. Therefore, the aim of characterisation is to minimise the deviation between the sensor's outputs and the true values by finding an optimal set of weights. To have a better understanding of the proposed statistical models and related algorithms in later chapters, we introduce the background of our considered industrial application and fundamental principles behind the sensor technology.

2.4.1 Background

The resonator sensor under study is an in-line device that measures the density of flowing liquid. One common way of measuring the density of liquid ρ is to collect liquid mass m and corresponding volume V, and the density is given by $\rho = m/V$. However, the sensors in the considered situation are used to measure densities of the flowing liquids, which can be determined by the following equation,

$$\rho = \frac{C_1^2}{f^2} \times TC(\Delta P, \Delta T) - C_2, \tag{2.4.1}$$

where f denotes the resonance frequency. Constants C_1 and C_2 can be determined at a known condition when pressure changes $\Delta P = 0$ and temperature changes $\Delta T = 0$. The term $TC(\Delta P, \Delta T)$ is called the *temperature compensation* (TC) dependent on ΔP and ΔT.

Based on the resonance phenomenon, the pipe inside the sensor will vibrate when there is liquid passing through. The frequency of resonance can be detected and collected by a transducer installed inside of the sensor. Also, the experimental conditions ΔP and ΔT are varied during the characterisation process.

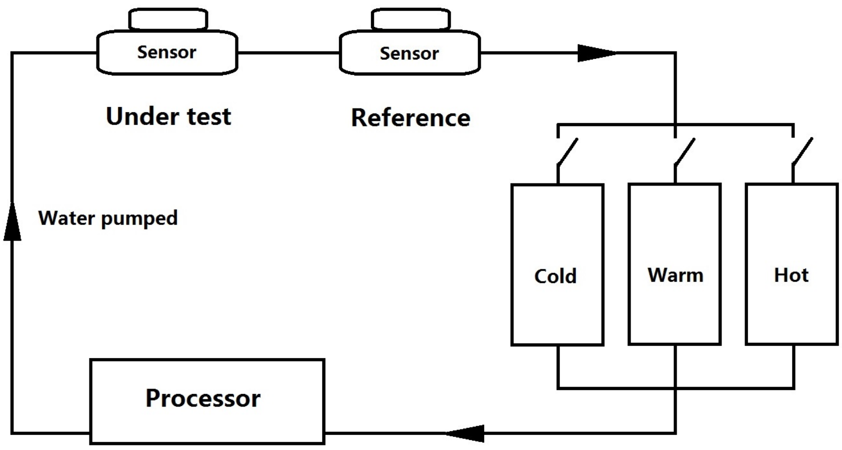


Figure 2.1: Sensor calibration rig.

According to Figure 2.1, when water is pumped through the sensor under test, demanded physical factors $\{f, \Delta P, \Delta T\}$ are detected inside of the sensor. In order to measure liquid density ρ, the value of $TC(\Delta P, \Delta T)$ needs to be estimated by varying the sensing conditions for a fluid with well-known density characteristics. Given a specified condition with ΔP and ΔT, the temperature compensations can be determined through the following linear model,

$$TC(\Delta P, \Delta T) = \beta_1 + \beta_2 \Delta P + \beta_3 \Delta T + \beta_4 \Delta P \Delta T + \beta_5 (\Delta T)^2, \tag{2.4.2}$$

where $\beta = \{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}^T$ is a 5-dimensional critical parameter. The reasons of using such a polynomial regression model to determine the temperature compensations are given by Bayliss and Downes [2003].

The aim for characterisation is to find an optimal value of β, which obviously plays a critical role for the accuracy of liquid density estimates. Therefore, we need to find the true values of temperature compensation under certain environmental conditions specified by ΔP and ΔT. During characterisation process, water was used due to its availability and its well-studied density characteristics. Hence, the value of temperature compensation can be determined by rearranging Equation (2.4.1)

$$TC(\Delta P, \Delta T) = \frac{\rho_w + C_1}{C_2^2} \times f^2, \tag{2.4.3}$$

where ρ_w represents the true density of water, which is known as a constant under a certain environmental condition. Therefore, with collected elementary physical variables $\{f, \Delta P, \Delta T\}$ and true values of $TC(\Delta P, \Delta T)$, a unique value of β can be estimated by using regression techniques.

To make sure that the characterised sensor can consistently provide accurate estimates of liquid densities under various conditions, different combinations of ΔP and ΔT are set during the characterisation process. Here, we briefly illustrate the data collection procedure to elaborate the characterisation process. Water is pumped through a sensor with three different temperature levels: low, medium and high. For each temperature level, measurements are recorded under three different pressure levels: low, medium and high. Hence, there are totally 9 combinations of pressure and temperature levels. During the characterisation process, the temperature of water follows a sequence: low temperature, followed by medium, then high temperature, followed by medium and back to low temperature. Within each temperature bands, measurements are collected twice under each pressure levels for repeatability. The measured temperature pattern is illustrated in Figure 2.2.

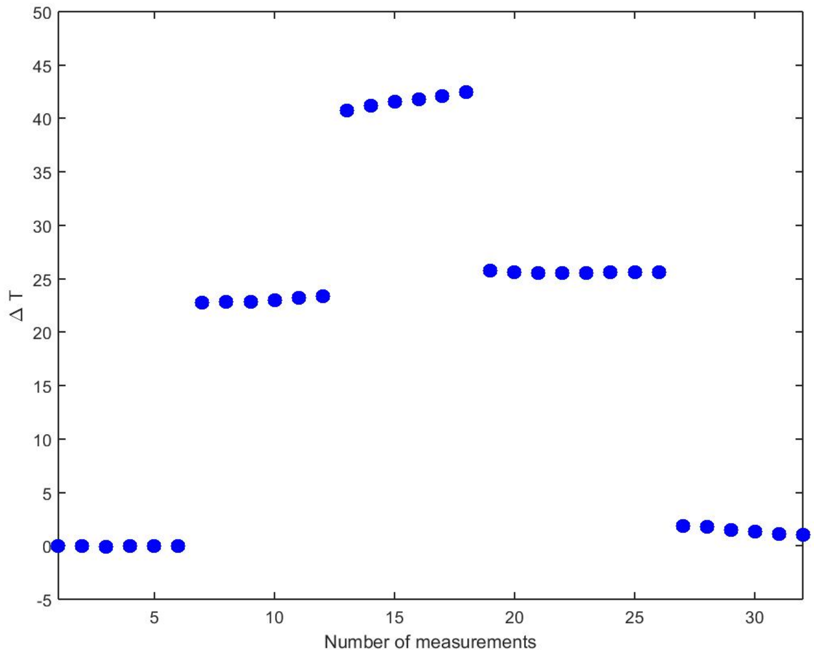


Figure 2.2: Data collection procedure with the temperature pattern: low, medium, high, medium and low.

The x-axis represents the number of measurements including 12 measurements from cold temperatures, 14 measurements from warm temperatures and 6 measurements from hot temperatures. Thus, totally 32 measurements are collected during characterisation. On the y-axis, the values of ΔT are determined by taking the differences between the temperatures of water passing through the sensor and the recorded room temperature.

Let p denotes the number of collected measurements and $y = (y_1, y_2, \ldots, y_p)^T$ denotes the values of temperature compensation. By using Equation (2.4.3), y can be determined by frequencies $f = (f_1, f_2, \ldots, f_p)^T$ and constants C_1 and C_2. Meanwhile, a $p \times 5$ matrix $X = (X_1, X_2, \ldots, X_5)$ is formed by vectors $X_j = (x_{j,1}, \ldots, x_{j,p})^T$ for $j = 1, \ldots, 5$. Vector X_1 contains regression intercepts, X_2, X_3, X_4 and X_5 are pressure changes ΔP, temperature changes ΔT, interactions $\Delta P \Delta T$ and temperature changes squared $(\Delta T)^2$, respectively.

The value of β is estimated by the ordinary least squares (OLS) method, which is easy to implement and fast to execute. The linear relationship described in Equation (2.4.2) can be simplified to

$$y = X\beta + \varepsilon, \tag{2.4.4}$$

where ε represents a $p \times 1$ error vector. Then, the OLS estimate of β is achieved by minimising the sum of error squared,

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Then, the characterisation errors can be calculated by

$$\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta}.$$

A sensor is considered to be well characterised only if the magnitude of all errors $\hat{\varepsilon}$ lie within a prescribed range. Company's quality specifications are various depending on different types of sensors.

The total time of a complete characterisation process takes 7 hours including installing and removing sensors from the testing facility. In order to collect measurements of ΔT, we have to wait approximately 2 hours for the whole characterisation system reaching a stable temperature status. The actual shape and the size of the sensor also affect the characterisation time.

2.4.2 Bayesian Linear Approach

A Bayesian linear regression is proposed by Eves et al. [2015] to reduce the number of measurements required. The historical information is used to replace parts of the measurements needed to be collected for the current sensor without affecting the accuracy of liquid density estimation. Using the accumulated dataset which contains a large number of previously characterised products, the characterisation time could be remarkably reduced by taking a fewer number of measurements per sensor.

The measurements at mid-range temperatures are targeted to be removed from the current sensor characterisation. Similar to the linear relationship stated in Equation (2.4.4), the values of temperature compensation y^- given reduced measurements X^- follows a Normal distribution,

$$f(y^-|X^-, \beta) \sim \mathcal{N}(X^-\beta, \sigma^2 \mathbb{I}_{p^-}), \tag{2.4.5}$$

where σ^2 is fixed as a extreme small constant (of the order $\times 10^{-7}$) and $\mathbb{I}_{p^-}$ represents a $p^- \times p^-$ identity matrix. Rather than assigning a Gamma prior on σ^2, it is reasonable to fix the value of σ^2, as it is estimated with very high precision.

A key step of the proposed Bayesian approach is to extract information of the critical parameter β from the historical dataset of the same type of sensors. By treating OLS estimates $\{\hat{\beta}_i\}_{i=1}^n$ of n previously characterised sensors as observations, we can constructed a multivariate Normal distribution,

$$\beta \sim \mathcal{N}\left(\mu_0, \Sigma_0\right), \tag{2.4.6}$$

where μ_0 and Σ_0 are the mean and the variance-covariance matrix with $\mu_0 = n^{-1}\sum_{i=1}^{n}\hat{\beta}_i$ and $\Sigma_0 = \mathrm{diag}(\sigma_j^2)$, and $\sigma_j^2 = (n-1)^{-1}\sum_{i=1}^{n}(\hat{\beta}_{ij} - \mu_{0j})^2$ for each dimension $j = 1, \ldots, 5$ of each observation $i = 1, \ldots, n$.

Then, using the estimated Normal distribution as prior, we can draw the posterior inference of β for the current sensor with reduced measurements through a linear regression model. The posterior of β given $(y^-, X^-, \sigma^2, \mu_0, \Sigma_0)$ follows a 5-dimensional multivariate Normal distribution of parameter β,

$$\beta \sim \mathcal{N}\left(\widetilde{\mu},\ \widetilde{\Sigma}_\beta\right),$$

where

$$\widetilde{\mu} = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X^{-T}X^-\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}X^{-T}y^-\right), \widetilde{\Sigma}_\beta = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X^{-T}X^-\right)^{-1}.$$

The posterior inference of β is basically derived by combining prior distribution in Equation (2.4.6) constructed from historical dataset with full measurements and likelihood function in Equation (2.4.5) from the current sensor with reduced measurements.

As demonstrated by Eves et al. [2015], the proposed Bayesian linear approach using reduced measurements can successfully characterise current sensors by incorporating historical information with full measurements from the same type. Since measurements at mid range temperatures are no longer required, the characterisation time using the proposed Bayesian approach can be reduced from 7 hours to 5 hours.

As we have mentioned earlier, with 7 hours as a typical time for processing a single sensor by the OLS method, there are only two runs of calibration can be completed per day. For example, an operator starts at 6 AM to start the test, and the test cycle will finish at 1 PM. Thus, only two batches of sensors can be tested per day. The project we considered in this thesis is aimed at increasing the manufacturing capacity, by reducing the calibration time per sensor. Hence, the costs will also be significantly reduced. Since the goal of the proposed Bayesian method is to eliminate one temperature level, the calibration time can be potentially reduced from 7 hours to 5 hours. Therefore, it can allow three test runs per day. In theory, the Bayesian approach increases the manufacturing capacity by 50%.

Chapter 3

Bayesian Linear Model with Dirichlet Process Mixture Prior

3.1 Introduction

Earlier in Section 2.4, we described the importance of the characterisation process for sensors to provide reliable outputs. The main task of characterisation can be viewed as estimating the critical parameter β using measurements collected during characterisation. The goal of this chapter is to estimate β by using a reduced number of measurements; particularly when multi-modality is discovered from the historical estimates of β. Following the idea introduced in Eves et al. [2015], the information stored in previously tested sensors can be used to replace information needed to be gathered on the current sensor without sacrificing the accuracy of the estimate.

It is worth noting that the historical estimates of β are obtained via OLS method using full measurements, and the company has accumulated a huge historical dataset of information on this type of sensors. We can model the historical information and use it as the prior information to draw the posterior inference of β in order to reduce the number of measurements needed per sensor. The proposed Bayesian method was a success in a certain way, where the historical estimates of β are assumed to follow a Normal distribution. However, when this method is applied to a different type of sensor dataset, the proposed Bayesian method completely failed as the historical estimates of β exhibit a multi-model structure.

In this chapter, we focus on the type of sensors where a multi-modal structure is displayed in the historical parameter estimates. Two approaches are proposed to characterise this type of sensors. The first one is a Bayesian linear model with Dirichlet process mixture prior (BL-DPMP), where the Bayesian linear regression relationship is used to describe the response variable and the covariates, and the regression coefficient β is modelled by a Dirichlet process mixture (DPM) model. Following the BL-DPMP approach, the reduction in individual sensor measurements can be achieved in two stages. In the first stage, the unknown structure of historical information is modelled by a hierarchical nonparametric model. Specifically, we assume the historical estimates of β follow a Dirichlet process mixture (DPM) of Normal distributions, instead of a single Normal distribution. Using the historical estimates of β as the data, the multi-modal structure of the historical information can be captured by a mixture model with a sufficient and finite number of Normal distributions.

In the second stage of the BL-DPMP approach, one Normal component from the constructed mixture model is assigned to be the prior distribution for the current sensor with reduced measurements. The Bayesian posterior inference of β for the current sensor can then be derived by combining the prior information that has been derived from the historical dataset with full measurements, and the likelihood function with reduced measurements. Following the proposed BL-DPMP procedure, we can characterise the problem type of sensors with reduced measurements to meet the required high precision. Based on the numerical evidence, the proposed BL-DPMP approach is verified by achieving a satisfactory pass rate.

The second proposed approach is a DPM regression model with data-driven prediction (DPMR-DDP) procedure, which uses a DPM regression model to describe the relationships between the response variable and the covariates from the historical dataset, and which improves the prediction accuracy following a data-driven prediction procedure. The DPM regression approach was first proposed by Müller et al. [1996], which modelled the joint distribution of response variables y and covariates X as a DP mixture of multivariate Normals to obtain the conditional density of y given X. However, researchers noticed that, in some cases, this model can lead to poor curve fitting and poor prediction. Recent developments on the topic can be found in Shahbaba and Neal [2009] and Wade et al. [2014]. The poor prediction issue becomes crucial when dealing with high-precision sensors in the considered situation. It is shown that predictions drawn from the standard DPM regression model are not accurate enough to meet the company's quality specifications. In addition, we propose a data-driven prediction (DDP) procedure on the DPM regression model in order to enhance the accuracy of prediction when characterising the current sensor with reduced measurements. The slice sampling algorithm introduced by Kalli et al. [2011] is applied to estimate the nonparametric mixture density for both proposed approaches.

The chapter is organised as follows: Section 3.2 provides a preliminary study on the type of sensors with multi-modality displayed from the historical estimates of the critical parameter β. Section 3.3 describes the details of both the proposed BL-DPMP and DPMR-DDP approaches mentioned above, which will be used to reduce the number of measurements required for the current sensor characterisation. The results obtained by applying the proposed approaches to the problem sensor dataset are presented in Section 3.4, along with the results obtained from other methods. Section 3.5 discusses the drawbacks of the proposed approaches and suggests directions for further research.

3.2 Motivation

As mentioned in Section 2.4, the aim of our project is to use the historical information and advance statistical models to reduce the number of measurements required for the current sensor characterisation. This is equivalent to estimate the critical parameter β of the current sensor using a reduced number of measurements and to embed the estimated value of β in the current sensor, so that the sensor output can achieve the required high level of accuracy. Eves et al. [2015] proposed a Bayesian way to model the historical estimates of critical parameter β by a Normal distribution, and use the historical information as the prior distribution to estimate the posterior distribution of β for the current sensor. However, the proposed method completely failed to a different type of sensors where the historical parameter estimates exhibit a multi-modal structure.

For the sensor considered in this project, the company has accumulated huge historical databases of information on the S1 type of sensors. There are $N = 895$ previously characterised S1 sensors with full measurements $\{(y_i, X_i)\}_{i=1}^{895}$ included in the S1 dataset, where each y_i is a 32×1 response variable and each X_i is a 32×5 covariate variable.

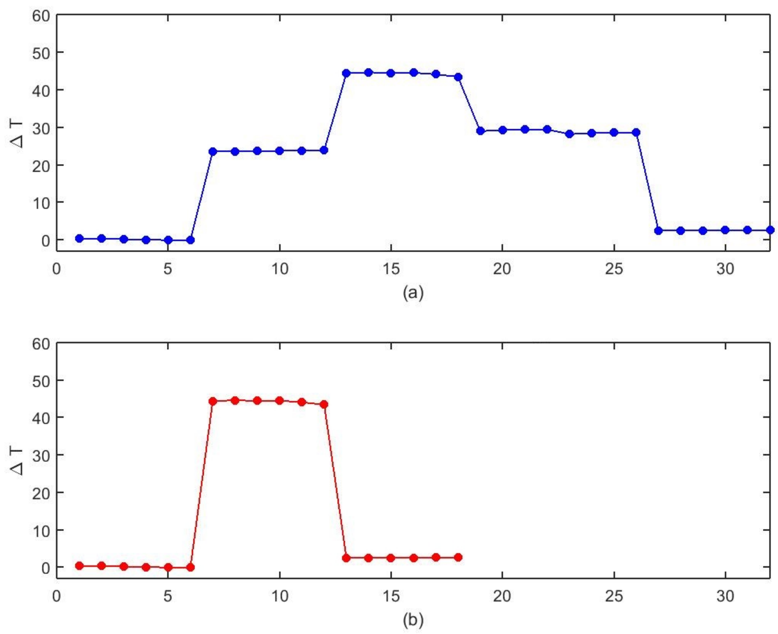


Figure 3.1: Top plot (a) is the full test procedure including 32 measurements for OLS; bottom plot (b) is the test procedure without measurements at the mid-range temperatures including 18 measurements for the proposed Bayesian method.

It is worth noting that the x-axis in both plots represents the number of measurements which needs to be collected during characterisation process. As illustrated in Figure 3.1(a), measurements are recorded at low, medium and high temperatures. At each range of temperatures, measurements are recorded twice under 3 different pressure levels: low, medium and high. For the S1 type of sensors, two extra measurements are taken at the second warm temperatures. Hence, there are 32 measurements in total that need to be collected to characterise the S1 type of sensors.

For each sensor i, a linear relationship between y_i and X_i is demonstrated to be appropriate by the company, which can be expressed as

$$y_i = X_i\beta_i + \varepsilon_i,$$

where ε_i is a 32×1 error vector for $i = 1, \ldots, 895$. Thus, the estimates $\{\beta_i\}_{i=1}^{895}$ of the regression coefficient β can be obtained via the OLS method, where each $\beta_i = (\beta_{i,1}, \ldots, \beta_{i,5})^T$ is a 5-dimensional vector. This 32 number of measurements can be treated as the minimum requirement for OLS estimates of β to achieve the required high level of accuracy. On the other hand, the proposed Bayesian method removes the measurements under the mid-range temperatures from the three temperatures currently used. It uses only 18 measurements to characterise the current S1 sensors as shown in Figure 3.1(b).

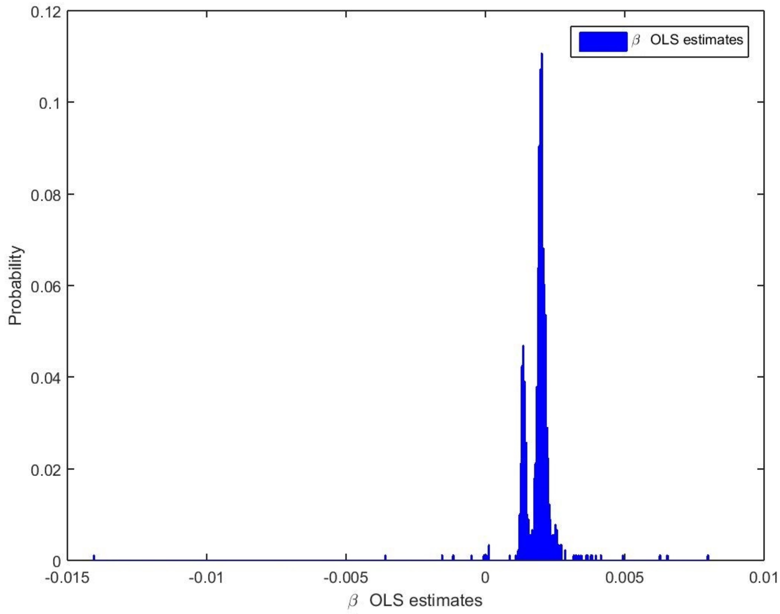


Figure 3.2: Histogram of the historical estimates $\{\beta_{i,2}\}_{i=1}^{895}$ from the S1 type of sensors.

In order to reveal the multi-modal structure of the regression coefficient β, we focus on the second dimension of β estimates $\{\beta_{i,2}\}_{i=1}^{895}$ and present the histogram plot in Figure 3.2. In Figure 3.2, there are two important features shown on the histogram plot. First, the OLS estimates $\{\beta_{i,2}\}_{i=1}^{895}$ are well spread and extreme values can be observed on both sides of the main clustered area. Hence, fitting a single Normal distribution leads to a flat bell-shaped curve as the curve needs to cover all values of $\{\beta_{i,2}\}_{i=1}^{895}$. As a result, the variance-covariance matrix of the fitted Normal distribution becomes too large, whereas most of β estimates are actually concentrated in the central area. In this case, a Normal distribution will not be able to correctly capture the multi-modal structure of parameter β. Similar phenomena are discovered on other dimensions of β estimates.

The second feature becomes more noticeable after increasing the number of bins of the histogram plot upto 100 as presented in Figure 3.3. Then, we can observe a clear multi-modal structure in the histogram plot. The black curve represents a single Normal distribution, which is specified by the sample mean and the covariance matrix. As shown by the black dotted vertical line, the sample mean is located in a region with low likelihood. Thus, modelling the parameter estimates of β into a Normal distribution as the prior distribution can potentially mislead the posterior estimate of β. Such an inaccurate model is unable to provide sufficient prior information to accurately estimate the posterior distribution of β for the current sensor. As a result, the current sensor with reduced measurements would be recognised as a failed product. However, this sensor would be recognised as a qualified product if it is characterised by the current OLS method with full measurements. The red curve in Figure 3.3 apparently has a better fitting for the multi-modality of β estimates. In this chapter, we propose a nonparametric mixture model to capture the unknown structure of β and use the estimated mixture model as prior information to compensate the information needed to be gathered on the current sensor.

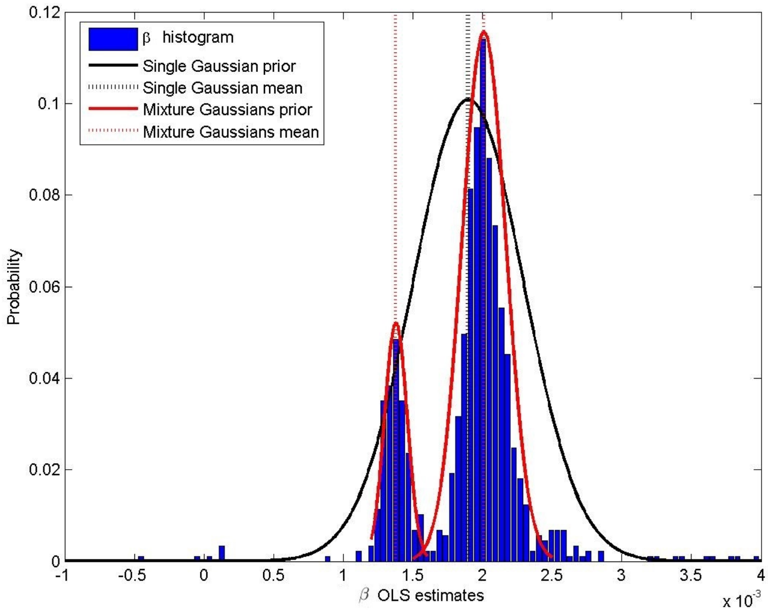


Figure 3.3: Histogram of the historical estimates $\{\beta_{i,2}\}_{i=1}^{895}$ with curve fitting by a Normal distribution (black) and a mixture of Normal distributions (red).

In order to demonstrate the proposed Bayesian approaches in this chapter, a historical dataset is formed to contain the first n number of β estimates $\{\beta_i\}_{i=1}^{n}$, which are obtained via OLS method with full measurements $\{(y_i, X_i)\}_{i=1}^{n}$. We estimate the mixture density of β from the historical dataset, and use the estimated mixture density as the prior information to estimate the posterior distribution of

β for the current sensor with reduced measurements. Therefore, the size of the historical dataset needs to be carefully chosen to be able to reflect the structure of β. To determine the size of the historical dataset, we apply the Kullback-Leibler divergence (KLD) to measure the distance between a pair of probability distributions $f_{1:i}$ and $g_{1:i+1}$, which are specified by the first i and $i+1$ estimates of β for i=1, . . . , 895. The KLD for each pair can be calculated as,

$$\mathrm{KLD}(f_{1:i} \mid\mid g_{1:i+1}) = \int f(y) \log \left(\frac{f(y)}{g(y)} \right) \mathrm{d}y.$$

By increasing the value of i, we can use the increment of KLD to judge how much extra information is carried forward by the newly included estimate β_i. The accumulated KLD increments for each value of i are plotted in Figure 3.4.

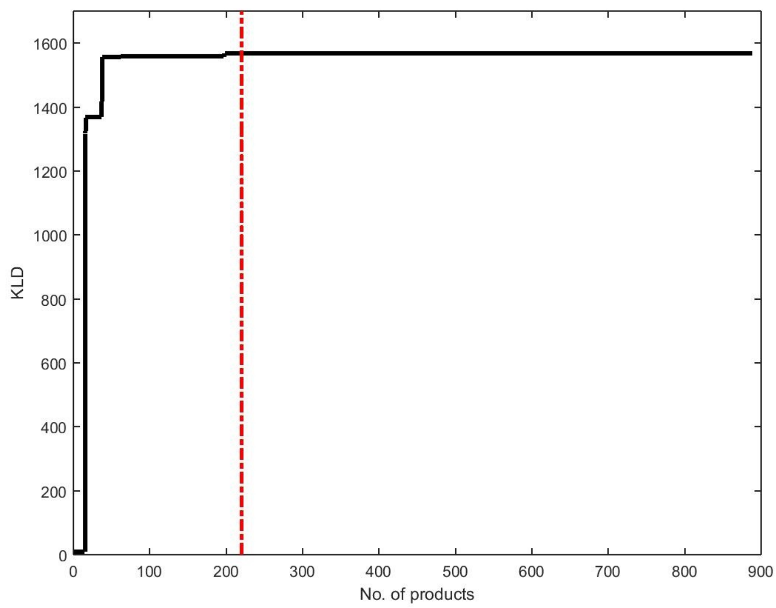


Figure 3.4: Cumulative Kullback Leibler divergence obtained from the historical estimates of $\{\beta_i\}_{i=1}^{895}$ *for the S1 type of sensors.*

The black solid line represents the cumulated KLD values, which increases as more estimates β_i are included in f and g. On the right hand side of the red vertical dashed line, there is hardly any increment on the KLD value can be observed. Hence, the first 240 OLS estimates of β are chosen to form the historical dataset. Meanwhile, we notice that the first 240 sensors includes some β estimates with little increment. In theory, those estimates with little KLD increment could be moved out of the historical dataset. In this chapter, we keep the arriving order of the first 240 β estimates since every sensor is characterised immediately once it was manufactured.

In Figure 3.5, we present the histogram plot of the first 240 β OLS estimates. As the black curve shows, modelling the historical estimates $\{\beta_i\}_{i=1}^{240}$ by a single Normal distribution is considered to be inappropriate for this type of sensors. The mixture density of $\{\beta_i\}_{i=1}^{240}$ will be estimated by the Bayesian approaches introduced in the Section 3.3.

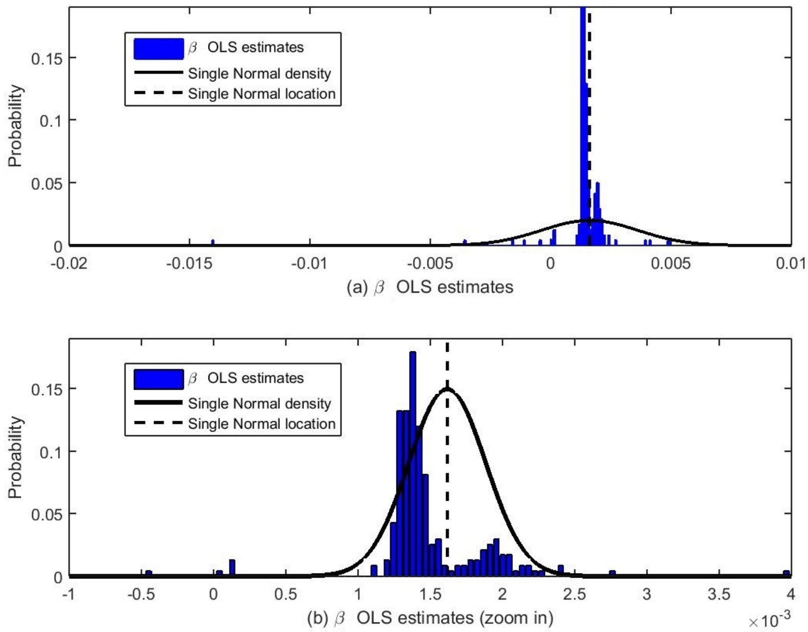


Figure 3.5: Top plot (a) shows the first 240 $\{\beta_{i,2}\}_{i=1}^{240}$ OLS estimates are fitted by one Normal distribution, with location represented by a dash line; bottom plot (b) zooms in to the most clustered range; fitting a Normal distribution as prior still produces misleading informations for both mean and variance-covariance.

The 895 S1 sensors data with full measurements $\{(y_i, X_i)\}_{i=1}^{895}$ are separated into three parts. The first part of data contains the first $n = 240$ sensors data with full measurements $\{(y_i, X_i)\}_{i=1}^{240}$, which is used to calculate the OLS estimates $\{\beta_i\}_{i=1}^{240}$ and to construct the mixture density of β.

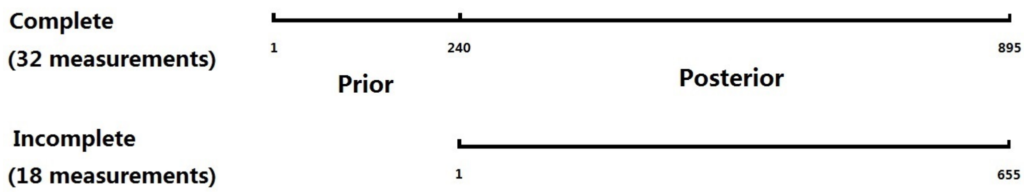


Figure 3.6: Historical sample size with full measurements and the posterior sample size with full and reduced measurements of the S1 type of sensors.

The second part of data contains the rest 655 sensors information with reduced 18 measurements $\{(y_i^-, X_i^-)\}_{i=241}^{895}$, which need to be characterised via the proposed Bayesian approaches. Then, the estimated mixture density from the historical dataset is used as the prior information to derive the posterior inference for the sensors in the second part of data with reduced measurements.

In order to validate the proposed Bayesian method, we need to assess the accuracy of the estimated β values for sensors with reduced measurements. Therefore, we form the third part of data which contains the rest 655 sensors information with full measurements. We embed the obtained β estimate via the proposed Bayesian approach to the current sensor, and recognise the current sensor as a qualified product if the estimated response variable achieve the required high level of precision given the full measurements as the covariate variable. Further details will be given in Section 3.4.

3.3 Methodology

3.3.1 Bayesian Linear Model with Dirichlet Process Mixture Prior

The main purpose of the sensor characterisation can be simplified by estimating the value of the critical parameter β through minimising the difference between the estimated response variable and the true value. As shown in Section 3.2, a multi-modal structure of β is discovered from the historical dataset of the S1 type of sensors. In this section, we propose a Bayesian linear model with Dirichlet process mixture prior (BL-DPMP), which uses the historical information to reduce the number of measurements needed per sensor without sacrificing accuracy of the β estimate.

In industry, sensor characterisation relies on a linear model

$$y = X\beta + \varepsilon, \tag{3.3.1}$$

where response variable y is a $p \times 1$ vector, covariates X is a $p \times q$ matrix, regression coefficient β is a $q \times 1$ vector. The error vector ε follows a multivariate Normal distribution $\varepsilon \sim \mathcal{N}(\varepsilon|\ 0, \sigma^2\mathbb{I}_p)$ with a fixed value of σ^2, and $\mathbb{I}_p$ denotes a $p \times p$ identity matrix. For high-precision sensors, the estimates of the regression coefficient β are highly accurate and the variation is extremely small (of the order $\times(10^{-7})$), hence we fix the value of σ^2 as a constant.

Instead of constructing a single Normal distribution as the prior information for the regression coefficient β, as proposed by Eves et al. [2015]. We provide a new way to provide more accurate prior information for the current sensor through a Dirichlet process mixture (DPM) model. The idea is to use n number of OLS estimates $\{\beta_i\}_{i=1}^n$ as observations to build a DPM of Normal distributions, where the OLS estimates are obtained from the historical dataset $\{(y_i, X_i)\}_{i=1}^n$ with full measurements. We can then describe β by a finite number of distinct Normal components, which can be used as the prior choices for the current t^{th} sensor with reduced measurements y_t^- and X_t^- for $t = n+1, \ldots, N$.

Given historical OLS estimates $\{\beta_i\}_{i=1}^n$, the regression coefficient β can be modelled by infinite mixture of multivariate Normal distributions as follows,

$$f_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Sigma}}(\beta) = \sum_{j=1}^{\infty} \omega_j\ \mathcal{N}(\beta|\ \mu_j, \Sigma_j), \tag{3.3.2}$$

where $\boldsymbol{\omega} = \{\omega_j\}_{j=1}^{\infty}$ represent the stick-breaking weights,

$$\omega_1 = v_1 \quad \text{and} \quad \omega_j = v_j \prod_{\{l<j\}} (1 - v_l) \quad \text{for} \quad j > 1. \tag{3.3.3}$$

Take $\{v_j\}_{j=1}^{\infty}$ to be independent and identically distributed variables from a Beta distribution,

$$\pi(v) \sim \text{Beta}(v|M_1, M_2),$$

with fixed constants of M_1 and M_2. Let component-specific parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{(\mu_j, \Sigma_j)\}_{j=1}^{\infty}$ denote means and variance-covariance matrices for every mixing Normal distribution, and the prior distributions for mean and variance-covariance parameters are given by

$$\begin{aligned} \pi(\mu|\Sigma) \sim & \mathcal{N}(\mu|m_0, \tau_0\Sigma) \\ \pi(\Sigma^{-1}) \sim & \text{Wishart}(\Sigma^{-1}|s_0, S_0^{-1}). \end{aligned} \tag{3.3.4}$$

The initial value of m_0 is obtained as the mean of n OLS estimates of β, such as $m_0 = n^{-1} \sum_{i=1}^{n} \beta_i$, τ_0 is initialised as 1, and S_0 is the precision matrix of $\{\beta_i\}_{i=1}^{n}$.

In order to truncate the infinite mixture model to be finite, we introduce a latent variable U such that the joint density of (β, u) given $(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written as

$$f_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Sigma}}(\beta, u) = \sum_{j=1}^{\infty} \mathbb{1}_{\{u<\xi_j\}} \frac{\omega_j}{\xi_j} \mathcal{N}(\beta|\ \mu_j, \Sigma_j), \tag{3.3.5}$$

where $\boldsymbol{\xi} = \{\xi_j\}_{j=1}^{\infty}$ is generated as a deterministic, positive and decreasing sequence. It is used as an envelope function to improve the efficiency of sampling without changing the mixture model. It is clear that the infinite expression of the desired density $f_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Sigma}}(\beta, u)$ in Equation (3.3.2) can be returned by integrating over u. Hence, the marginal density for U is given by

$$f_{\boldsymbol{\xi}}(u) = \sum_{j=1}^{\infty} \mathbb{1}_{\{u<\xi_j\}} = \sum_{j=1}^{\infty} \xi_j \ \text{Unif}(u|0, \xi_j),$$

where $\text{Unif}(\cdot)$ denotes a uniform distribution. A collocation of j can be obtained as

$$A_{\boldsymbol{\xi}}(u) = \{j : \xi_j > u\}.$$

The collection $A_{\boldsymbol{\xi}}(u)$ is a finite set for all $u > 0$. Then, the joint density of (β, u) in Equation (3.3.5) can be equally written as

$$f_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Sigma}}(\beta, u) = \sum_{j \in A_{\boldsymbol{\xi}}(u)} \frac{\omega_j}{\xi_j} \mathcal{N}(\beta|\mu_j, \Sigma_j).$$

The conditional density of β given u is stated as

$$f_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Sigma}}(\beta|u) = \frac{1}{\sum_j \mathbb{1}_{\{u<\xi_j\}}} \sum_{j \in A_{\boldsymbol{\xi}}(u)} \frac{\omega_j}{\xi_j} \mathcal{N}(\beta|\ \mu_j, \Sigma_j).$$

We introduce the latent variable U as a threshold to rule out those mixing components with insignificant contributions, and have a finite mixture model with equal weights $1/(\sum_j \mathbb{1}_{\{u<\xi_j\}})$.

Furthermore, a second latent variable $\boldsymbol{\gamma}$ is introduced to indicate allocations of observations. We write $\gamma_i = j$ to state that the i^{th} observation is allocated to the j^{th} mixing component, such that $\gamma_i = \gamma_{i'} = j$ if observations β_i and $\beta_{i'}$ share common parameters $\mu_i = \mu_{i'} = \mu_j$ and $\Sigma_i = \Sigma_{i'} = \Sigma_j$. Thus, the joint density of (β, u, γ) can be written as

$$f_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Sigma}}(\beta, u, \gamma) = \mathbb{1}_{\{u<\xi_\gamma\}} \frac{\omega_\gamma}{\xi_\gamma} \mathcal{N}(\beta|\ \mu_\gamma, \Sigma_\gamma).$$

The configuration information of $\gamma_i = j$ means that observation β_i is generated from the j^{th} Normal component with parameters (μ_j, Σ_j). By introducing the latent variable $\boldsymbol{\gamma}$ significantly improves efficiency of the sampling scheme.

With observations $\{\beta_i\}_{i=1}^n$ and latent variables $\{u_i\}_{i=1}^n$ and $\{\gamma_i\}_{i=1}^n$, the complete likelihood function can be written as,

$$\mathrm{L}_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Sigma}}(\{(\beta_i, u_i, \gamma_i)\}_{i=1}^n) = \prod_{i=1}^n \mathbb{1}_{\{u_i<\xi_{\gamma_i}\}} \frac{\omega_{\gamma_i}}{\xi_{\gamma_i}} \mathcal{N}(\beta_i|\ \mu_{\gamma_i}, \Sigma_{\gamma_i}).$$

Also, the full posterior of component-specific parameters is proportional to

$$p(\mu, \Sigma | \{(\beta_i, u_i, \gamma_i)\}_{i=1}^n) \propto \pi(\mu|\Sigma)\ \pi(\Sigma) \prod_{i=1}^{n} \mathbb{1}_{\{u_i < \xi_{\gamma_i}\}}\ \frac{\omega_{\gamma_i}}{\xi_{\gamma_i}}\ \mathcal{N}(\beta_i|\ \mu_{\gamma_i}, \Sigma_{\gamma_i}).$$

where $\pi(\mu|\Sigma)$ and $\pi(\Sigma)$ denote the prior distributions in Equations (3.3.4).

To be able to implement the slice sampler introduced by Kalli et al. [2011] to the model, a set of full conditional density function is required. We initialise the chain by assigning $\{\gamma_i = i\}_{i=1}^n$ and the slice sampler can be proceeded by following a Gibbs sampling scheme.

Step 1. The latent variable U can be sampled from a Uniform distribution according to the configuration indicators $\boldsymbol{\gamma}$,

$$u_i \sim \mathrm{Unif}(u_i | 0, \xi_{\gamma_i}),$$

where the sequence $\{\xi_{\gamma_i}\}_{i=1}^n$ is determined by $\xi_{\gamma_i} = \exp\{-b \cdot \gamma_i\}$ with a pre-fixed constant b. For each observation β_i, we need to sample a sufficient number of $\{\xi_j\}_{j=1}^{k_i}$ until we can no longer find a $\xi_{k_i+1} > u_i$. Then, the number of mixture components k_i can be determined by,

$$k_i = \lfloor -\frac{\log(u_i)}{b} \rfloor, \tag{3.3.6}$$

where $\lfloor x \rfloor$ stands for the greatest integer that is less than or equal to x. Hence, the total number of mixture components k^* can be obtained by $k^* = \max_{\{i=1,\ldots,n\}}\{k_i\}$.

Step 2. We sample v_j from its conditional posterior distribution, which follows a Beta distribution,

$$(v_j|\ \{\beta_i, u_i, \gamma_i\}_{i=1}^n, M_1, M_2) \sim \mathrm{Beta}\left(v_j|\ n_j + M_1, m_j + M_2\right).$$

The term $n_j = \sum_{i=1}^n \mathbb{1}_{\{i:\gamma_i = j\}}$ represents the number of observations in the j^{th} component, and $m_j = \sum_{i=1}^n \mathbb{1}_{\{i:\gamma_i > j\}}$ sums the number of observations which have not been allocated to any of the previous $1, \ldots, j-1$ components yet. The component weight parameter ω_j can be updated by substituting the posterior of v_j to the stick-breaking expression mentioned in Equation (3.3.3).

Step 3. The conditional posterior distributions of parameters μ_j and Σ_j can be derived as

$$p(\mu_j, \Sigma_j | \{(y_i, X_i)\}_{\{i:\gamma_i=j\}}, m_0, \tau_0, s_0, S_0) \propto \pi(\mu|\Sigma)\pi(\Sigma) \prod_{\{i:\gamma_i=j\}} \mathcal{N}(\beta_i | \mu_{\gamma_i}, \Sigma_{\gamma_i}),$$

where $\pi(\mu|\Sigma)$ and $\pi(\Sigma)$ represent prior distributions stated in Equations (3.3.4). Hence, component parameters μ_j and Σ_j can be sampled from

$$\begin{aligned}(\mu_j|\beta_{\{i:\gamma_i=j\}}, \Sigma_j, m_0, \tau_0) &\sim \mathcal{N}(\mu_j | m_j,\ \tau_j\Sigma_j),\\ (\Sigma_j^{-1}|\beta_{\{i:\gamma_i=j\}}, \mu_j, s_0, S_0) &\sim \mathcal{W}\left(\Sigma_j^{-1} | n_j + s_0 + 1,\ S_j\right),\end{aligned}$$

where $\mathcal{W}(\cdot)$ denotes the Wishart distribution, and hyperparameters can be updated as

$$\begin{aligned} m_j =& \tau_j \left(\sum_{\{i:\gamma_i=j\}} \beta_i + \frac{1}{\tau_0} m_0 \right), \quad \tau_j = \left(n_j + \frac{1}{\tau_0} \right)^{-1}, \\ \text{and} \ \ S_j =& \left(S_0 + \sum_{\{i:\gamma_i=j\}} (\beta_i - \mu_j)(\beta_i - \mu_j)^T + \frac{1}{\tau_0}(\mu_j - m_0)(\mu_j - m_0)^T \right)^{-1}. \end{aligned}$$

Step 4. Given updated parameters of each component, the conditional distribution of γ_i is proportional to

$$p(\gamma_i = j | \beta_i, \omega_j, \mu_j, \Sigma_j) \propto \ w_j \ \mathcal{N}(\beta_i|\mu_j, \Sigma_j).$$

The value of γ_i for each observation β_i lies in between 1 to k^*, where k^* is the largest number of components obtained in **Step 1**. For a component with no allocation yet, the component parameters are independently sampled from the corresponding prior distributions.

Therefore, all the conditional densities are available in closed forms and the Markov chain is complete. The described Gibbs sampling scheme is iteratively proceeded through **Step 1 - 4** until convergence is reached.

Until now, the observations $\{\beta_i\}_{i=1}^n$ are allocated to k^* distinct Normal components with parameters $\{(\mu_j, \Sigma_j)\}_{j=1}^{k^*}$ and corresponding weights $\{\omega_j\}_{j=1}^{k^*}$. When we observe new sensors data with reduced measurements $\{(y_t^-, X_t^-)\}_{t=n+1}^N$, these k^* Normals are treated as prior choices for the regression coefficients β in the linear model stated in Equation (3.3.1). Set $\{\gamma_t\}_{t=n+1}^N$ denote the prior allocation indicators of $\{(y_t^-, X_t^-)\}_{t=n+1}^N$. Denote $\gamma_t = j$ as the j^{th} Normal component is assigned to be the prior distribution of the t^{th} sensor. For each sensor t, the probability of the prior selection $\gamma_t = j$ is proportional to

$$p(\gamma_t = j|\ y_t^-, X_t^-, \omega_j, \mu_j) \propto\ \omega_j\ \mathcal{N}(y_t^-|\ X_t^-\mu_j,\ \sigma^2\mathbb{I}_{p^-}).$$

The terms ω_j and μ_j are the weight and the mean respectively for the corresponding j^{th} Normal component, and p^- is the dimension of response variable y_t^- with reduced measurements. Therefore, the prior allocation for the current sensor γ_t is selected as the component j with the highest probability $p(\gamma_t = j)$. Given the prior allocation $\gamma_t = j$, the prior distribution of the regression coefficient β_t is given by

$$\pi(\beta_t) \sim \mathcal{N}(\beta_t|\mu_j, \Sigma_j).$$

Then, the posterior distribution of β_t is derived as,

$$p(\beta_t|\ y_t^-, X_t^-, \gamma_t = j, \mu_j, \Sigma_j) \propto \pi(\beta_t)\mathcal{N}\left(y_t^-|\ X_t^-\beta_t, \sigma^2\mathbb{I}_{p^-}\right),$$

which follows a Normal distribution,

$$(\hat{\beta}_t|\ y_t^-, X_t^-, \gamma_t = j, \mu_j, \Sigma_j)\ \sim \mathcal{N}\left(\hat{\beta}_t|\ \mu_t, \Sigma_t\right), \tag{3.3.7}$$

where

$$\mu_t = \left(\Sigma_j^{-1} + \frac{1}{\sigma^2} X_t^{-T} X_t^{-}\right)^{-1} \left(\Sigma_j^{-1}\mu_j + \frac{1}{\sigma^2} X_t^{-T} y_t^{-}\right),$$
$$\Sigma_t = \left(\Sigma_j^{-1} + \frac{1}{\sigma^2} X_t^{-T} X_t^{-}\right)^{-1}.$$

We use the posterior inference of $\hat{\beta}_t$ given in Equation (3.3.7) as the characterised value of the critical parameter β for the t^{th} sensor with reduced measurements (y_t^-, X_t^-).

To validate the proposed BL-DPMP approach, we substitute the estimate $\hat{\beta}_t$ to the the following equation,

$$\varepsilon_t = y_t - X_t \hat{\beta}_t,$$

where (y_t, X_t) denote the full measurements recorded for the current t^{th} sensor. Hence, the resulting error ε_t is calculated from the complete data by using $\hat{\beta}_t$ obtained from the BL-DPMP model. We use the error terms $\{\varepsilon_t\}_{t=n+1}^{N}$ to asses the performance of the proposed BL-DPMP approach, where details are described in Section 3.4.

3.3.2 Dirichlet Process Mixture Regression Model with Data-Driven Prediction Procedure

In this section, we propose a Dirichlet process mixture regression model with data-driven prediction (DPMR-DDP) procedure to describe the relationship between the response variables $\{y_i\}_{i=1}^{n}$ and covariates $\{X_i\}_{i=1}^{n}$ in the historical dataset and enhance the prediction accuracy by adding a data-driven prediction procedure. We have a nonparameter regression model for $\{(y_i, X_i)\}_{i=1}^{n}$, which can be stated as

$$f_{\boldsymbol{\omega},\boldsymbol{\beta}}(y_i|X_i) = \sum_{j=1}^{\infty} \omega_j \ \mathcal{N}(y_i|X_i\beta_j, \sigma^2\mathbb{I}_p), \tag{3.3.8}$$

where X_i is a $p \times q$ matrix of covariates, and y_i is a $p \times 1$ vector of response variable for $i = 1, \dots, n$. The mixing weights of components are denoted by $\boldsymbol{\omega} = \{\omega_j\}_{j=1}^{\infty}$, which follows the stick-breaking process as described in Equations (3.3.3). The regression coefficients $\boldsymbol{\beta} = \{\beta_j\}_{j=1}^{\infty}$ follow a Normal prior distribution $\mathcal{N}(\beta|\mu_0, \Sigma_0)$, where μ_0 is obtained as the average of the n OLS estimates of $\{\beta_i\}_{i=1}^{n}$ and Σ_0 is the variance-covariance matrix of $\{\beta_i\}_{i=1}^{n}$. We fix variance σ^2 to be a small constant for the same reason mentioned in Section 3.3.1.

Now, we need to find the corresponding regression version of the DPM model. Similarly, two sets of latent variables U and $\boldsymbol{\gamma}$ are introduced to transform the infinite mixture model to be

$$f_{\boldsymbol{\omega},\boldsymbol{\beta}}(y, u, \gamma |\ X) = \mathbb{1}_{\{u<\xi_\gamma\}} \frac{\omega_\gamma}{\xi_\gamma} \mathcal{N}(y|\ X\beta_\gamma, \sigma^2\mathbb{I}_p).$$

Given observed pairs $\{(y_i, X_i)\}_{i=1}^{n}$, we can write down the full likelihood function as

$$\mathrm{L}_{\boldsymbol{\omega},\boldsymbol{\beta}}(\{(y_i, X_i, u_i, \gamma_i)\}_{i=1}^{n}) = \prod_{i=1}^{n} \mathbb{1}_{\{u_i<\xi_{\gamma_i}\}} \frac{\omega_{\gamma_i}}{\xi_{\gamma_i}} \mathcal{N}(y_i|\ X_i\beta_{\gamma_i}, \sigma^2\mathbb{I}_p).$$

Hence, the conditional posterior of regression coefficient β can be derived as

$$p(\beta|\{(y_i, X_i, u_i, \gamma_i)\}_{i=1}^{n}) \propto \pi(\beta) \prod_{i=1}^{n} \mathbb{1}_{\{u_i<\xi_{\gamma_i}\}} \frac{\omega_{\gamma_i}}{\xi_{\gamma_i}} \mathcal{N}(y_i|\ X_i\beta_{\gamma_i}, \sigma^2\mathbb{I}),$$

where $\pi(\beta)$ denotes the prior distribution stated in model (3.3.8).

Similar to the slice sampling scheme described in Section 3.3.1, we initialise the chain by assigning $\{\gamma_i = i\}_{i=1}^{n}$, and proceed the iterative procedure as follows.

Step 1. The latent variable U can be sampled from a Uniform distribution according to the configuration indicator $\boldsymbol{\gamma}$,

$$u_i \sim \mathrm{Unif}(u_i|0, \xi_{\gamma_i}).$$

The sequence $\boldsymbol{\xi} = \{\xi_{\gamma_i}\}_{i=1}^{n}$ is determined by $\xi_{\gamma_i} = \exp\{-b \cdot \gamma_i\}$ with a fixed constant b, where the value of b controls the decreasing speed of $\boldsymbol{\xi}$. Given the sampled value of u_i, we can truncate a sufficient number of $\{\xi_j\}_{i=1}^{k_i}$ and determine the number of mixture components k_i by Equation (3.3.6). Thus, the total number of mixture components k^* can be found by $k^* = \max_{\{i=1,\ldots,n\}}\{k_i\}$.

Step 2. Given that v_j follows a Beta prior distribution Beta$(v_j|\ M_1, M_2)$, we can sample v_j from the conditional posterior distribution,

$$(v_j|\ \{(y_i, X_i, u_i, \gamma_i)\}_{i=1}^n, M_1, M_2) \sim \mathrm{Beta}\left(v_j|\ n_j + M_1,\ m_j + M_2\right),$$

where $n_j = \sum_{i=1}^n \mathbb{1}_{\{i:\gamma_i=j\}}$ and $m_j = \sum_{i=1}^n \mathbb{1}_{\{i:\gamma_i>j\}}$. Then, the weight parameter ω_j can be obtained by substituting the posterior of v_j into the stick-breaking expression stated in Equation (3.3.3).

Step 3. According to the configuration indicator $\boldsymbol{\gamma}$, we can find a set of observations indices in the j^{th} component. The conditional posterior distribution of β_j can be derived as

$$p(\beta_j|\ \{(y_i, X_i)\}_{\{i:\gamma_i=j\}}, \mu_0, \Sigma_0) \propto \mathcal{N}(\beta|\ \mu_0, \Sigma_0) \prod_{\{i:\gamma_i=j\}} \mathcal{N}(y_i|\ X_i\beta_{\gamma_i}, \sigma^2\mathbb{I}_p).$$

Hence, we can sample β_j from

$$(\beta_j|\ \{(y_i, X_i)\}_{\{i:\gamma_i=j\}}, \mu_0, \Sigma_0) \sim\ \mathcal{N}(\beta_j|\ \mu_j, \Sigma_j), \tag{3.3.9}$$

$$\text{where}\ \ \mu_j = \left(\frac{1}{\sigma^2}\sum_{\{i:\gamma_i=j\}} X_i^T X_i + \Sigma_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\sum_{\{i:\gamma_i=j\}} X_i^T y_i + \Sigma_0^{-1}\mu_0\right),$$
$$\text{and}\ \ \Sigma_j = \left(\frac{1}{\sigma^2}\sum_{\{i:\gamma_i=j\}} X_i^T X_i + \Sigma_0^{-1}\right)^{-1}.$$

Step 4. Given updated parameters $\{(\omega_j, \mu_j, \Sigma_j)\}_{j=1}^{k^*}$ of every existing component, the conditional distribution of γ_i is proportional to

$$p(\gamma_i = j|\ \{(y_i, X_i)\}_{i=1}^n, \omega_j, \beta_j) \propto\ \frac{w_j}{\xi_j}\,\mathcal{N}(y_i|\ X_i\beta_j, \sigma^2\mathbb{I}_p).$$

for $j = 1, \ldots, k^*$. The value of γ_i is chosen to be the component with the greatest probability.

Therefore, the chain is completed and the slice sampler can be proceeded until convergence is reached. Up to this stage, the historical data $\{(y_i, X_i)\}_{i=1}^n$ are allocated into a DPM regression model with k^* distinct regression coefficients $\{\beta_j\}_{j=1}^{k^*}$.

Next, we need to predict the value of y_t given X_t with full measurements, when only the incomplete pair of data (y_t^-, X_t^-) is available. We consider the prediction obtained from the standard DPM regression model to be not accurate enough for high-precision sensor characterisation. Here, we briefly discuss the reasons and propose a data-driven prediction (DDP) procedure to improve the accuracy of the prediction.

Our aim is to predict the value of response variable y_t given the covariates X_t with full measurements. In general, based on the estimated DPM regression model with k^* components, the predictive distribution of y_t given X_t can be derived as

$$\begin{aligned} f(y_t|X_t) =& \frac{\alpha_0}{\alpha_0+n+1}\int \mathcal{N}(y_t|\ X_t\beta_0, \sigma^2\mathbb{I}_p)\mathcal{N}(\beta_0|\ \mu_0, \Sigma_0)\mathrm{d}\beta_0 \\ &+ \sum_{j=1}^{k^*}\frac{n_j}{\alpha_0+n+1}\int \mathcal{N}(y_t|\ X_t\beta_j, \sigma^2\mathbb{I}_p)\mathcal{N}(\beta_j|\ \mu_j, \Sigma_j)\mathrm{d}\beta_j, \\ =& \frac{\alpha_0}{\alpha_0+n+1}\mathcal{N}\left(y_t|\ X_t\mu_0,\ \sigma^2\mathbb{I}_p + X_t\Sigma_0 {X_t}^T\right) \\ &+ \sum_{j=1}^{k^*}\frac{n_j}{\alpha_0+n+1}\mathcal{N}\left(y_t|\ X_t\mu_j,\ \sigma^2\mathbb{I}_p + X_t\Sigma_j {X_t}^T\right), \end{aligned} \tag{3.3.10}$$

where the term of μ_0 and Σ_0 represent the initial values of mean and covariance matrix in the mixture regression model stated in Equation (3.3.8). As a result, the prediction of y_t is simply an average of all component-specific predictions $\{X_t\mu_j\}_{j=0}^{k^*}$ with weights given by the Pólya urn scheme described in Section 2.1.3. Following this way, the prediction of y_t does not involve any information carried in the available reduced measurements (y_t^-, X_t^-). As we mentioned earlier, each sensor needs to be individually characterised in order to achieve the required high precision. Therefore, the prediction of y_t obtained from Equation (3.3.10) is not considered to be accurate enough as it is not specified for the t^{th} sensor by using the information from (y_t^-, X_t^-).

In order to improve the prediction performance, we can specify a suitable regression relationship from the estimated mixture regression model for each pair of reduced measurements (y_t^-, X_t^-). Let $\gamma_t = j$ to represent that the j^{th} regression component is assigned to the observed pair of (y_t^-, X_t^-). For each sensor t, the predictive probability is proportional to

$$p(\gamma_t = j | y_t^-, X_t^-, \beta_j) \propto \mathcal{N}(y_t^- |\ X_t^- \mu_j,\ \sigma^2 \mathbb{I}_{p^-}), \tag{3.3.11}$$

for $j = 1, \ldots, k^*$. Then, the predictive distribution of y_t given X_t with full measurements can be derived as

$$f(y_t | X_t, \gamma_t = j) \sim \mathcal{N}(y_t | X_t \mu_j, \sigma^2 \mathbb{I}_p + X_t \Sigma_j X_t^T). \tag{3.3.12}$$

In this way, the information carried in reduced measurements (y_t^-, X_t^-) is incorporated in $\gamma_t = j$ in order to identify which regression relationship is the most suitable one in the estimated mixture regression model. The predictive distribution of y_t given X_t is draw based on the selected regression coefficient β_j with parameters (μ_j, Σ_j).

However, following the procedure described above, the prediction of y_t given X_t is simply estimated by using the regression coefficient μ_j obtained from (3.3.9). The observed data (y_t^-, X_t^-) becomes more likely to be allocated to the largest component due to the clustering feature of the Dirichlet process. Two different sensors X_t^- and $X_{t'}^-$ may share the same value of regression coefficient $\beta_t = \beta_{t'} = \mu_j$, if we find $\gamma_t = \gamma_{t'} = j$ for another sensor $t' \neq t$. In this case, the incomplete information of (y_t^-, X_t^-) is only participated during the selection stage of $\gamma_t = j$. For high-precision sensors in the considered situation, the prediction drawn by the shared value of μ_j may also fail the sensor to meet the required high precision.

This prediction accuracy issue of the DPM regression model is also noticed by other researchers: modified versions of the DPM regression model can be found in Shahbaba and Neal [2009] and Wade et al. [2014], and proposals to allow for covariate-dependent weights can be found in Griffin and Steel [2006] and Dunson and Park [2008]. In this section, we propose a data-driven prediction procedure to further enhance the accuracy of prediction of y_t given X_t, by incorporating the information carried in reduced measurements (y_t^-, X_t^-) to the regression coefficients $\{\beta_j\}_{j=1}^{k^*}$. Once the mixture regression model with k^* distinct values of regression coefficients $\{\beta_j\}_{j=1}^{k^*}$ is obtained from the historical dataset $\{(y_i, X_i)\}_{i=1}^{n}$, the information carried in the reduced measurements (y_t^-, X_t^-) can be used to update every regression coefficient β_j for $j = 1, \ldots, k^*$ through the following way,

$$(\beta_j^{(t)}|y_t^-, X_t^-, \mu_j, \Sigma_j) \ \sim \mathcal{N}(\beta_j^{(t)}|\ \mu_j^{(t)},\ \Sigma_j^{(t)}), \tag{3.3.13}$$

$$\text{where}\ \ \mu_j^{(t)} = \left(\frac{1}{\sigma^2} {X_t^-}^T y_t^- + \Sigma_j^{-1}\mu_j\right)\left(\frac{1}{\sigma^2} {X_t^-}^T X_t^- + \Sigma_j^{-1}\right)^{-1},$$

$$\text{and}\ \ \Sigma_j^{(t)} = \left(\frac{1}{\sigma^2} {X_t^-}^T X_t^- + \Sigma_j^{-1}\right)^{-1},$$

where the posterior distributions of parameters μ_j and Σ_j are given in Equation (3.3.9). Now we have updated k^* distinct regression coefficients $\{\beta_j^{(t)}\}_{j=1}^{k^*}$. We then use a maximisation procedure to determine the allocation of the t^{th}, where the probability of $\gamma_t = j$ is proportional to

$$p(\gamma_t = j|\ y_t^-, X_t^-, \beta_j^{(t)}) \propto \mathcal{N}(y_t^-|\ X_t^-\beta_j^{(t)},\ \sigma^2\mathbb{I}_{p^-}),$$

for each $j = 1, \ldots, k^*$. Therefore, given $\gamma_t = j$, the predictive distribution of y_t given X_t with full measurements can be obtained as,

$$f(y_t|\ X_t) \sim \mathcal{N}\left(y_t|\ X_t\mu_j^{(t)},\ \sigma^2\mathbb{I}_p + X_t\Sigma_j^{(t)}X_t^T\right). \tag{3.3.14}$$

Following the proposed data-driven prediction procedure, the accuracy of prediction is significantly improved and can outperforms the standard DPM prediction.

The numerical evidence is presented in Section 3.4.

3.4 Results Analysis

In this section, we demonstrate how to apply both proposed BL-DPMP and DPMR-DDP approaches to reduce the required number of measurements needed per sensor. We also provide evidence to show that the Bayesian linear approach with single Normal prior introduced by Eves et al. [2015] is unlikely to provide accurate estimates of β for some types of sensors with multi-modality shown in historical parameter estimates. By taking the S1 dataset as an example, we present performances obtained from both of the proposed approaches when characterising the S1 type of sensors with reduced measurements.

To assess and to compare the performances of the proposed approaches, we count the number of sensors failed to meet the required high precision from different methods and calculate the corresponding pass rates. Meanwhile, the number of failures obtained via OLS method with complete data can be used as a benchmark. To verify those failed sensors, we present the characterisation testing procedure first.

Recall that the S1 dataset contains information of 895 previously characterised sensors in total, the first n sensors with full measurements $\{(y_i, X_i)\}_{i=1}^{n}$ are used to calculate the parameter estimates $\{\beta_i\}_{i=1}^{n}$ and to construct the prior information of β. Then, we need to characterise the rest $(895 - n)$ sensors with reduced measurements $\{(y_t^-, X_t^-)\}_{t=n+1}^{895}$. Then, we use the obtained estimates $\{\beta_t\}_{t=n+1}^{895}$ to calculate the error terms,

$$\varepsilon_t = y_t - X_t\beta_t, \tag{3.4.1}$$

where ε_t is a 32×1 vector $(\varepsilon_{1,t}, \ldots, \varepsilon_{32,t})^T$ for $t = n+1, \ldots, 895$. Based on the temperature pattern shown in Figure 3.2(a), the absolute values of average errors are calculated for every temperature band,

$$\bar{\varepsilon}_{\text{low1},t} = \frac{1}{6}\sum_{l=1}^{6}|\varepsilon_{l,t}|, \quad \bar{\varepsilon}_{\text{medium1},t} = \frac{1}{6}\sum_{l=7}^{12}|\varepsilon_{l,t}|, \quad \bar{\varepsilon}_{\text{high},t} = \frac{1}{6}\sum_{l=13}^{18}|\varepsilon_{l,t}|,$$
$$\bar{\varepsilon}_{\text{medium2},t} = \frac{1}{8}\sum_{l=19}^{24}|\varepsilon_{l,t}|, \quad \bar{\varepsilon}_{\text{low2},t} = \frac{1}{6}\sum_{l=25}^{32}|\varepsilon_{l,t}|. \tag{3.4.2}$$

For a S1 sensor to pass the characterisation test, the absolute values of average errors for all temperature bands $(\bar{\varepsilon}_{\text{low1},t}, \bar{\varepsilon}_{\text{medium1},t}, \bar{\varepsilon}_{\text{high},t}, \bar{\varepsilon}_{\text{medium2},t}, \bar{\varepsilon}_{\text{low2},t})$ calculated in Equation (3.4.2) must lie within a range of $[0, 0.3\%]$. Otherwise, the sensor will be recognised as a failure and sent back for further investigation. For both approaches, we can count the number of sensors which meet the high precision level and the corresponding pass rates. The proposed approaches will be approved to be valid if they reach a comparable level of pass rate to the OLS pass rate obtained with full measurements.

Now, we present the number of failures obtained by the Bayesian linear approach proposed by Eves et al. [2015] with the single Normal prior assumption. The first $n = 30$ OLS estimates $\{\beta_i\}_{i=1}^{30}$ are used to construct the prior information of β, where initial values are obtained in Equations (2.4.6). Hence, there are 865 sensors need to be characterised with reduced measurements, and we count the number of failed sensors and calculate the pass rate. We also present the numbers of failures obtained from the currently applied OLS method with full and reduced measurements.

Table 3.1: Numbers of failures obtained among OLS approach with full measurements, Bayesian linear and OLS approaches with reduced measurements, corresponding pass rates and execution time for the S1 type of sensors.

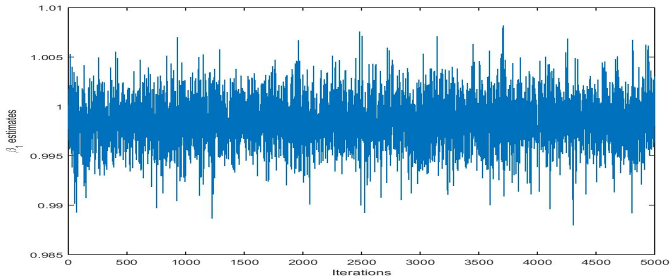
	OLS (full)	Bayesian linear (reduced)	OLS (reduced)
No. of failure at 0.1%	19	527	538
No. of failure at 0.2%	18	343	366
No. of failure at 0.3%	17	233	250
Pass rate	98.03%	73.06%	71.10%
Time (second)	0.04	0.1	0.04

As shown in Table 3.1, the numbers of failures are collected under three different errors limits 0.1%, 0.2% and 0.3%. We identify 17 failed sensors at 0.3% level after being characterised by OLS method with full measurements. The pass rate 98.03% can be treated as the benchmark for other methods, as it is obtained by using the current practice in the real characterisation line. On the other hand, the Bayesian linear approach with reduced measurements gives 233 failures with pass rate 73.06%, and the OLS approach with reduced measurements gives a slightly worse result: 250 failures with pass rate 71.1%.

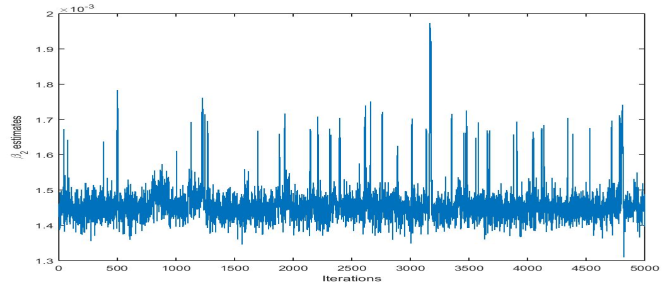
It is obvious that when dealing with the sensor data with reduced measurements, the pass rates drop significantly for both OLS approach and Bayesian linear approach with the single Normal prior assumption. As we expected, OLS method does not perform well since it can not achieve any information about measurements at mid-range temperatures. Meanwhile, the Bayesian linear approach proposed by Eves et al. [2015] suffers from the drawback of assuming a single Normal prior distribution for datasets with multi-modality shown in the historical parameter estimates.

BL-DPMP

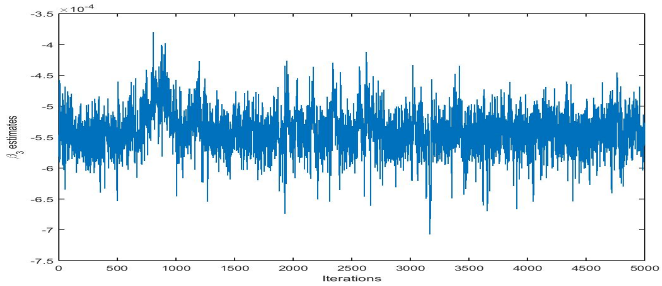
Following the procedure described in Section 3.3.1, we take the first $n = 240$ OLS estimates $\{\beta_i\}_{i=1}^{240}$ as observations, and apply the slice sampler to estimate the Normal mixture density of β. The initial values are chosen as stated in Equations (2.4.6), and we run the slice sampler for 10,000 iterations with 5,000 iterations as the burn-in period. In practice, the slice sampler mixed fast and the trace plot of $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$ estimates for one component after the burn-in period are given in Figure 3.7.



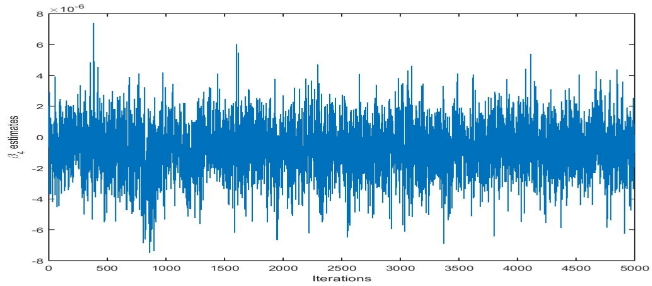
(a) β_1



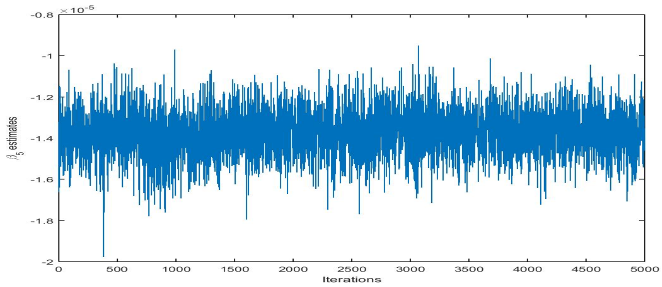
(b) β_2



(c) β_3



(d) β_4



(e) β_5

Figure 3.7: Trace plots of $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$ *estimates for the last 5000 iterations.*

The histogram plot of the second dimension of the OLS estimates $\{\beta_{i,2}\}_{i=1}^{240}$ and the curve fitting are presented in Figure 3.8.

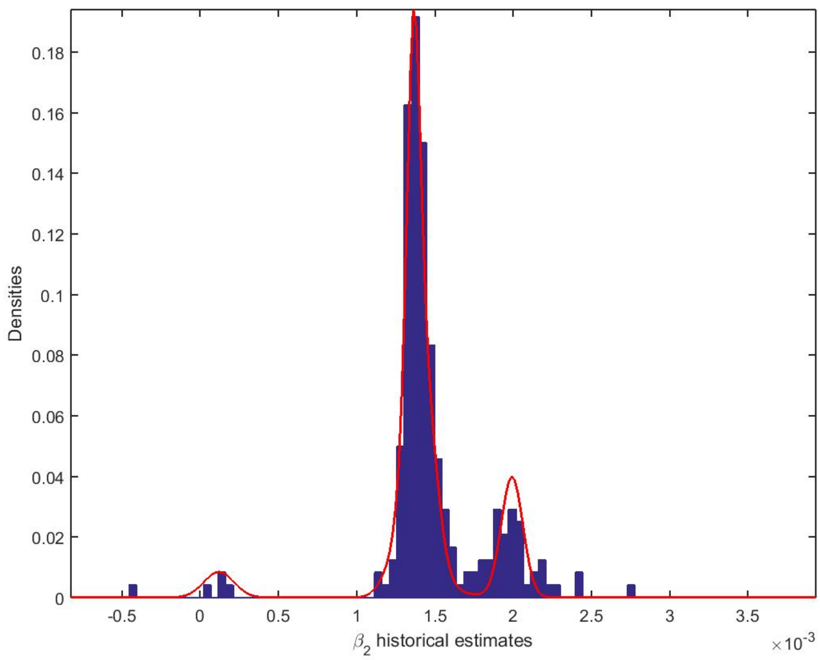


Figure 3.8: Histogram of the historical estimates $\{\beta_{i,2}\}_{i=1}^{240}$ and the curve fitting obtained from the BL-DPMP approach with slice sampler of the S1 type of sensors.

Table 3.2 shows the number of allocations in each component and the corresponding weights displayed in Figure 3.8. According to Figure 3.8, the mixture components are well spread out, and the size of each Normal distribution matches its weight shown in the Table 3.2.

Table 3.2: Number of allocations and corresponding weight for each mixing Normal component obtained from the slice sampler of the S1 historical dataset.

Comp.	No. of allocations	Weight	Comp.	No. of allocations	Weight
1	56	0.2333	2	35	0.1458
3	18	0.075	4	16	0.0667
5	29	0.1208	6	30	0.125
7	13	0.0542	8	8	0.0333
9	7	0.0292	10	11	0.0458
11	5	0.0208	12	1	0.0042
13	8	0.0333	14	2	0.0083
15	1	0.0042			

According to Table 3.2, there are 15 Normal components that survive with actual observation(s) allocated. We set the initial values of $M_1 = 1$ and $M_2 = 3$ for the Beta prior distribution from the stick-breaking process described in Equation (3.3.3). Judging by the curve fitting presented in Figure 3.8, the estimated mixture density of historical dataset is potentially over-fitted. For this considered industrial application, we prefer to keep those components with only a couple of observations. Since the estimated mixture density stated in Table 3.2 will be used as the historical information to compensate the missed information for characterising the next sensor without measurements at warm temperatures, the more number of components actually provide more prior choices for sensor characterisations. Hence, it is more likely for a future sensor to find an appropriate prior distribution.

The estimated mixing components become the prior choice for estimating β_t given reduced measurements (y_t^-, X_t^-). The final estimate of β_t can be drawn according to Equation (3.3.7), then we can calculate the error term ε_t by Equation (3.4.1) and the corresponding average errors by Equations (3.4.2). A comprehensive comparison of the performances achieved by different approaches with both complete and incomplete datasets are presented in Table 3.3.

Table 3.3: Comparison performances obtained from OLS, Bayesian Linear approach and BL-DPMP approach with slice sampler on both the complete and incomplete datasets of the S1 type of sensors.

		No. of failures (655)			Pass rate	Time (sec)
		0.1%	0.2%	0.3%		
OLS	Complete	4	3	2	99.69%	0.04
	Incomplete	432	290	199	69.62%	0.05
Bayesian Linear	Complete	4	3	2	99.69%	0.1
	Incomplete	431	290	199	69.62%	0.1
BL-DPMP	Complete	6	3	2	99.69%	1350
	Incomplete	162	24	8	98.78%	1936

In Table 3.3, all methods provide 2 failures out of 655 sensors with full measurements. It shows that the proposed BL-DPMP approach works well when data is fully observed. If we focus on the 0.3% column, both OLS and Bayesian linear methods provide 199 failures with incomplete data. On the other hand, the proposed BL-DPMP model is able to reduces the failure number to 8, which pushes the pass rate from 69.62% up to 98.78%. It is worthy to note that the 98.78% pass rate is achieved from sensors without having measurements at mid-range temperatures. Similar results are observed from other two columns with 0.1% and 0.2% error limits. Therefore, when dealing with incomplete data, BL-DPMP approach consistently provide much fewer numbers of failures compare to OLS and Bayesian linear approaches.

To further investigate the performance of these three methods, we take one sensor as an example to compare the corresponding error terms.

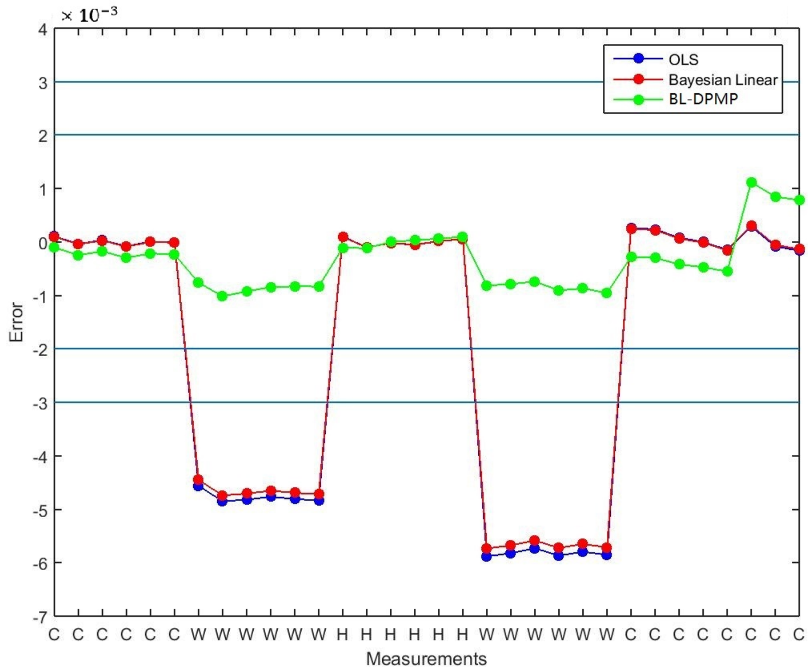


Figure 3.9: Errors comparison for one product among OLS, Bayesian linear and BL-DPMP approach with slice sampler of the S1 type of sensors.

In Figure 3.9, the errors calculated from different methods with incomplete data are plotted with different colours. The horizontal lines represent ±0.2% and ±0.3% error limits. For the current S1 sensor with 32 measurements and 5 temperature bands, both OLS (blue) and Bayesian linear (red) approaches are failed, since the resulting average errors for medium temperature bands are below the -0.3% error limit. It is clear that this sensor can be considered as a failed product by using OLS method or the Bayesian linear approach. On the other hand, the errors from BL-DPMP approach (green) are successfully narrowed into the $\pm 0.1\%$ error interval, which implies that the sensor is well characterised.

Here, I would like to briefly discuss the normality assumption of the error term $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{32})'$. We assumed that ε follows a Normal distribution $\mathcal{N}(\varepsilon|0, \sigma^2 \mathbb{I}_{32})$ with fixed value of σ^2 in Equation (3.3.1). However, in Figure 3.9, the estimated errors do not seem to perfectly fluctuate around zero. This phenomenon is simply caused by the manufacturing reasons in the real world industrial application. It is not possible to have a clear Normal distribution with only 32 error points. The overall error does follow a Normal distribution, this is simply because of the large sample size. Therefore, it is reasonable to assume the error term ε following a Normal distribution with extreme small variance.

Based on the results shown above, the proposed BL-DPMP method is capable to reduce the number of measurements required per sensor without losing much of the estimation precision. The 98.78% pass rate from the proposed approach is considered to be tolerable, given the subsequent benefits from the reduction in both costs and time for the characterisation process.

DPMR-DDP

By taking the first 240 full measurements $\{(y_i, X_i)\}_{i=1}^{240}$ as observations, we build a DPM regression model to describe the relationship between response variable y and covariates X, by executing the slice sampler for 10,000 iterations and 9,800 iterations as the burn-in period. Then, given the estimated mixture regression model with distinct k^* regression coefficients, the prediction of y_t conditioning on X_t with full measurements is obtained by incorporating the available information carried in the reduced measurements (y_t^-, X_t^-) for the current t^{th} sensor via the proposed data-driven prediction procedure.

As discussed in Section 3.3.2, the prediction obtained from the DPM regression model is not considered to be accurate enough for high-precision sensor in this industrial case. We compare the prediction performances of DPM regression with and without data-driven prediction procedure, along with methods mentioned previously. The overall results from different methods with complete and incomplete datasets are summarised in Table 3.4.

Table 3.4: Comparison of number of failures caused by OLS, Bayesian Linear, BL-DPMP, DPMR and DPMR-DDP for both the complete and incomplete datasets of the S1 type of sensors.

		No. of Failures (655)			Pass rate	Time (sec)
		0.1%	0.2%	0.3%		
OLS	Complete	4	3	2	99.69%	0.04
	Incomplete	432	290	199	69.62%	0.05
Bayesian Linear	Complete	4	3	2	99.69%	0.1
	Incomplete	431	290	199	69.62%	0.1
BL-DPMP	Complete	6	3	2	99.69%	1350
	Incomplete	162	24	8	98.78%	1936
DPMR	Complete	408	120	57	91.3%	12220
	Incomplete	590	314	135	79.39%	10990
DPMR-DDP	Complete	4	3	2	99.69%	4318
	Incomplete	309	120	46	92.98%	4522

When dealing with the S1 sensor data with full measurements, all methods perform as good as the OLS method, except for the DPM regression model. The DPM regression model provides 57 failed sensors, which is considered to be too many to meet the company's requirement. The reason is that when the DPM regression model draws prediction of y_t given X_t, it actually uses the same set of regression coefficient $\beta_{\gamma_t=j}$ for all observations with reduced measurements X_t^- that are allocated in this j^{th} mixing component. For instance, if an observation is allocated to the j^{th} Normal component and its location is actually on the edge of the selected Normal curve, it shares the same set of regression coefficient $\{\beta_t\}_{\{t:\gamma_t=j\}}$ as the same as the other observations in the same j^{th} component. Thus, the distance between the component location and the observation causes the sensor to fail the characterisation test. As shown in Equation (3.3.12), the estimated mean μ_j does not carry information from the reduced measurements (y_t^-, X_t^-). The unique information about the t^{th} sensor only participates to the selection stage through Equation(3.3.11). According to the resulting number of failures, this extent of information participation in the standard DPM regression model is not enough to be able to deliver accurate prediction of y_t given X_t.

The inaccurate selection might be another reason for the poor prediction from the DPM regression model. According to Equation (3.3.11), the incomplete pair of (y_t^-, X_t^-) turns out to be the driving force in deciding the allocation of the t^{th} sensor. Extra concern is arisen such that the selected component may not be the most suitable one since the information of the t^{th} sensor only partially contained in reduced measurements (y_t^-, X_t^-). Therefore, the DPM regression model are not considered to be very suitable for this high precision sensor characterisation problem.

To enhance the accuracy of the prediction, we propose a data-driven prediction procedure for the DPM regression model. When data is incomplete for the current sensor, the data-driven prediction incorporates the information of (y_t^-, X_t^-) into the coefficient regression $\beta_j^{(t)}$ by Equation (3.3.14). In this way, the information stored in (y_t^-, X_t^-) provides more contributions of drawing prediction of y_t given X_t. In the proposed data-driven prediction, the allocation of the t^{th} sensor is dependent on $\beta_j^{(t)}$ rather than β_j, where $\beta_j^{(t)}$ has combined the complete information from the historical dataset and the incomplete information for the current sensor. Therefore, the data-driven prediction leads to a better selection for the current sensor characterisation and a better prediction. The numbers of failures obtained from the DPM regression model with and without data-driven prediction procedure are both considered to be too high in the company's point of view. However, the proposed data-driven prediction procedure improves the accuracy of prediction significantly compared to the standard DPM regression model.

In summary, when the data is incomplete, the proposed BL-DPMP approach can achieve a satisfied pass rate comparing to currently applied OLS method with complete data. The results obtained from the DPMR-DDP procedure suggests that this DDP procedure has potential to be further improved. The commonly applied DPM regression model without data-driven prediction procedure ends with the largest number of failures among all methods, which shows it is not suitable for high-precision sensor characterisation considered in this industrial case. The OLS and Bayesian linear model with single Normal prior both collapse for the S1 dataset.

Based on the results we have presented above, the proposed BL-DPMP approach shows the possibility that there is no need to take measurements at the mid-range temperatures for the S1 sensors characterisation in the future. Despite the common drawback of high computational cost of the MCMC type of algorithms, the results approve that our proposed BL-DPMP approach performs as good as the current OLS method with a fewer number of measurements. The potential benefits on cost and time can be introduced to the manufacturer by applying the proposed method on the real characterisation line.

In addition, we applied the proposed BL-DPMP approach to characterise other two types of sensors with reduced measurements. The historical datasets from both types of sensors exhibit single modal structure. In those cases, the proposed BL-DPMP approach was able to accurate describe the historical information and provided 100% pass rate on both types of sensors. Therefore, the BL-DPMP approach is shown to be powerful enough to deal with sensors which exhibit unknown structure of historical datasets.

3.5 Discussion

In this chapter, we proposed two Bayesian approaches, which will be used to reduce the required number of measurements for characterising one type of sensors where a multi-modal structure is exhibited in the historical parameter estimates. The first proposed BL-DPMP approach captured the multi-modality of historical estimates of the regression coefficient $\{\beta_i\}_{i=1}^n$ by a Dirichlet process mixture of

Normal distributions. Following this, one of the mixing components is selected and used to provide the prior information for the current sensor with reduced measurements. The posterior distribution of the coefficient regression β for the current sensor was derived through a Bayesian linear regression model. The proposed BL-DPMP approach achieved a satisfactory pass rate level in comparison to that achieved by OLS method with full measurements as the reference pass rate.

Instead of modelling the historical estimates of the regression coefficient β, we also proposed a DPM regression model to describe the relationship between the response variable $\{y_i\}_{i=1}^n$ and the covariate $\{X_i\}_{i=1}^n$ from the historical dataset and added a data-driven prediction (DPMR-DDP) procedure to enhance the accuracy of the prediction. Since the prediction obtained from the standard DPM regression model was not able to reach the satisfactory level of accuracy in the considered situation, we proposed a data-driven prediction procedure used to draw prediction of y_t given X_t with full measurements for the current sensor by incorporating the incomplete information carried in (y_t^-, X_t^-). According to the results obtained from the proposed DPMR-DDP approach, the proposed prediction procedure significantly improved the prediction performance of the DPM regression model. However, it still produced too many sensors that failed to meet the company's quality specifications.

One concern that arose when applying the proposed BL-DPMP approach to the real characterisation line in industry is how to update the prior information. There might be a slight shift in the β estimates as time goes by. Hence, we need to update the β estimates for the current sensor with reduced measurements to the prior information. This updated prior information will be used to characterise the next sensor with reduced measurements. In future work, the proposed BL-DPMP approach can potentially be extended to apply to a generalised linear model. In this case, the linear relationship between the response variable y and the covariates X in the historical dataset no longer hold. Additionally, the OLS estimators are replaced by the maximum likelihood estimations (MLE).

Other concerns arise when applying the MCMC method to estimate the mixture density. One drawback of MCMC methods is that they are quite computationally intensive, which counteracts the time saved from collecting a fewer number of measurements in the characterisation process. The proposed BL-DPMP approach costs 1936 seconds, more than half an hour, to run 10,000 iterations. In comparison, the OLS method currently used in industry can be executed in under a second. The expensive computational cost would be a major hurdle that we need to address before we can apply the proposed BL-DPMP model to the real characterisation line. Another concern is the repeatability; if we run the sampling method multiple times, the estimated mixture densities always differ slightly from each other. Therefore, we will seek fast alternatives to MCMC methods that can provide deterministic mixture density estimations without losing much precision.

Chapter 4

Hybrid Variational Bayesian Algorithm for Mixture Models

4.1 Introduction

In Chapter 3, we proposed a Bayesian linear model with Dirichlet process mixture prior (BL-DPMP) to reduce the number of measurements required for sensor characterisations. For a certain type S1 sensors, we modelled the historical estimates of parameter β by a DPM of Normal distributions. Hence, the estimated mixture density provides sufficient prior information for characterising sensors with reduced measurements.

A key step in the BL-DPMP approach is to estimate the mixture density for the unknown multi-modal structure from the previous estimates of parameter β. The slice sampler we applied in Chapter 3 can be categorised as a stochastic technique. In general, such a method draws samples under the exact posterior distributions. Hence, given sufficiently large samples, inference drawn from the MCMC approach is reliable and can be treated as a reference for comparisons with other algorithms. However, one of the drawbacks of the MCMC approach is that the sampling techniques usually require expensive computational effort. As a result, the time saved by taking fewer measurements is counteracted by the extra time needed to draw samples. Moreover, the sampling approach cannot guarantee to consistently provide the same estimate of the mixture density. Moreover, if we run the slice sampler multiple times, some observations may be assigned to different mixture components in each run, which may slightly change the posterior inference of the mixing component parameters. This repeatability problem presented by the proposed BL-DPMP approach might be a concern from the manufacturer's point of view, especially when applied to for characterising high-precision sensors.

Instead of the sampling technique, we apply an alternative class of techniques called deterministic approximations to estimate the mixture densities. Without losing much accuracy, the intractable posterior distribution can be analytically approximated by a well-studied parametric form. However, the mixture densities estimated by deterministic approximation techniques can never be exact. We can therefore view it as a complementary approach of MCMC, since it requires less computational effort to provide a deterministic result, but only an approximation. In this chapter, we focus on a family of approximation techniques called *Variational Bayes* (VB) or *Variational Inference* (VI).

The main idea behind Variation Bayes approach is to approximate the exact posterior distribution by positing a family of approximating densities over unobserved variables, and exploring the member of that family that minimises the Kullback-Leibler divergence (KLD) of the posterior distribution. Thus, Variational Bayes turns inference from a simulation problem into an optimisation problem. As one of the critical points of drawing successful variational inference, the chosen approximating family must be flexible enough to give a close approximation to the posterior distribution. Meanwhile, the form of the chosen family should not be too complicated for efficient optimisation. Summaries of contemporary variational approximations can be found in the works of Jordan et al. [1999] and Titterington [2004]. New variational approximation methodologies for particular applications have been proposed by McGrory and Titterington [2007]. The statistical properties of estimators obtained via variational approximation are presented in Wang and Titterington [2006].

One common drawback when applying Variational Bayes to estimate mixture model densities is the local maxima problem. When applying VB to the S1 historical dataset with multi-modality, we discovered that, sometimes, VB inference results in a local maximum rather than in the global maximum. Another well-known shortcoming of VB is that the variance of the posterior distributions is generally underestimated, as discussed in Wang and Titterington [2005]. Various methods have been proposed by researchers in order to overcome these issues. For example, Ueda and Ghahramani [2002] proposed a split and merge procedure to optimise an objective function, which allows simultaneous estimates of the parameters and the number of components in a Normal mixture model. By using the component-elimination property noted by Attias [1999] and Corduneanu and Bishop [2001], a variational optimisation technique is described by McGrory and Titterington [2007], where the algorithm is initialised with a large number of components, and those components with sufficiently small weights are dropped out during the optimisation process. Constantinopoulos and Likas [2007] introduced a progressive mean splitting procedure, which starts with two components. Wu et al. [2012] developed a split and eliminate VB for highly heterogeneous spiky spatial data patterns.

In this chapter, we propose a hybrid Variational Bayes (HVB) algorithm to approximate the DPM of Normal densities. The HVB algorithm starts with an over fitted number of components, and it converges to a suitable number of components according to the component-elimination property. Then, HVB follows a greedy searching scheme which explores every possible model structure. In this way, the risk of becoming trapped in a local maximum is reduced. We can further reduce this risk by incorporating two types of splitting operations, where details are given in Section 5.3.3. This proposed HVB approach deterministically estimates the mixture density by consuming much less computational resources. HVB allows for accurate descriptions of each Normal component and provides flexibility in terms of the number of mixture components needed in order to avoid local maxima problems. Judging by the results from the empirical study, the proposed HVB approach is able to provide accurate approximations of β with reduced measurements with much less the computational effort, in contrast to that required by MCMC approaches.

The remainder of this chapter is organised as follows. In Section 4.2, we describe basic Variational Bayes approximations for Normal mixture density estimation. The proposed HVB algorithm is introduced in Section 4.3. In Section 4.4, we illustrate the HVB algorithm through a simulation study of Normal mixtures. Details of how we apply HVB to the S1 dataset and a comparison with VB and overfitting VB approaches are also included. Section 4.5 discusses the potential drawbacks of HVB and directions for further research.

4.2 Variational Bayes

A central task in Bayesian statistics is to evaluate the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ of the unobserved parameters $\boldsymbol{\theta}$ given the observed data variables $\mathbf{x}$, especially if the posterior is not easy to compute. Consider a joint density

$$p(\mathbf{x}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})$$

with observations $\mathbf{x} = \{x_i\}_{i=1}^n$ and parameters $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^k$, where $p(\boldsymbol{\theta})$ and $p(\mathbf{x}|\boldsymbol{\theta})$ are known as prior and likelihood respectively in a Bayesian model. The inference of $\boldsymbol{\theta}$ can be drawn by computing the conditional density

$$p(\boldsymbol{\theta}|\mathbf{x}) = p(\mathbf{x}, \boldsymbol{\theta})/p(\mathbf{x}).$$

The necessary marginal density $p(\mathbf{x})$ can be evaluated by marginalising out the parameters from the joint density, such as $p(\mathbf{x}) = \int p(\mathbf{x}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$. For many applications in practice, the integral is not available in analytical form or it requires prohibitive computational effort. In this section, we introduce a class of deterministic approximation algorithms based on variational Bayes. The basic idea is to pick an variational distribution $q(\boldsymbol{\theta})$ which follows some tractable family, and then to try to make this $q(\boldsymbol{\theta})$ as close as possible to the exact posterior $p(\boldsymbol{\theta}|\mathbf{x})$. The log marginal density can be equally written as

$$\log p(\mathbf{x}) = \log \int q(\boldsymbol{\theta})\ \frac{p(\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\ \mathrm{d}\theta.$$

Recall that if a function g is concave, then we have $g(\mathbb{E}[\mathbf{x}]) \geq \mathbb{E}[g(\mathbf{x})]$ from Jensen's inequality. Thus, the log marginal density of data can be expressed as

$$\log p(\mathbf{x}) \geq \int q(\boldsymbol{\theta})\ \log\left(\frac{p(\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{\theta},$$

where the right hand side of the inequality is known as the *evidence lower bound* (ELBO),

$$\mathrm{ELBO}(q) = \int q(\boldsymbol{\theta})\ \log\left(\frac{p(\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{\theta}. \tag{4.2.1}$$

The difference between the marginal density $\log p(\mathbf{x})$ of data and ELBO can be written as,

$$\begin{aligned} \log p(\mathbf{x}) - \mathrm{ELBO}(q) &= \log p(\mathbf{x}) - \int q(\boldsymbol{\theta}) \log\left(\frac{p(\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \log\left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{x})}\right)\ \mathrm{d}\boldsymbol{\theta} \\ &= \mathrm{KLD}(q||p), \end{aligned} \tag{4.2.2}$$

which is recognised as a Kullback-Leibler divergence (KLD). The $\mathrm{KLD}(q||p)$ measures the distance between two probability distributions $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{x})$, where the value of KLD is always non-negative and vanishes when $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{x})$. Hence, we can decompose the log marginal likelihood into two parts:

$$\log p(\mathbf{x}) = \mathrm{ELBO}(q) + \mathrm{KLD}(q||p).$$

The quantity $\mathrm{ELBO}(q)$ is tractable to compute with a suitable choices of $q(\boldsymbol{\theta})$. Then, we can maximise the evidence lower bound with respect to the variational distribution $q(\boldsymbol{\theta})$, which is equivalent to minimizing the KLD stated in Equation

(4.2.2).

Assume that parameters $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^k$ are mutually independent, and each of them is governed by a distinct factor in the variational density, such as $q(\boldsymbol{\theta}) = \prod_{j=1}^k q_j(\theta_j)$. It is worth noting that no further assumption is made about the variational distribution $q(\boldsymbol{\theta})$ and no restriction is placed on the functional form of an individual $q_j(\theta_j)$. The optimal value of the ELBO is achieved by maximising with respect to each of the $q_j(\theta_j)$ in turn. Substituting $q(\boldsymbol{\theta}) = \prod_j q_j(\theta_j)$ into the ELBO(q) expression in Equation (4.2.1), we focus on one group $q_l(\theta_l)$ and dependent terms for $l = 1, \ldots, k$. Denoting $q_l(\theta_l)$ by q_l for simplicity, the evidence lower bound can be written as

$$\begin{aligned}
\text{ELBO}(q) &= \int q(\boldsymbol{\theta}) \;\log\left(\frac{p(\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{\theta} \\
&= \int \prod_j q_j \left(\log p(\mathbf{x}, \boldsymbol{\theta}) - \sum_j \log q_j\right) \mathrm{d}\boldsymbol{\theta} \\
&= \int q_l \int \prod_{j \neq l} q_j \left(\log p(\mathbf{x}, \boldsymbol{\theta}) - \log q_l - \sum_{j \neq l} \log q_j\right) \mathrm{d}\theta_{\{j:j\neq l\}} \mathrm{d}\theta_l \\
&= \int q_l \left(\int \log p(\mathbf{x}, \boldsymbol{\theta}) \prod_{j \neq l} q_j \mathrm{d}\theta_j\right) \mathrm{d}\theta_l - \int q_l \log q_l \;\mathrm{d}\theta_l + \text{const},
\end{aligned}$$

where terms unrelated to θ_l are absorbed in the constant term. From right hand side of the expression, the term in bracket $\int \log p(\mathbf{x}, \boldsymbol{\theta}) \prod_{j\neq l} q_j \mathrm{d}\theta_j$ can be treated as the expectation of $\log p(\mathbf{x}, \boldsymbol{\theta})$ with respect to variational distributions over all parameters θ_j for $j \neq l$. We can write down the evidence lower bound and rearrange it as

$$\begin{aligned}
\text{ELBO}(q) &= \int q_l \;\mathbb{E}_{j\neq l}[\log p(\mathbf{x}, \boldsymbol{\theta})] \;\mathrm{d}\theta_l - \int q_l \log q_l \;\mathrm{d}\theta_l + \text{const} \\
&= \int q_l \log\left(\frac{\mathbb{E}_{j\neq l}[\log p(\mathbf{x}, \boldsymbol{\theta})]}{q_l}\right) \;\mathrm{d}\theta_l + \text{const} \\
&= -\;\text{KLD}(q_l || \mathbb{E}_{j\neq l}[\log p(\mathbf{x}, \boldsymbol{\theta})]) + \text{const}.
\end{aligned}$$

Therefore, the optimal solution of $q_l^*(\theta_l)$ is achieved when the Kullback-Leibler divergence between $q_l(\theta_l)$ and $\mathbb{E}_{j\neq l}[\log p(\mathbf{x}, \boldsymbol{\theta})]$ is minimised to zero with other factors $\{q_j\}_{\{j:j\neq l\}}$ fixed,

$$\log q_l^*(\theta_l) = \mathbb{E}_{j\neq l}[\log p(\mathbf{x}, \boldsymbol{\theta})] + \text{const}. \tag{4.2.3}$$

In words, this means that the log of the optimal solution for $q_l(\theta_l)$ is simply obtained by considering the log of the joint distribution over all observed variables $\mathbf{x}$ and unobserved parameters $\boldsymbol{\theta}$, and then taking the expectation with respect to all of the other variational factors $\{q_j(\theta_j)\}_{\{j:j\neq l\}}$. The additive constant is calculated by normalising the distribution $\log q_l^*(\theta_l)$. Thus, the general solution to optimisation of $q_l^*(\theta_l)$ is obtained as

$$q_l^*(\theta_l) = \frac{\exp\{\mathbb{E}_{j\neq l}[\log p(\mathbf{x}, \boldsymbol{\theta})]\}}{\int \exp\{\mathbb{E}_{j\neq l}[\log p(\mathbf{x}, \boldsymbol{\theta})]\}\ \mathrm{d}\theta_l},$$

by taking the exponential on both sides of Equation (4.2.3).

4.2.1 VB for Mixtures of Normals

In this section, we take a mixture of Normal distributions as an example to explain the intractability of the mixture model and demonstrate how to construct a Variational Bayes algorithm to estimate a mixture of Normals. VB for Normal mixtures is well studied by Beal and Ghahramani [2003] and Biship [2006]. Let $\mathbf{x} = \{x_i\}_{i=1}^n$ be a set of D-dimensional observations, which is generated from a finite mixture model of k Normal distributions with mixing weights $\boldsymbol{\omega} = \{\omega_j\}_{j=1}^k$, means $\boldsymbol{\mu} = \{\mu_j\}_{j=1}^k$ and precision matrices $\boldsymbol{\Lambda} = \{\Lambda_j\}_{j=1}^k$. Thus, the model can be written as,

$$p(x_i|\ \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{j=1}^k \omega_j\ \mathcal{N}(x_i|\ \mu_j, \Lambda_j^{-1}).$$

We assign a Dirichlet distribution to be the prior for weight parameters $\boldsymbol{\omega}$,

$$\pi(\boldsymbol{\omega}) \sim \mathrm{Dir}(\boldsymbol{\omega}|\boldsymbol{\alpha}) = C(\boldsymbol{\alpha}) \prod_{j=1}^k \omega_j^{\alpha_j - 1}, \tag{4.2.4}$$

where $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^{k}$ and the normalisation constant $C(\boldsymbol{\alpha})$ is the multivariate Beta function, which is $C(\boldsymbol{\alpha}) = \Gamma(\sum_{j=1}^{k} \alpha_j)/(\prod_{j=1}^{k} \Gamma(\alpha_j))$. For the component parameters, we assume the following factored conjugate joint prior distribution,

$$\pi(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \pi(\boldsymbol{\mu}|\boldsymbol{\Lambda})\pi(\boldsymbol{\Lambda}) = \prod_{j=1}^{k} \mathcal{N}(\mu_j|\ m_0, (\tau_0\Lambda_j)^{-1})\mathcal{W}(\Lambda_j|\ s_0, S_0), \tag{4.2.5}$$

with pre-fixed initial values (m_0, τ_0, s_0, S_0).

As we have mentioned earlier, the marginal density $f(x)$ is necessary, and it is derived by integrating over all parameters $(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$. Given conjugate prior distributions, each individual integral is computable. However, the marginalisations involve summing over all possible configurations of the unobserved parameters, and there are k^n of integrals needs to calculate in total. The required computational effort remains exponential in k, hence it is considered to be intractable in practice.

To overcome this issue, we introduce latent configuration variables $\mathbf{z} = \{z_i\}_{i=1}^{n}$ to indicate allocations of observations $\mathbf{x} = \{x_i\}_{i=1}^{n}$. Each z_i is a binary vector $z_i = (z_{i1}, \ldots, z_{ik})$ and we write $z_{ij} = 1$ if the i^{th} observation is allocated to the j^{th} component for $i = 1, \cdots, n$ and $j = 1, \cdots, k$; otherwise $z_{ij} = 0$. Therefore, the joint density of $(\mathbf{x}, \mathbf{z})$ given $(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ can be written as,

$$f_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Lambda}}(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n}\prod_{j=1}^{k} \omega_j^{z_{ij}}\ \mathcal{N}(x_i|\ \mu_j, \Lambda_j^{-1})^{z_{ij}}.$$

The conditional distribution of $\mathbf{z}$ given the mixing weights $\boldsymbol{\omega}$ can be expressed as,

$$p(\mathbf{z}|\boldsymbol{\omega}) = \prod_{n=1}^{n}\prod_{j=1}^{k} \omega_j^{z_{ij}}. \tag{4.2.6}$$

Therefore, the conditional distribution of observations $\mathbf{x}$ given latent variables $\mathbf{z}$ and Normal component parameters $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is given by,

$$p(\mathbf{x}|\ \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{i=1}^{n}\prod_{j=1}^{k} \mathcal{N}(x_i|\ \mu_j, \Lambda_j^{-1})^{z_{ij}}.$$

We can also write down the full joint distribution as,

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\mathbf{z}|\boldsymbol{\omega})\pi(\boldsymbol{\omega})\pi(\boldsymbol{\mu}|\boldsymbol{\Lambda})\pi(\boldsymbol{\Lambda}). \tag{4.2.7}$$

Given the configuration z_i indicates which component x_i comes from, the marginal density of $f(\mathbf{x})$ can be derived by

$$f(\mathbf{x}) = \sum_{\mathbf{z}} \int_{\boldsymbol{\Lambda}} \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\omega}} \pi(\boldsymbol{\omega})\pi(\boldsymbol{\mu})\pi(\boldsymbol{\Sigma}) \prod_{i=1}^{n} \left(\prod_{j=1}^{k} \omega_j^{z_{ij}} \ \mathcal{N}(x_i| \ \mu_j, \Lambda_j^{-1})^{z_{ij}} \right) \ \mathrm{d}\boldsymbol{\omega}\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\Lambda}.$$

For each observation x_i, we only need to marginalise over parameters from one component with $z_{ij} = 1$, which greatly reduces the required computational effort.

We now consider a variational distribution which factorises between the latent variables and the parameters, such as,

$$q(\mathbf{z}, \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{z})q(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda}). \tag{4.2.8}$$

Updating configuration variables z.

According to the joint density stated in (4.2.7), the posterior of $\mathbf{z}$ given observations $\mathbf{x}$ is proportional to

$$p(\mathbf{z}|\mathbf{x}, \boldsymbol{\omega}) \propto p(\mathbf{z}|\boldsymbol{\omega}) \ p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}).$$

The optimal solution for the variational inference of $q^*(\mathbf{z})$ can be obtained by following the general result derived in Equation (4.2.3). By collecting those terms related to $\mathbf{z}$ and absorbing the rests to the normalised constant, we obtain the updating equation of latent variables as,

$$\log q^*(\mathbf{z}) = \mathbb{E}_{\boldsymbol{\omega}}[\log p(\mathbf{z}|\boldsymbol{\omega})] + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[\log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})]. \tag{4.2.9}$$

The first expectation in Equation (4.2.9) is derived by using the conditional distribution $p(\mathbf{z}|\boldsymbol{\omega})$ in Equation (4.2.6),

$$\mathbb{E}_{\boldsymbol{\omega}}[\log p(\mathbf{z}|\boldsymbol{\omega})] = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\ \mathbb{E}[\log \omega_j].$$

Since observations $\{x_i\}_{i=1}^n$ follow a mixture of Normal distributions, the second expectation in Equation (4.2.9) is expressed as,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}[\log p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})] \\
&\quad = \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log\left(\left((2\pi)^{-\frac{D}{2}}|\Lambda_j^{-1}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x_i-\mu_j)^T\Lambda_j(x_i-\mu_j)\}\right)^{z_{ij}}\right)\right] \\
&\quad = z_{ij}\left(-\frac{D}{2}\log(2\pi)+\frac{1}{2}\mathbb{E}\left[\log|\Lambda_j|\right]-\frac{1}{2}\mathbb{E}_{\mu_j,\Lambda_j}[(x_i-\mu_j)^T\Lambda_j(x_i-\mu_j)\right]\right).
\end{aligned}$$

Therefore, the updating equation for $q(\mathbf{z})$ is obtained as

$$\log q^*(\mathbf{z}) = \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\log\rho_{ij} + \text{const}, \tag{4.2.10}$$

by defining

$$\log\rho_{ij} = \mathbb{E}[\log\omega_j] - \tfrac{D}{2}\ln(2\pi) + \tfrac{1}{2}\mathbb{E}[\log|\Lambda_j|] - \tfrac{1}{2}\mathbb{E}_{\mu_j,\Lambda_j}[(x_i-\mu_j)^T\Lambda_i(x_i-\mu_j)]$$

$$\begin{aligned}
\mathbb{E}[\log\omega_j] =& \psi(\alpha_j) - \psi\left(\sum_j \alpha_j\right) \\
\mathbb{E}[\log|\Lambda_j|] =& \sum_{d=1}^{D}\psi\left(\frac{s_j+1-d}{2}\right) + D\log 2 + \log|S_j| \\
\mathbb{E}_{\mu_j,\Lambda_j}[(x_i-\mu_j)^T\Lambda_j(x_i-\mu_j)] =& D\tau_j^{-1} + s_j(x_i-m_j)^T S_j(x_i-m_j),
\end{aligned}$$

where ψ denotes the *digamma* function, which is defined as the log derivative of the gamma function, e.g. $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log(\Gamma(x))$. Parameters α_j are subject to the constraint $0 < \alpha_j < 1$ in order to ensure that the distribution can be normalised. Taking the exponential on both sides of $\log q^*(\mathbf{z})$ in Equation (4.2.10), it can be standardised as

$$q^*(\mathbf{z}) \propto \prod_{i=1}^{n} \prod_{j=1}^{k} \rho_{ij}^{z_{ij}}$$
$$= \prod_{i=1}^{n} \prod_{j=1}^{k} r_{ij}^{z_{ij}},$$

where $r_{ij} = \rho_{ij} / \sum_{j=1}^{k} \rho_{ij}$. For a latent variable $z_i = \{z_{ij}\}_{j=1}^{k}$, each z_{ij} is a bivariate variable and they sum to 1 such as $\sum_{j=1}^{k} z_{ij} = 1$. Therefore, the optimal solution $q^*(z_i)$ is given by,

$$\mathbb{E}[z_{ij}] = r_{ij}. \tag{4.2.11}$$

Back to our proposed variational distribution stated in Equation (4.2.8), the second term on the right hand side can be further factorised as

$$q(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\omega}) \prod_{j=1}^{k} q(\mu_j, \Lambda_j).$$

Using the general result in Equation (4.2.3), we have

$$\begin{aligned}\log q^*(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\ &= \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{z}|\boldsymbol{\omega})] + \log p(\boldsymbol{\omega}) + \sum_{j=1}^{k} \sum_{i=1}^{n} \mathbb{E}[z_{ij}] \log \mathcal{N}(x_i| \ \mu_j, \Lambda_j^{-1}) \\ &\quad + \sum_{j=1}^{k} \log p(\mu_j, \Lambda_j) + \text{const.}\end{aligned}$$

Updating weight parameters $\boldsymbol{\omega}$.

To obtain the optimal solution of ω, we gather related terms to the following equation,

$$\begin{aligned}\log q^*(\boldsymbol{\omega}) &= \mathbb{E}_{\mathbf{z}}[\log p(\mathbf{z}|\boldsymbol{\omega})] + \log p(\boldsymbol{\omega}) \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} r_{ij} \log \omega_j + (\alpha_0 - 1) \sum_{j=1}^{k} \log \omega_j + \text{const} \\ &= \sum_{j=1}^{k} \log \omega_j \left(\alpha_0 - 1 + \sum_{i=1}^{n} r_{ij} \right),\end{aligned}$$

where $n_j = \sum_{i=1}^{n} r_{ij}$ summarises the number of observations allocated in the j^{th} component. Taking the exponential on both sides, we have

$$q^*(\boldsymbol{\omega}) = \prod_{j=1}^{k} \omega_j^{\alpha_0 - 1 + n_j},$$

which is recognised a Dirichlet distribution

$$q^*(\boldsymbol{\omega}) \sim \mathrm{Dir}(\boldsymbol{\omega}|\boldsymbol{\alpha}), \tag{4.2.12}$$

where $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^{k}$ with $\alpha_j = \alpha_0 + n_j$.

Updating component parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$.

The the optimal solution of variational posterior of $q^*(\mu_j, \Lambda_j)$ for the j^{th} component can be derived as,

$$\log q^*(\mu_j, \Lambda_j) = \sum_{i=1}^{n} \mathbb{E}[z_{ij}] \log \mathcal{N}(x_i|\ \mu_j, \Lambda_j^{-1}) + \sum_{i=1}^{n} \log p(\mu_j, \Lambda_j) + \text{const.}$$

The component parameters can be represented as $q^*(\mu_j, \Lambda_j) = q^*(\mu_j|\ \Lambda_j) q^*(\Lambda_j)$. Hence, the log variational posterior can be derived as,

$$\begin{aligned}
\log q^*(\mu_j, \Lambda_j) &= \log q^*(\mu_j|\ \Lambda_j) + \log q^*(\Lambda_j) \\
&= \sum_{i=1}^{n} \mathbb{E}[z_{ij}] \log \mathcal{N}(x_i|\ \mu_j, \Lambda_j^{-1}) + \log \mathcal{N}(\mu_j|\ m_j, (\tau_j \Lambda_j)^{-1}) \\
&\quad + \log \mathcal{W}(\Lambda_j|\ v_j, S_j).
\end{aligned}$$

By keeping terms relevant to μ_j, the log conditional posterior distribution of μ_j given Λ_j can be written as

$$\begin{aligned}
\log q^*(\mu_j|\Lambda_j) = -\frac{1}{2} &\left[\mu_j^T \left(\tau_0 + \sum_{i=1}^{n} \mathbb{E}[z_{ij}] \right) \Lambda_j \mu_j + \mu_j^T \Lambda_j \left(\tau_0 m_0 + \sum_{i=1}^{n} \mathbb{E}[z_{ij}] x_i \right) \right] \\
&+ \text{const.}
\end{aligned}$$

Taking the exponential on both sides, the posterior of $q^*(\mu_j)$ follows a Normal distribution

$$q^*(\mu_j|\Lambda_j) \sim \mathcal{N}(\mu_j|\ m_j, \tau_j\Lambda_j), \tag{4.2.13}$$

where $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^n r_{ij}x_i$, $\tau_j = \tau_0 + n_j$ and $m_j = \frac{1}{\tau_j}(\tau_0 m_0 + n_j\bar{x}_j)$.

Similarly, by gathering terms related to Λ_j, we obtain the log conditional posterior distribution of $q^*(\Lambda_j)$ as,

$$\begin{aligned}
\log q^*(\Lambda_j) &= \log q^*(\mu_j, \Lambda_j) - \log q^*(\mu_j|\Lambda_j) \\
&= \frac{v_0 - D - 1 + \mathbb{E}[z_{ij}]}{2}\log|\Lambda_j| - \frac{1}{2}\mathrm{Tr}\left[\Lambda_j(S_0^{-1} + \tau_0(\mu_j - m_0)(\mu_j - m_0)^T\right. \\
&\quad \left. + \sum_{i=1}^n \mathbb{E}[z_{ij}](x_i - \mu_j)(x_i - \mu_j)^T - \tau_j(\mu_j - m_j)(\mu_j - m_j)^T)\right].
\end{aligned}$$

Since Λ_j is independent of μ_j, the expression can be further simplified until all terms related to μ_j are cancelled out. Eventually, the posterior of $q^*(\Lambda_j)$ follows a Wishart distribution,

$$q^*(\Lambda_j) \sim \mathcal{W}(\Lambda_j|\ s_j, S_j), \tag{4.2.14}$$

where
$$\begin{aligned}
s_j =& s_0 + n_j, \\
S_j^{-1} =& S_0^{-1} + n_j W_j + \frac{\tau_0 n_j}{\tau_j}(\bar{X}_j - m_0)(\bar{x}_j - m_0)^T, \\
W_j =& \frac{1}{n_j}\sum_{i=1}^n r_{ij}(x_i - \bar{x}_j)(x_i - \bar{x}_j)^T.
\end{aligned}$$

The algorithm of **Variational Bayes** can be described as

Step 1. Initialise the prior parameter values and randomly assign each observation to a prefixed number of mixture components.

Step 2. Estimate the model parameters $(\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ using Equations (4.2.12), (4.2.13), (4.2.14). Find the allocation probabilities for each observation based on the updated parameters using Equation (4.2.11).

Step 3. Iteratively perform **Step 2** until convergence reached.

Updating the ELBO.

The convergence of the Variational Bayes algorithm can be assessed by applying the evidence lower bound that we have mentioned in (4.2.1),

$$
\begin{aligned}
\mathrm{ELBO}(q) &= \sum_{\mathbf{z}} \int\int\int q(\mathbf{z},\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Lambda}) \log\left\{\frac{p(\mathbf{X},\mathbf{z},\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Lambda})}{q(\mathbf{z},\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Lambda})}\right\} \mathrm{d}\boldsymbol{\omega}\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\Lambda} \\
&= \mathbb{E}[\log p(\mathbf{X},\mathbf{z},\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Lambda})] - \mathbb{E}[\log q(\mathbf{z},\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Lambda})] \\
&= \mathbb{E}[\log p(\mathbf{X}|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})] + \mathbb{E}[\log p(\mathbf{z}|\boldsymbol{\omega})] + \mathbb{E}[\log p(\boldsymbol{\omega})] + \mathbb{E}[\log p(\boldsymbol{\mu},\boldsymbol{\Lambda})] \\
&\quad - \mathbb{E}[\log q(\mathbf{z})] - \mathbb{E}[\log q(\boldsymbol{\omega})] - \mathbb{E}[\log q(\boldsymbol{\mu},\boldsymbol{\Lambda})].
\end{aligned}
$$

Individual expectations are expressed as the following:

$$
\begin{aligned}
\mathbb{E}[\log p(\mathbf{x}|\mathbf{z},\boldsymbol{\mu},\boldsymbol{\Lambda})] &= \frac{1}{2}\sum_{j=1}^{k} n_j\{\mathbb{E}[\log|\Lambda_j|] - D\tau_j^{-1} - s_j\mathrm{Tr}[W_j S_j] \\
&\quad - s_j(\bar{x}_j - m_j)^T S_j(\bar{x}_j - m_j) - D\log(2\pi)\}, \\
\mathbb{E}[\log p(\mathbf{z}|\boldsymbol{\omega})] &= \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\mathbb{E}[\log\omega_j], \\
\mathbb{E}[\log p(\boldsymbol{\omega})] &= \log C(\alpha_0) + (\alpha_0 - 1)\sum_{j=1}^{k}\mathbb{E}[\log\omega_j], \\
\mathbb{E}[\log p(\boldsymbol{\mu},\boldsymbol{\Lambda})] &= \frac{1}{2}\sum_{j=1}^{k} n_j\left\{D\log(\frac{\tau_0}{2\pi}) + \mathbb{E}[\log|\Lambda_j|] - \frac{D\tau_0}{\tau_j}\right. \\
&\quad \left. -\tau_0 v_j(m_j - m_0)^T S_j(m_j - m_0)\right\} + k\log B(S_0, s_0) \\
&\quad + \frac{s_0 - D - 1}{2}\sum_{j=1}^{k}\mathbb{E}[\log|\Lambda_j|] - \frac{1}{2}\sum_{j=1}^{k} s_j\mathrm{Tr}[S_0^{-1}S_j], \\
\mathbb{E}[\log q(\mathbf{z})] &= \sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}\log r_{ij}, \\
\mathbb{E}[\log q(\boldsymbol{\omega})] &= \sum_{j=1}^{k}(\alpha_j - 1)\mathbb{E}[\log\omega_j] + \log C(\boldsymbol{\alpha}), \\
\mathbb{E}[\log q(\boldsymbol{\mu},\boldsymbol{\Lambda})] &= \sum_{j=1}^{k}\left\{\frac{1}{2}\mathbb{E}[\log|\Lambda_j|] + \frac{D}{2}\log\left(\frac{\tau_j}{2\pi}\right) - \frac{D}{2} - \mathrm{H}[q(\Lambda_j)]\right\},
\end{aligned}
$$

where $\mathrm{H}[q(\Lambda_j)]$ is the entropy of the Wishart distribution

$$\mathrm{H}[\Lambda_j] = -\log B(S,s) - \frac{s-D-1}{2}\mathbb{E}[\log|\Lambda_j|] + \frac{sD}{2},$$

and the coefficients $C(\boldsymbol{\alpha})$ and $B(S,v)$ are given by

$$\begin{aligned} C(\alpha) =& \frac{\Gamma(\sum_{j=1}^{k}\alpha_j)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)}, \\ B(S,s) =& |S|^{-s/2}\left(2^{sD/2}\pi^{D(D-1)/4}\prod_{d=1}^{D}\Gamma\left(\frac{s+1-d}{2}\right)\right)^{-1}. \end{aligned}$$

As previously mentioned, the values of the evidence lower bound should monotonically increases, and the difference of ELBO at two consecutive iterations can be used to assess the convergence. If the difference is less than a pre-fixed criterion, then the Variational Bayes is claimed to have converged.

4.3 Hybrid Variational Bayes

Fitting a Normal mixture model via the standard VB algorithm introduced in Section 4.2 involves choosing a initial value for the number of components k. In most cases, the total number of components is unknown. In order to obtain a sensible estimate of the number of components, we can execute VB with different choices of $k = 1, 2, \ldots,$ and collect the corresponding values of the ELBO for each k. Since a higher ELBO indicates a better fitting of the data, a sensible value k^* can be determined by the resulting mixture model with the largest ELBO. Obviously, this approach could be computationally expensive if VB converges slowly.

Another commonly recognised drawback of VB is the so-called local maxima problem. The problem has been well studied in the Expectation-Maximisation (EM) algorithm proposed by Dempster et al. [1977]. It is shown that the EM algorithm always converges to a local maximum, but this local maximum is not guaranteed to be the same as the global maximum. Similarly, the VB algorithm, which can be treated as an extension of the EM algorithm, often results in a suboptimal solution rather than the global optimal solution. In the case of mixture models, the local maxima problem arises when there are too many Normal components in one area of the space and too few in another. For the standard VB algorithm introduced previously, it is not possible to move the component from the overestimated region to the underestimated region without passing through areas with low values of the marginal likelihood.

One way to alleviate the local maxima problem is to run the standard VB algorithm for multiple times with randomly generated initial values of parameters. Then we can pick the one with the largest value of ELBO as the final density estimate. Similar to the model selection problem mentioned above, when the number of repeats increases, it is more likely to find a global maximum, but, the computational efforts also increases. Therefore, for datasets with relatively slow convergence rates, researchers usually adopt the *multiple restart* strategy. This involves running VB for multiple randomly generated initial parameter values for a fixed number of steps, then the one with the highest evidence lower bound is chosen and that VB is continued until convergence.

In this section, we propose a hybrid Variational Bayes (HVB) algorithm to estimate the posterior of a mixture model. The HVB algorithm follows a greedy searching scheme which explores every possible model complexity. Hence, the optimal number of components k^* is decided by the data with minimum restrictions. The proposed greedy searching scheme incorporates the component elimination property of VB. Meanwhile, the local maxima problems can be made less likely by applying two types of splitting operations.

4.3.1 Component-Elimination Property

The Variational Bayes algorithm assumes that the number of components k is fixed at the start of the run. Instead of repeating VB with different choices of k, an alternative approach is to over-fit a relatively large number of components as a starting value, and allow VB to converge to a mixture density with a suitable value of k. This mechanism is known as the *component-elimination property* of VB. Basically, if some of the components fall in the same region in the data space during maximisation of the lower bound, then there is a strong tendency in the model to eliminate those redundant components once the data in the region are sufficiently described by fewer components.

According to Attias [1999], this important component-elimination property protects VB away from the well-known problem with the ordinary EM algorithm for mixture models. For instance, if a component is centred at only one observation, the component covariance will equal zero and the model likelihood will be infinity. Thus, the wrong model might achieve a higher likelihood than the correct one. The component-elimination property of VB can automatically rule out this problem by assign a prior distribution on the weight parameter. Moreover, Attias [1999] stated that if the number of observations assigned to a component is one or less, the posterior mean of the mixing weight of that component is zero, which means it will be automatically eliminated.

To illustrate the component-elimination property, recall that the difference between $\log p(\mathbf{x})$ and evidence lower bound can be measured by the Kullback-Leibler divergence between the variational distribution $q(\boldsymbol{\theta})$ and marginal likelihood $p(\boldsymbol{\theta}|\mathbf{x})$,

$$\mathrm{KLD}(q||p) = \log p(\mathbf{x}) - \mathrm{ELBO} = \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{x})} \mathrm{d}\boldsymbol{\theta},$$

where parameters $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ include mixing weights and component parameters. Assuming the chosen conjugate priors can be factorised over the number of components, the KLD is calculated by the sum of the number of components,

$$\int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{x})} \mathrm{d}\boldsymbol{\theta} = \sum_{j=1}^{k} \int q(\theta_j) \log \frac{q(\theta_j)}{p(\theta_j|\mathbf{x})} \mathrm{d}\theta_j.$$

Adopting the mean-field approximation, which implies $q(\theta) = q(\omega)q(\mu, \Lambda)$, each term related to θ_j on the right-hand side of the equation can be written as

$$\begin{aligned}&\int q(\theta_j)\log\frac{q(\theta_j)}{p(\theta_j|\mathbf{x})}\mathrm{d}\theta_j\\ &\quad=\int q(\omega_j)\log\frac{q(\omega_j)}{p(\omega_j|\mathbf{x})}\mathrm{d}\omega_j+\int\int q(\mu_j,\Lambda_j)\log\frac{q(\mu_j,\Lambda_j)}{p(\mu_j,\Lambda_j|\mathbf{x})}\mathrm{d}\mu_j\mathrm{d}\Lambda_j.\end{aligned}$$

We take mixing parameter ω_j as an example. Imagine there are only a few observations allocated to the j^{th} component, the posterior of variational distribution $q(\omega_j)$ follows a Dirichlet distribution $\mathrm{Dir}(\omega_j|\alpha_j)$ as stated in Equation (4.2.12), where $\alpha_j = \alpha_0 + \sum_{i=1}^{n} n_j$, α_0 is a constant and $n_j = \sum_{i=1}^{n} r_{ij}$ counts the number of observations in the j^{th} component. Then, the posterior distribution of $q(\omega_j)$ is dominated by the prior $p(\omega_j)$, which leads the value of $\int q(\omega_j)\log(q(\omega_j)/p(\omega_j|\mathbf{x}))\mathrm{d}\omega_j$ to zero. Similar results hold for the rest terms related to μ_j and Λ_j. Therefore, as the number of data that are available to the j^{th} component tends to zero, the corresponding KLD also tends to zero. Consequently, the KLD is decreased by removing a component that covers few observation, and the ELBO increases as the component is eliminated.

Corduneanu and Bishop [2001] introduced an approach which estimates the number of components by optimising the mixing weight through a type-2 maximum likelihood. The unwanted components are removed if their weights are less than 10^{-5}. They also noticed the importance of initial choice of mixture parameters, and proposed to use the K-means clustering to set the initial means. The standard VB algorithm will lead to suboptimal models if the initial number of components is chosen smaller than required. Thus, a natural approach for addressing the model selection problem is to fit a mixture model initialised with a large number of components to start off. In general, such a strategy is not recommended for other types of approaches but VB, because the following reasons: first, the VB approximation has a low requirement on computational resources compared to sampling approaches. This is true even for repeating multiple times. Second, the VB algorithm does not suffer from the label-switching problem. Third, the component-elimination property enhance the possibility of finding the global optimal solution. With the component-elimination property, redundant components are removed as VB converges to a solution, thereby determining model complexity automatically. However, this tactic is not computational efficient when an arbitrarily large number of components is initialised. Hence, it is useful to perform a preliminary investigation of the dataset to have some sense of the number of components needed.

4.3.2 Greedy Searching Scheme

For estimating the unknown structure of mixture density, the component-elimination property of VB suggests a natural *overfitting* strategy for the determination of the number of component. However, VB is shown sensitive to the initialisations of the hyperparameters and the initial allocations of observations. It is possible that the VB converges to a suboptimal model complexity as a result of the initialisation choices. In addition, the eliminating operation is not reversible during VB updating procedure. Hence, the overfitting strategy can not guarantee the detection of the global optimum in a single run.

In this section, we introduce a VB algorithm with greedy searching scheme to resolve the model selection problem and the VB does not depend on the initialisation. It is basically an incremental method which attempts to split all k existing components, and stops if there is no further improvement of the evidence lower bound. Splitting tests are applied to each existing component, which will be described in Section 4.3.3. Once a component is qualified to be split, an additional component is generated, and the observations in that component are re-allocated into those two subcomponents. Then, the VB updating equations are applied to the new mixture model starting with $k + 1$ components. If VB converges with a higher value of the ELBO, we keep the resulting mixture density, which implies the existence of more than one components for the testing data region; otherwise, we return to the initial structure with k mixtures and continue searching for the next potential candidate.

Given VB converged to a mixture model with k component and a value of ELBO, the greedy searching scheme can be briefly described as follows.

Greedy searching scheme of the HVB algorithm

Step 1. Perform the splitting tests for each existing component j for $j = 1, \ldots, k$. Stop testing once one component is found satisfying the splitting criteria.

Step 2. Observations allocated in the j^{th} component are re-assigned into two subcomponents, while allocations in other components are kept fixed.

Step 3. Run VB updating equations for this new mixture model with $k + 1$ components until convergence. Evaluate the value of ELBO for the new model.

Step 4. If new value of ELBO is larger than the previously obtained one, we keep the new structure of mixture density. Otherwise, we return to the original model and move on to the next component.

Step 5. Repeat **Step 2 - 4** until either no further component satisfy the testing criteria, or no further improvement of ELBO can be achieved.

For the first step of searching candidate components to be split, we performs two types of splitting tests with orthogonal directions, and those two types of splitting operations are applied to the component highlighted in **Step 2**. Details of splitting tests with specific criteria and corresponding splitting operations are described in the next section. The greedy searching scheme can be viewed as a driving force for the VB to explore other possible model structures, and thus increases the chance of finding a global maximum posterior of a mixture model.

4.3.3 Two Splitting Directions

We incorporate two types of splitting tests and two splitting operations in the greedy searching scheme, which allows us to have a better description for different regions of data, and thus leads to a more accurate estimate of the mixture density.

The first one is referred as the *mean split* operation, which splits a component into two side-by-side subcomponents with unequal means $\mu_j \neq \mu_{j+1}$. The type of splitting operation is commonly applied in mixture density estimation, see Ueda et al. [2000] and Constantinopoulos and Likas [2007] as examples.

Additionally, we also introduce the *variance split* operation, which splits a component into two overlapping subcomponents such that $\|\Sigma_j\| \ll \|\Sigma_{j+1}\|$ with similar means $\mu_j \simeq \mu_{j+1}$. For heterogeneous multi-modal pattern, some regions of data can be better described by a mixture of a concentrating "peaky" component and a relative "flat" component, rather than fitting one single Normal distribution. A component is split into two subcomponents, *inners* and *non-inners*, in situations where there exists a high concentration of observations close to the component mean. Next, we introduce the mean and the variance split in details.

Following the first step of the greedy searching scheme, we first run standard VB with a overly fitted number of components. Based on the component-elimination property, VB converged to a local maximum of mixture density with a number of component k^*. The corresponding component parameters are updated to $\{\omega_j^*, \mu_j^*, \Lambda_j^*\}_{j=1}^{k^*}$ according to the allocation matrix Z^*. The allocation matrix Z^* is formed by $n \times k^*$ elements, $z_{ij} = 1$ implies the i^{th} observation is allocated to the j^{th} component, and $\sum_{j=1}^{k^*} z_{ij} = 1$ for each observation.

Next, we attempt to examine all k^* number of existing mixing components by two splitting operations.

Mean split

The idea of mean split is to identify a component which covers two or more groups of data, then split the observations in the highlighted component into two subcomponents. For those observations which are less than the component mean, they will be re-allocated to the left-hand side subcomponent, and the rest observations are automatically moved to the right-hand side subcomponent.

In general, the mean split operation splits components into two side-by-side subcomponents, without considering the direction of split for high-dimensional datasets. For the considered multivariate dataset, we apply principal component analysis (PCA) to determine the directions of the mean split, where PCA is an orthogonal transformation to convert linearly correlated variables into a set of uncorrelated principal components. For D-dimensional observations, we can transform the component variance-covariance parameters into a vector containing D eigenvalues. It allows us to assess the variation for the component according to the weight of the largest variance. If there is too much variation on the major direction, it suggests that we might have a better fit with more than one component for that data region.

For the $D \times D$ covariance matrix $\Sigma_j^* = \Lambda_j^{-1*}$ of the j^{th} mixing component, we can decompose it into D principle components $(p_1, \ldots, p_D)'$, where p_1 with the largest eigenvalue λ_1 represents as much of the data variation as possible. The component j is highlighted by the mean splitting test if

$$\frac{\lambda_1}{\sum_{d=1}^{D} \lambda_d} > c_1 \quad \text{and} \quad \sigma_j^{\max} > c_2,$$

where $\sigma_j^{\max} = \sqrt{\max_{\{1:D\}}\{\text{diag}(\Sigma_j)\}}$ represents the largest standard deviation of the j^{th} component of interest and $\text{diag}(\Sigma_j)$ denote the diagonal elements of the covariate matrix Σ_j. A component is considered as suitable to split to two subcomponents, if the data variation along p_1 is larger than c_1, and the largest value of the standard deviations $\sigma_j^{\max}$ is greater than c_2, where constants c_1 and c_2 are carefully chosen according to each specific dataset. It is worth noting that we should chose the value of c_1 reasonably large, as this reflects a large difference in the eigenvalue ratio which usually implies a poorly fitted component. Meanwhile, if we chose a relative small value of c_2, it means that the component would be split more likely, and hence leads to a higher number of mixtures.

After identifying the component by the mean splitting test, we project all observations in the flagged component on the p_1 axis, and calculate the sample mean m_{p_1} on the major splitting direction p_1. Then, we can re-assign observations to either left or right hand side subcomponents according to m_{p_1}. Hence, the number of columns in the allocation matrix Z^* becomes $k^* + 1$, and the corresponding component parameters can be determined by the sample mean and the sample covariance,

$$\mu_{k^*+1} = \frac{1}{n_{k^*+1}} \sum_{i=1}^{n} x_{z_{i,k^*+1}=1},$$
$$\Lambda_{k^*+1} = \mathrm{cov}(\{x_{z_{i,k^*+1}=1}\}_{i=1}^{n})^{-1},$$

where n_{k^*+1} denotes the number of observations in the new component $k^* + 1$.

This newly obtained allocation matrix will be used as the starting point to proceed the VB until convergence. If the resulting value of ELBO of the new model is larger than the previous one, we keep the resulting mixture density; otherwise, we return to the original mixture model and perform the mean splitting test to the next component.

Variance split.

In contrast to the mean split, we also examine each component for the variance splitting test. For a component which is highlighted to be variance split, we identify those observations which are concentrated in the central area as *inners*, and the rest observations which are scattered around as *non-inners*. Then, we split the highlighted component into two overlapping subcomponents with one peaky subcomponent and one broad subcomponent.

We identify inner observations using the Mahalanobis distance (MD), which measures the distance between an observation x and a distribution P. The distance becomes zero if x is at the mean of P and increases as x moves away from the mean of P. For high-dimensional data, it measures the number of standard deviations from x to the mean of P along each principal component axis. For the considered case, assuming an observation x_i is allocated to the j^{th} Normal component, the distance $\mathrm{MD}_i^{(j)}$ is given by

$$\mathrm{MD}_i^{(j)} = \sqrt{(x_i - \mu_j^*)^T \Lambda_j^* (x_i - \mu_j^*)},$$

where (μ_j, Λ_j) are component mean and precision parameters. We can compare the value of $\mathrm{MD}_i^{(j)}$ with the theoretical results from a chi-square distribution with D degree of freedom, e.g. $(\mathrm{MD}_i^{(j)})^2 \sim \chi^2_{df=D}$. This relationship exists when x_i is assumed to be normally distributed. Hence, an observation is identified as an inner if

$$\mathrm{MD}_i^{(j)} < \sqrt{\chi^2_{df=D,c_3}}, \tag{4.3.1}$$

where c_3 represents the cumulative probability of the area under the curve of the chi-square distribution. Typically, a small value of c_3 is preferred, as we only collect inner observations concentrating to the component mean. Meanwhile, if the value of c_3 is chosen to be too close to zero, the algorithm may miss some potential inner observations since the observations are not close enough to the fitted component mean.

We consider a component to have a better description with two overlapping subcomponents, if

$$n_j^{\text{inners}}/n_j > c_4,$$

where n_j^{inners} and n_j are the number of identified inners according to condition (4.3.1), and the total number of observations in the j^{th} component. The value of c_4 is chosen as a probability such that $0 < c_3 < c_4 < 1$. When the inners proportion c_4 is larger than the theoretical expected proportion c_3, that means the actual number of observations that lie in the centre region of the component is larger than the expected number. Therefore, we highlight the j^{th} component for a variance split.

Once a component is flagged as suitable to be replaced by two overlapping subcomponents, inner observations will still be allocated in the j^{th} component, and non-inner observations will be re-allocated to the $(k^*+1)^{\text{th}}$ component. Then, the component parameters are initialised as the follows,

$$\begin{aligned} &\mu_{\text{inners}} = \mu_{\text{non-inners}} = \mu_j, \\ &\Lambda_{\text{non-inners}} = \Lambda_j \quad \text{and} \quad \Lambda_{\text{inners}} = c_5 \times \Lambda_j, \end{aligned} \tag{4.3.2}$$

where c_5 is a fixed constant which must be larger than 1. According to Equations (4.3.2), half weights are assigned on both newly proposed subcomponents, and the inner subcomponent is initialised with a higher precision. Meanwhile, the component mean remains the same.

After the variance split, we obtain a new allocation matrix Z^* with $k^* + 1$ number of columns. It will be used as a starting point for applying the VB updating procedure until convergence. We keep the resulting mixture model if the corresponding value of the ELBO is larger than the previous one. Otherwise, we return to the original mixture model and apply the variance splitting test to the next component.

4.3.4 Hybrid Variational Bayes Algorithm

We propose a hybrid Variational Bayes (HVB) algorithm to estimate densities of mixture models by combining all three features mentioned above. The HVB algorithm starts with an overfitting strategy. Fixing a relatively large number of component, VB converges to a local maximum model estimates. Even though such a local maximum is not guaranteed to be the global optimal solution, it provides a sensible estimate of the number of components k as a initial value for the greedy searching scheme. Then, HVB progressively attempts to split every existing mixture components. Moreover, each component is examined by two types of splits, the mean split and variance split, which potentially provide a more thorough description of the mixture model. The greedy searching scheme is designed to avoid local maxima and to find the global optimal solution.

The hybrid Variational Bayes algorithm will terminate when either no component satisfies any of the split criteria or the ELBO is the same within a tolerance level compared to the previous value. Instead of running standard VB for multiple times, HVB only needs to execute once and it is independent of initialisations. Therefore, HVB effectively reduces the possibility of becoming trapped into a local maximum, and highly likely to deliver a global maximum of the mixture density estimation problem. In summary, the proposed hybrid VB algorithm can be described by the following steps.

Hybrid Variational Bayes algorithm:

Step 1. Set initial number of component k relative large, randomly assign observations to k components and run VB until convergence. Start splitting test from $l_{\text{split}} = 1$.

Step 2. If component l_{split} is highlighted by the variance/mean splitting test, proceed with variance/mean split and get the new allocation matrix. Using the new allocation as a starting point, apply the VB updating equations until convergence. If the new ELBO is larger than the original, keep the resulting mixture model and update all the posterior parameters. If this component does not qualify for both splitting criteria, move on to the next component $l_{\text{split}} = l_{\text{split}} + 1$.

Step 3. Repeat **Step 2** until no further component can pass splitting tests or no further split can improve the ELBO from the current mixture density.

The corresponding pseudo code is given as follows.

Algorithm 1 Hybrid Variational Bayes Algorithm

Set k freely and $\text{ELBO}_{\text{increment}} = 1^{-20}$. Let $l_{split} = 1$.
Run VB with randomly allocation of observations in k components. Obtain an estimated mixture density with k^* components.
while $l_{split} < k^*$ **do**
  Apply the variance split, and obtain a new allocation R.
  Use the new R to run VB updating until convergence is reached.
  if $\text{ELBO}_{\text{new}} - \text{ELBO} > \text{ELBO}_{\text{increment}}$ **then**
    Keep the new mixture density and the corresponding posterior parameters.
    Set $l_{split} = l_{split}$.
  else
    Apply the mean split, and obtain a new allocation R.
    Use the new R to run VB updating until convergence is reached.
    if $\text{ELBO}_{\text{new}} - \text{ELBO} > \text{ELBO}_{\text{increment}}$ **then**
      Keep the new mixture density and the corresponding posterior parameters.
      Set $l_{split} = l_{split}$.
    else
      Set $l_{split} = l_{split} + 1$.
    end if
  end if
end while

4.4 Result Analysis

In this section, we carry out a simulation study to show that the local maxima problems can be highly likely avoided in mixture density estimation by using the hybrid Variational Bayes algorithm proposed in Section 4.3.4, whereas the standard Variational Bayes can be easily caught. In Section 4.4.2, we also demonstrate the HVB algorithm for the industrial application mentioned in Chapter 3. The performance of the proposed HVB algorithm is compared with results from those previously presented MCMC methods. As a fast alternative of the MCMC, the HVB algorithm consistently provides reliable mixture density estimation on both simulated datasets and real-world applications.

4.4.1 Simulation Study

In this section, we apply the proposed hybrid Variational Bayes algorithm to a simulated dataset with multi-modality. One hundred bivariate observations $\mathbf{y} = \{y_i\}_{i=1}^{100}$ are generated from a mixture of 3 Normal distributions,

$$f_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Sigma}}(y_i) = \sum_{j=1}^{3} \omega_j \; \mathcal{N}(y_i|\mu_j, \Sigma_j).$$

Mixing components weights $\boldsymbol{\omega}$ and locations $\boldsymbol{\mu}$ are defined as

$$\boldsymbol{\omega} = (0.3, 0.5, 0.2), \quad \boldsymbol{\mu} = \begin{pmatrix} -0.5 & 0 & 3 \\ 0 & 2 & 2.5 \end{pmatrix},$$

and covariance are given as $\Sigma_1 = 0.02 \times \mathbb{I}_2$, $\Sigma_2 = 0.05 \times \mathbb{I}_2$, $\Sigma_3 = 0.01 \times \mathbb{I}_2$, where $\mathbb{I}_2$ indicates a 2×2 identity matrix. Initial values of parameters stated in Equations (4.2.4) and (4.2.5) are fixed as follows: $\alpha_0 = 1$, $m_0 = \sum_{i=1}^{n} y_i$, $\tau_0 = 1$, $s_0 = 3$ and $S_0 = \mathbb{I}_2$. The HVB terminates when the increment of the ELBO between two consecutive iterations is less than the pre-fixed constant 10^{-20}. HVB finishes in 7 seconds and the curve fitting is plotted in Figure 4.1.

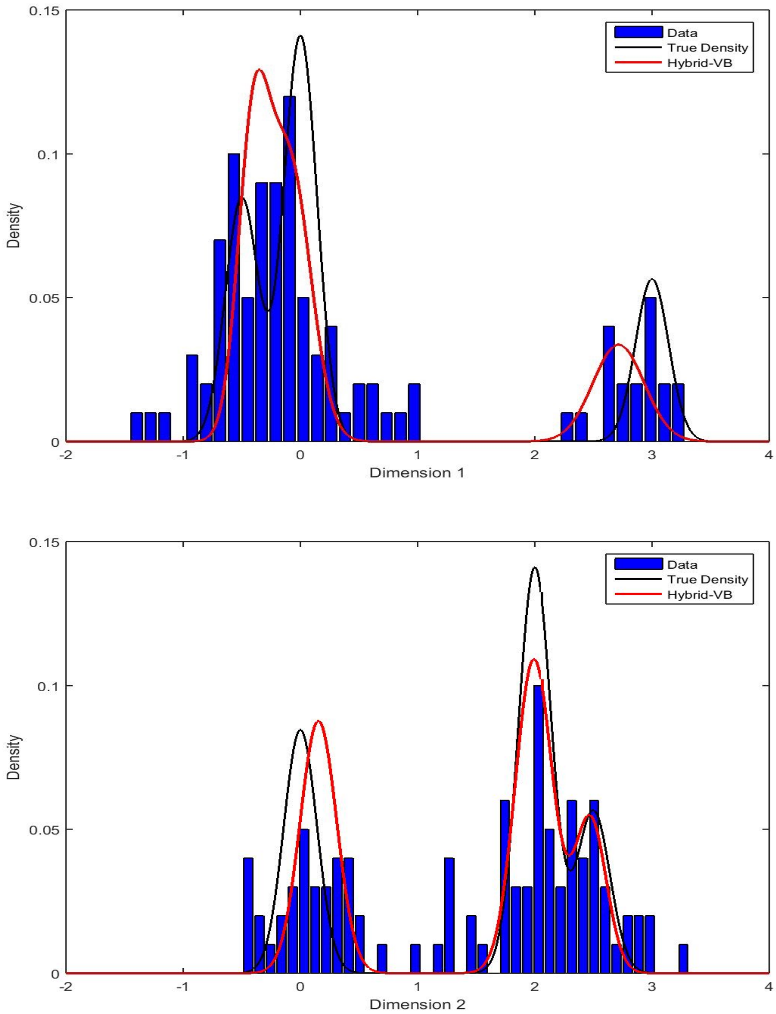


Figure 4.1: Density estimation of a simulated dataset on both dimensions by hybrid Variational Bayes.

The histogram on each dimension shows the locations of the 100 generated observations, the dashed black curve represents the true density of the simulated data, and the solid red curve is the density estimation obtained from the hybrid Variational Bayes algorithm. As shown in Figure 4.1, two close Normal components are generated for each dimension, and HVB correctly catches the multi-modal structure of the simulated dataset. Numerical evidence can be found by comparing the estimated values of mean and weight parameters,

$$\boldsymbol{\omega} = (0.3415, 0.4343, 0.1843) \quad \text{and} \quad \boldsymbol{\mu} = \begin{pmatrix} -0.4018 & -0.09248 & 2.7145 \\ -0.1539 & 1.9923 & 2.4724 \end{pmatrix},$$

which are fairly close to their true values. On the other hand, the standard VB with fixed number of component $k = 3$ is clearly stuck to a local maximum and failed to deliver the global optimal density estimation, as shown in Figure 4.2.

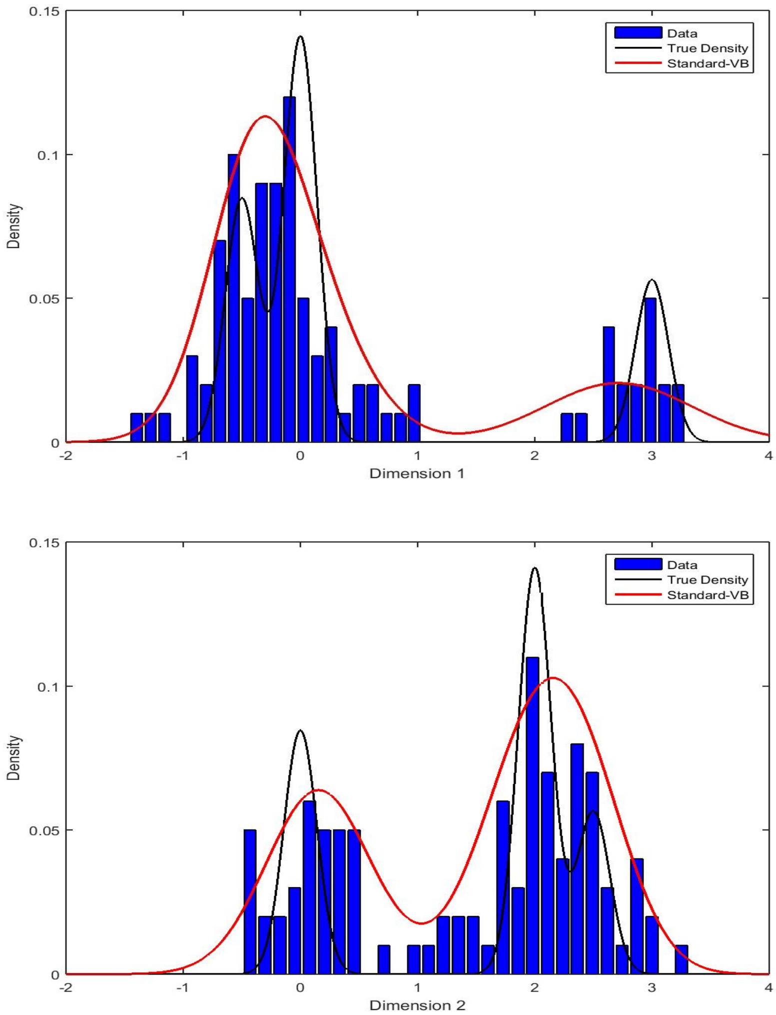


Figure 4.2: Density estimation of a simulated dataset on both dimensions by standard Variational Bayes algorithm.

The estimated values of parameters $(\boldsymbol{\omega}, \boldsymbol{\mu})$ are given by

$$\boldsymbol{\omega} = (0.006, 0.6322, 0.3618) \quad \text{and} \quad \boldsymbol{\mu} = \begin{pmatrix} 0.3309 & -0.2992 & 1.4677 \\ 1.4437 & 0.9496 & 2.2737 \end{pmatrix}.$$

It can be seen from the weight estimates, the first component has a much smaller weight than the other two. This is because the number of components is fixed to be 3 and VB has converged to a local mixture density with only 2 components.

When compared to the true density, the estimates of mixture locations are obviously inaccurate. For standard VB with an unknown number of components, the general approach for this model selection problem is to fit VB with different numbers of components separately. We also fixed $k = 2$ and $k = 4$ to run VB, and both runs highly likely deliver inaccurate density estimates of the true mixture model.

In addition, we compare the values of the Kullback-Leibler divergence from HVB and VB algorithms. The KLD between the estimated density and the true density is calculated by

$$\mathrm{KLD}(f_{\mathrm{est}} \mid\mid g_{\mathrm{true}}) = \int f(y) \log\left(\frac{f(y)}{g(y)}\right) \mathrm{d}y.$$

The KLD from HVB algorithm equals to 0.0053, whereas the KLD from standard VB is calculated as 0.1336 which is 25 times larger. It is worth noting that the number of component k is inferred from HVB but fixed to be 3 for standard VB. According to both numerical and graphical evidence, the proposed hybrid Variational Bayes algorithm performs better than the standard VB algorithm, especially when dealing with multi-modal datasets.

4.4.2 Industrial Application

In Chapter 3, we proposed a Bayesian linear model with Dirichlet Process mixture prior (BL-DPMP) approach to reduce the number of measurements required per sensor. In order to obtain accurate estimates of the critical parameter β for the current sensor with reduced measurements, the historical estimates of β with full measurements is used as prior to compensate for the missed information.

The BL-DPMP approach takes the first $n = 240$ historical OLS estimates $\{\beta_i\}_{i=1}^{n}$ as observations, then estimates the mixture density of β as a finite number of mixture model,

$$f_{\boldsymbol{\omega},\boldsymbol{\mu},\boldsymbol{\Sigma}}(\beta_i) = \sum_{j=1}^{k} \omega_j \, \mathcal{N}(\beta_i | \, \mu_j, \Lambda_j^{-1}),$$

where $\boldsymbol{\omega} = \{\omega_j\}_{j=1}^k$ represent component weights which follows a Dirichlet distribution with initial value of $\alpha_0 = 0.3$. Component parameters $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ follow conjugate Normal and Wishart priors, which are stated in Equation (4.2.5) with initials $\tau_0 = 1$, $m_0 = n^{-1}\sum_{i=1}^n \beta_i$, $s_0 = 15$ and S_0 is the precision matrix of $\{\beta_i\}_{i=1}^n$. The prior value of α_0 is assigned empirically. A discussion about choosing initial value of α_0 can be found in Rousseau and Mengersen [2011], where they distinguished different asymptotic behaviours of posteriors corresponding to different relationships between α_0 and the number of components.

In order to apply the hybrid Variational Bayes algorithm to estimate the mixture density of β, we fix the number of components $k = 30$ at the start of the run. Constants involved in the mean split operation and the variance split operation are given as $(c_1, c_2, c_3, c_4, c_5) = (0.95, 10^{-3}, 0.75, 0.8, 1000)$ as stated in Section 4.3.4. HVB terminates when the ELBO increment between two consecutive iterations is less than 10^{-20}.

We present the resulting observation allocations and corresponding weights of the mixture density in Table 4.1 with a descending order.

Table 4.1: Number of allocations and corresponding weight for each mixing Normal component obtained from HVB of the S1 historical dataset.

Comp.	No. of allocations	Weight	Comp.	No. of allocations	Weight
1	156	0.6504	2	35	0.1458
3	16	0.0667	4	10	0.04167
5	3	0.125	6	2	0.0083
7	2	0.0083	8	2	0.0083
9	1	0.0042	10	1	0.0042
11	1	0.0042	12	1	0.0042
13	1	0.0042	14	1	0.0042
15	1	0.0042			

Table 4.1 shows that the historical observations are mainly clustered into four Normal components, which matches our visual judgement on the histogram plot of $\{\beta_{i,2}\}_{i=1}^n$ shown in Figure 4.3. The red curve successfully captures the multimodal structure of the β historical dataset. The solid line in the plot represents

the curve fitting based on estimated mixture density from HVB.

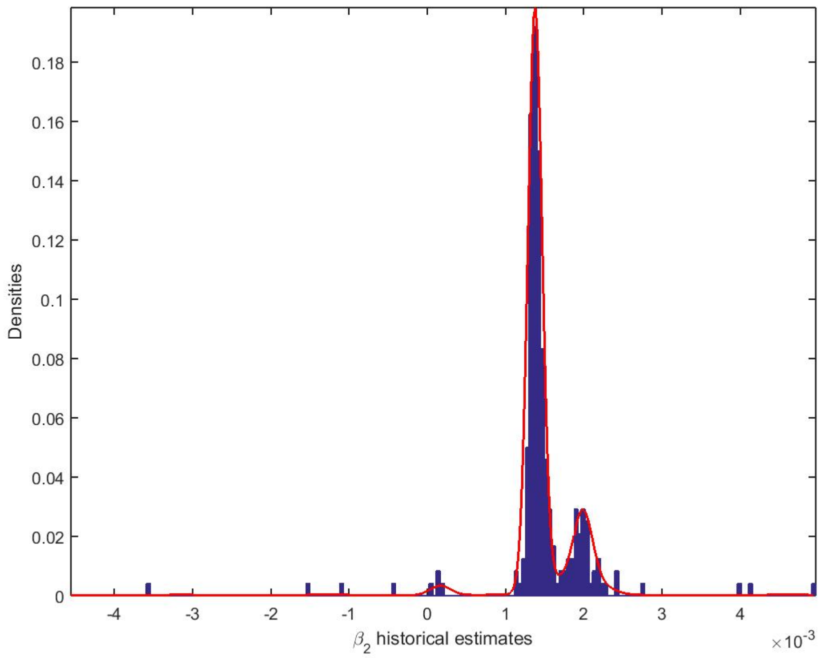


Figure 4.3: Histogram of the historical estimates $\{\beta_{i,2}\}_{i=1}^{240}$ and the curve fitting obtained from the BL-DPMP approach with HVB of the S1 type of sensors.

Next, when a new sensor with reduced measurements (y_t^-, X_t^-) is available, the estimated mixture density of critical parameter β obtained by HVB is used as prior choices for the regression coefficient β_t. The probability of selecting the most suitable Normal component is proportional to

$$\mathbb{P}(\gamma_t = j|\ y_t^-, X_t^-, \omega_j, \mu_j) \propto\ \omega_j\ \mathcal{N}(y_t^-|\ X_t^- \mu_j,\ \sigma^2 \mathbb{I}_{18}).$$

where response variable y_t^- is a 18×1 vector, covariates X_t^- is a 18×5 matrix. Let γ_t denote the allocation of the t^{th} sensor, ω_j and μ_j represent the weight and the mean for the j^{th} Normal component, the value of σ^2 is fixed to be extremely small, and $\mathbb{I}_{18}$ denotes a 18×18 identity matrix. Conditional on $\gamma_t = j$, the prior of β_t is assigned to be,

$$\pi(\beta_t) \sim \mathcal{N}(\beta_t|\ \mu_j, \Sigma_j).$$

Then, the posterior distribution of β_t conditional on $\gamma_t = j$ can be estimated through a Bayesian linear model

$$p(\beta_t|\ y_t^-, X_t^-, \gamma_t = j, \mu_j, \Sigma_j) \propto \pi(\beta_t)\mathcal{N}\left(y_t^-|\ X_t^-\beta_t, \sigma^2\mathbb{I}_{p^-}\right).$$

Hence, the posterior distribution of β_t conditional on γ_t follows a Normal distribution,

$$(\beta_t|\ y_t^-, X_t^-, \gamma_t = j, \mu_j, \Sigma_j)\ \sim \mathcal{N}\left(\beta_t|\ \mu_t, \Sigma_t\right),$$

where

$$\mu_t = \left(\Sigma_j^{-1} + \frac{1}{\sigma^2}X_t^{-T}X_t^-\right)^{-1}\left(\Sigma_j^{-1}\mu_j + \frac{1}{\sigma^2}X_t^{-T}y_t^-\right) \text{ and } \Sigma_t = \left(\Sigma_j^{-1} + \frac{1}{\sigma^2}X_t^{-T}X_t^-\right)^{-1}.$$

Eventually, we can calculate the characterisation error by $\varepsilon_t = y_t - X_t\beta_t$, where (y_t, X_t) represent the complete data with full measurements for the t^{st} sensor, and error term ε_t is a 32×1 vector.

According to the sensors quality specification stated in Equations (3.4.2), we calculate the absolute values of average errors for all temperature bands, and they must lie within a range of $[0, 0.3\%]$ to be a qualified product. In Table 4.2, we summarise the total numbers of failures from 655 sensors with reduced measurements obtained from the standard VB, the proposed HVB and the overfitting strategy of VB, along with previously described MCMC method and OLS approach with full data.

Table 4.2: Comparison performances obtained from OLS, Bayesian Linear approach, BL-DPMP approach with slice sampler, standard VB, HVB and overfitting VB on the incomplete dataset of the S1 type of sensors.

		No. of Failures (655)			Pass rate	Time (sec)
		0.1%	0.2%	0.3%		
OLS	complete	4	3	2	99.69%	0.04
OLS	incomplete	432	290	199	69.62%	0.05
Bayes Linear	incomplete	432	290	199	69.62%	0.1
BL - DPMP	incomplete	162	24	8	98.78%	8486
VB	incomplete	174	127	121	81.53%	4
HVB	incomplete	128	23	11	98.32%	225
Overfitting VB	incomplete	128	23	11	98.32%	496

As we explained in Section 2.4, the OLS method is easy to implement and fast to execute. The OLS estimates of β is accurate enough to meet the company's quality specifications when dealing with full measurements. However, when the mid-range temperatures are removed, the OLS collapses as expected, since it has no information about the missing measurements. The Bayesian linear approach with single Normal prior assumption, proposed by Eves et al. [2015], is not considered to be suitable for dealing with multi-modality datasets in our considered situation as illustrated in Chapter 3.

Using the Bayesian linear model with Dirichlet process mixture prior (BL-DPMP) approach, the resulting number of failures obtained by the slice sampler actually achieves a comparable level of accuracy (98.78%) to the OLS approach with complete data (99.69%). However, the drawbacks of MCMC is that it requires a high computational effort. In addition, the resulting estimated mixture densities are not exactly identical. As fast alternatives to the MCMC type of approaches, we applied standard VB, the proposed hybrid Variational Bayes algorithm and the overfitting strategy of VB to estimate the mixture density of β from the S1 historical dataset.

According to Table 4.2, standard VB gives 121 failures, and the pass rate 81.53% is considered to be too low to be accepted, as standard VB is easily to be caught by a local maximum and cannot move out. The proposed HVB approach provides 11 failures out 655 sensors with reduced measurements, which equals to pass rate 98.32%. Meanwhile, the required computational efforts of HVB is about 36 times less than the MCMC method, even though it costs more than the standard VB due to the greedy searching scheme. HVB only requires a single run to estimate the mixture model with great flexibility on the number of components and prevents local maxima through a greedy searching procedure. More importantly, it consistently provides reliable mixture model estimates and requires much less computational efforts. From a practical point of view, the hybrid Variational Bayes approach is more likely to be applied on the industrial characterisation line.

We also apply the overfitting strategy of VB by repeating the algorithm 100 times with fixed $k = 40$, and other initial values are kept the same as the HVB approach. From the previous investigation of fitting different initial numbers of components, the maximum value of the ELBO is achieved when the initial value of k lies into the range from 12 to 18, and there is no obvious penalty as long as k is chosen below 40. We then selecte the mixture density estimate with the minimum number of failures from 100 repeats.

From the last row of Table 4.2, the overfitting strategy of VB obtains the exact same results as the hybrid Variational Bayes approach. However, overfitting VB consumes twice time as much as HVB. One may argue that it is not necessary to use 100 repeats, in that case, less computational time would be required. In practice, we prefer to make sure that the repeating number is large enough to give a good probability of finding one global optimum, as it is difficult to choose the number of repeating for a specific dataset.

Judging by the number of failures obtained by the overfitting strategy of VB, it seems like to be a valid approach to the considered industrial application. The over fitted mixture model provides an acceptable number of failures. However, it requires multiple runs to achieve the best mixture density, which is computationally wasteful in terms of both time and storage. In the company's point of view, it is better to have a fast algorithm that consistently provides reliable density estimations with high repeatability.

To further illustrate the comparison among OLS, Bayesian linear, BL-DPMP model using slice sampler and the hybrid Variational Bayes approach, we present characterisation errors of one sensor as an example. The sensor is characterised without measurements at mid-range temperatures. The passing range is $\pm 0.3\%$ for average error of each temperature band.

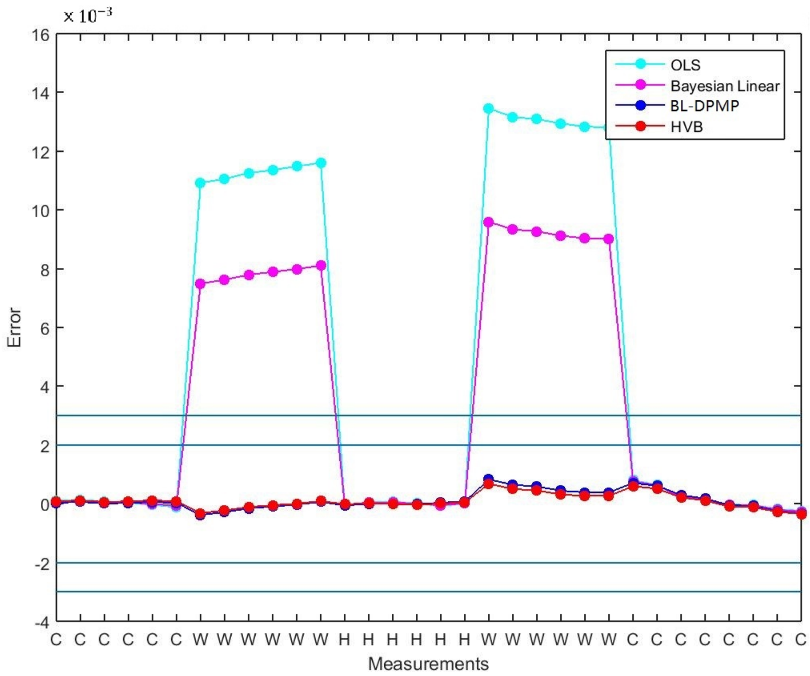


Figure 4.4: Errors comparison for one product among OLS, Bayesian linear, BL-DPMP approach with slice sampler and HVB of the S1 type of sensors.

For the current sensor with 32 measurements and 5 temperature bands, Figure 4.4 shows both OLS (cyan) and Bayesian linear (magenta) approaches failed to meet the company's quality specification. On the other hand, errors from both MCMC (blue) and HVB (red) lie within the $\pm 0.1\%$ error interval. Therefore, HVB successfully deliver an accurate estimate of β and it is considered to be a faster alternative of MCMC approach for estimating mixture densities.

4.5 Discussion

In this chapter, we proposed a hybrid Variational Bayes (HVB) algorithm as a way to estimate a mixture model of Normal distributions. As an alternative to the sampling approaches, the proposed approximation algorithm provides deterministic mixture density estimates and dramatically reduces the computational time. The HVB begins with an overfitting VB algorithm, then follows a greedy searching scheme by examining all existing components of both mean and variances splitting potentials. Overall, the purpose of the greedy searching scheme and two splitting operation introduced in Section 4.3 is to find a better a allocation matrix of observations. Once a new allocation matrix Z^* is proposed, we can calculate the component parameters and run VB to obtain a new value of ELBO. Judging by the increment of ELBO, we then decide to keep the resulting mixture density or not. By exploring every possible model structure, the HVB algorithm avoids becoming trapped in local maxima, and is highly likely to provide the global maximum judging by the increments of the ELBO. Results from both the simulation study and the industrial application support the validity of the proposed HVB approach in terms of numerical and graphical aspects.

On the other hand, there are a couple of concerns about the proposed HVB approach. The greedy searching process is not a very efficient design, because it progressively examines many different models. Running HVB once takes longer than would running a standard VB or an overfitting VB. In further research, we will propose an even faster sequential updating algorithm, which can estimate the unknown multi-modal structure in seconds, without losing much accuracy of the mixture density estimation.

Chapter 5

Optimal Permutation Sequential Algorithm for Mixture Models

5.1 Introduction

In the previous chapters, we presented a two-stage Bayesian nonparametric approach on an industrial application. The Bayesian nonparametric approach provides more flexibility when modelling unknown multi-modal structures, which becomes increasingly popular in many field of statistics. One of the most common constructions is sampling under the Dirichlet process (DP), introduced by Ferguson [1973]. When considering the density estimation problem on continuous functions, Lo [1984] proposed a class of priors called the Dirichlet process mixture (DPM) that assigns a Dirichlet process on the mixing distribution. The continuity of the density can be approximated by a mixture model with a countably infinite number of components. Hence, one can express the posterior distribution in a DPM by integrating out G as a product of the posterior for the partition multiplied by independent posteriors for each component. Instead of using the marginalisation of G to indirectly achieve such a structure, one could directly specify a model for the random partition by assuming conditional independence given the allocation to clusters. More details on DPM and other related asymptotic behaviours can be found in Chapter 2 of the book edited by Hjort et al.

[2010].

The main computational hurdle when applying the DPM models to estimate the mixture density is the intractability of posterior distributions. Recent research based on the Bayesian methods has made the implementation of the DPM model feasible and efficient. As one of the most commonly applied and widely studied approaches, the Markov Chain Monte Carlo (MCMC) methods provide inference by sampling under the exact posterior distribution. A review of both conjugate and non-conjugate sampling methods for DPM models can be found in Neal [2000].

The proposed BL-DPMP approach in Chapter 3 have been analysed using the slice sampler. In order to explore the posterior distribution of component parameters θ, MCMC requires a sufficient number of samples to be generated. In practice, MCMC can be computationally expensive, especially when dealing with large datasets. Therefore, there is an intense demand for fast alternatives that can be executed within minutes or seconds.

Several fast alternatives to MCMC have been developed by researchers recently. Liu [1996] proposed a sequential imputation method, which treats the hidden component allocation parameters $\boldsymbol{\gamma}$ as missing data and approximates the actually incomplete data posterior of the parameter vector by a mixture of the complete data posterior. The multiple complete datasets used in the mixture are created by this variation of importance sampling without requiring iterations. The recursive update procedure introduced by Newton and Zhang [1999] and Newton [2002] is designed to estimate the unknown mixing distribution based on sequential arriving observations. Wang and Dunson [2011] proposed a fast algorithm for DP mixture models, called the sequential updating and greedy search (SUGS) algorithm, which sequentially maximises the conditional posterior probability relaying on previously processed observations. Zhang et al. [2014] developed a variational SUGS (VSUGS) algorithm, which combines the idea contained in VB with the SUGS algorithm. VSUGS deals with finite mixture models, assigning a probability distribution on allocation, rather than the hard allocation used in SUGS.

All sequential techniques, including Newton and Zhang [1999], rely on the chosen ordering of the data. Unlike MCMC sampling approaches, the permutation invariant property for joint density of observations and allocations $f(\mathbf{y}, \boldsymbol{\gamma})$ does not hold for sequential approaches. Therefore, the posterior inference of component parameters θ is affected by the updating order of data. Different permutations of data may lead to different mixture density estimates. In order to eliminate or reduce the order-dependent effect, citetLiu1996 suggested to create multiple independent complete data copies and averaged them out to obtain a Monte Carlo type of estimator. The density estimates obtained using SUGS are found to be dependent on the order in which observations are added. As the order of the data is typically arbitrary, Wang and Dunson [2011] recommended repeating the SUGS algorithm for multiple random permutations of the data, as a way to eliminate the order-dependence. Similarly, VSUGS generates multiple density estimates with different permutations of data and uses the evidence lower bound (ELBO) for model selection instead of the pseudo-marginal likelihood (PML) recommended by SUGS. Wang and Walker [2017] proposed an optimal ordering (OO) method, which is designed to find the optimal ordering of data first, then samples multiple sets of allocations based on the determined order of data. The OO method averages over multiple sets of predictions based on sampled allocation variables.

In this chapter, we propose a fast *optimal permutation sequential* (OPS) algorithm, which provides a strategy to find an optimal permutation of observation. At the same time, it determines the corresponding observations' allocations. The component parameters can then be updated directly, once the optimal order of the data and the allocation are found. The proposed OPS algorithm only requires a single run, which makes it extremely fast when dealing with small or median sized datasets. Furthermore, we develop a large data strategy of OPS (L-OPS) to keep the computational advantage of the proposed procedure when the sample size n increases.

The remainder of this chapter is organised as follows. Section 5.2 provides details of the proposed OPS algorithm, including the large data strategy of OPS (L-OPS). The performance of OPS is demonstrated through both simulation studies and a real application; this is presented in Section 5.3, along with the comparisons with other algorithms. The conclusion and discussion are presented in Section 5.4.

5.2 Methodology

5.2.1 A Sequential Approach for Mixture Models

In this section, we start with the well-known Dirichlet process mixture model, which takes the form

$$f(y) = \int f(y|\theta)\, \mathrm{d}G(\theta), \tag{5.2.1}$$

where $f(y|\theta)$ is the density function for each θ, and $G(\cdot)$ is a DP prior with the presentation

$$G(\cdot) = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j}(\cdot) \tag{5.2.2}$$

where $\delta_{\theta_j}(\cdot)$ denotes the point mass at θ and $\{\omega_j\}_{j=1}^{\infty}$ represents the mixing weights, which is defined as before in Equations (3.3.3).

Suppose observations $y_1, y_2, \ldots$ arrive sequentially, at time i, we have allocated previous $\mathbf{y}^{(i-1)} = \{y_i\}_{i=1}^{i-1}$ observations according to $\boldsymbol{\gamma}^{(i-1)} = \{\gamma_i\}_{i=1}^{i-1}$ and an approximation of the true posterior $p(\theta_{\gamma_{i-1}}|\mathbf{y}^{(i-1)})$. Then, the conditional predictive distribution of allocation variable γ_i given by the previous allocations $\boldsymbol{\gamma}^{(i-1)}$ follows a multinomial distribution with probabilities

$$p(\gamma_i = j|\boldsymbol{\gamma}^{(i-1)}) = \begin{cases} \frac{\sum_{l=1}^{i-1} \mathbb{1}_{\{\gamma_i = j\}}}{\alpha+i-1}, & j = 1, \ldots, k_{i-1} \\ \frac{\alpha}{\alpha+i-1}, & j = k_{i-1}+1, \end{cases} \tag{5.2.3}$$

where $k_{i-1} = \max(\{\gamma_i\}_{i=1}^{i-1})$ represents number of mixing components after updating previous $i-1$ observations. With probability $\alpha/(\alpha+i-1)$, the j^{th} observation is allocated to a new component $k_{i-1}+1$. The number of components also depends on the constant α. The larger value of α we chose, the stronger tendency for a new observation to be allocated into a new component rather than those existing components. As mentioned in Section 2.1.3, we have the expected value of distinct mixing components $\mathbb{E}(k|\alpha, n) \approx \alpha \log(1+n/\alpha)$ for n moderately large.

With the current observation y_i, the sequential updating equation of allocation γ_i given previously observed $\mathbf{y}^{(i-1)}$ and determined allocations $\boldsymbol{\gamma}^{(i-1)}$ can be derived as

$$p(\gamma_i|y_i, \mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}) = \frac{f(y_i, \gamma_i|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})}{f(y_i|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})}, \tag{5.2.4}$$

where the joint density of y_i and γ_i on the numerator in Equation (5.2.4) can be obtained by

$$f(y_i, \gamma_i|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}) = p(\gamma_i|\boldsymbol{\gamma}^{(i-1)})f(y_i|\gamma_i, \mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}). \tag{5.2.5}$$

The marginal density of y_i on the denominator in Equation (5.2.4) can be obtained by summarising over all possible values of γ_i,

$$f(y_i|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}) = \sum_{j=1}^{k_{i-1}+1} f(y_i, \gamma_i = j|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}). \tag{5.2.6}$$

Then, we can calculate the probability $p(\gamma_i|y_i, \mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})$ in Equation (5.2.4), and the value of allocation γ_i can be assigned to the component with the largest conditional posterior density $p(\gamma_i|y_i, \mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})$.

Once we determine the allocation of γ_i, the conditional posterior distribution of θ_{γ_i} given the new observation y_i and the corresponding allocation γ_i can be updated through

$$p(\theta_{\gamma_i}|\mathbf{y}^{(i)}, \boldsymbol{\gamma}^{(i)}) \propto f(y_i|\gamma_i, \theta_{\gamma_i})p(\theta_{\gamma_i}|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}), \tag{5.2.7}$$

where $p(\theta_{\gamma_i}|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})$ is the posterior of θ_{γ_i} updated by previous observations $\mathbf{y}^{(i-1)}$ and corresponding allocations $\boldsymbol{\gamma}^{(i-1)}$. The conditional posterior distribution of θ_{γ_i} is available in analytical form if we assign conjugate prior distributions. Therefore, the sequential procedure can be proceeded by iteratively calculating allocation probabilities $p(\gamma_i|\mathbf{y}^{(i)}, \boldsymbol{\gamma}^{(i-1)})$ Equation in (5.2.4) and updating conditional posterior distribution $p(\theta_{\gamma_i}|\mathbf{y}^{(i)}, \boldsymbol{\gamma}^{(i)})$ through Equation (5.2.7). In each step, the sequential approach uses $p(\theta_{\gamma_i}, \gamma_i|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})$ to approximate the exact posterior distribution $p(\theta_{\gamma_i}, \gamma_i|\{y_i\}_{i=1}^n, \{\gamma_l\}_{\{l\neq i\}})$.

Next, we illustrate the details of the recursive updating scheme. By assigning the first observation y_1 to the first component $\gamma_1 = 1$, we can derive the conditional posterior of θ_1 as

$$p(\theta_1|y_1, \gamma_1) \propto f(y_1|\gamma_1, \theta)\pi_0(\theta), \tag{5.2.8}$$

where $\pi_0(\theta)$ represents the prior distribution. Given θ_1 and γ_1, we observe the next observation y_2. According to the DP prediction rule, the probability of assigning observation y_2 to the existing component and a new component is given by,

$$p(\gamma_2 = j|\gamma_1) = \begin{cases} \frac{1}{\alpha+1}, & j = 1, \\ \frac{\alpha}{\alpha+1}, & j = 2. \end{cases}$$

We can calculate the joint density of y_2 and γ_2 given y_1 and γ_1 by

$$\begin{aligned} f(y_2, \gamma_2|y_1, \gamma_1) &= \int f(y_2, \gamma_2|\theta_1)p(\theta_1|y_1, \gamma_1)\mathrm{d}\theta_1 \\ &= \int f(y_2|\gamma_2, \theta_1)p(\gamma_2|\gamma_1)p(\theta_1|y_1, \gamma_1)\mathrm{d}\theta_1. \end{aligned}$$

Thus, the posterior of γ_2 given y_2, y_1 and γ_1 is stated as

$$p(\gamma_2|y_2, y_1, \gamma_1) = \frac{f(y_2, \gamma_2|y_1, \gamma_1)}{f(y_2|y_1, \gamma_1)}$$

where the denominator in the above equation $f(y_2|y_1, \gamma_1)$ is the marginal density of y_2 given both y_1 and γ_1. The marginal density $f(y_2|y_1, \gamma_1)$ is obtained by summarising over $\gamma_2 = 1$ and $\gamma_2 = 2$,

$$f(y_2|y_1, \gamma_1) = \sum_{j=1}^{2} f(y_2, \gamma_2 = j|y_1, \gamma_1).$$

Following this, we can determine the allocation of y_2 by calculating the allocation probability $p(\gamma_2|y_2, y_1, \gamma_1)$. If $\gamma_2 = 1$, it means y_2 will be allocated into the existing component with parameter θ_1, and the corresponding component parameter θ_1 can be updated by

$$p(\theta_1|y_2, \gamma_2, y_1, \gamma_1) \propto f(y_2, \gamma_2|\theta_1)p(\theta_1|y_1, \gamma_1).$$

If $\gamma_2 = 2$, it means that the observation y_2 is assigned to a new component. Similar to Equation (5.2.8), θ_2 can be updated by

$$p(\theta_2|y_2, \gamma_2) \propto f(y_2, \gamma_2|\theta)\pi_0(\theta).$$

In general, when we observe y_i, for $i \geq 2$, there are $k_{i-1} + 1$ choices for the allocation variable γ_i, where k_{i-1} represents the number of distinct values obtained from previous allocations $\{\gamma_l\}_{l=1}^{i-1}$. Hence, according to Equation (5.2.4), the probability of allocation $\gamma_i = j$ is proportional to

$$p(\gamma_i|y_i, \mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}) \propto f(y_i, \gamma_i|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}), \quad \text{for} \quad j = 1, \ldots, k_{i-1} + 1.$$

Based on the value of γ_i, the conditional posterior distribution of the affected parameter θ_{γ_i} can be updated via Equation (5.2.7).

Here, we assume the density $f(y|\theta)$ follows a Normal distribution $\mathcal{N}(y|\theta)$ with component parameters $\theta = (\mu, \Lambda)$ in Equation (5.2.1). The component parameter θ is sampled from some prior distribution $G(\cdot)$ on $\mathbb{R} \times \mathbb{R}^+$, and $G(\cdot)$ itself is modelled as a Dirichlet process $\text{DP}(\alpha, G_0)$ with a precision parameter α and a base distribution G_0. In this model, the base distribution G_0 is known as

$$\pi_0(\mu, \Lambda) = \mathcal{N}(\mu|m_0, \tau_0\Lambda^{-1})\mathcal{W}(\Lambda|s_0, S_0), \tag{5.2.9}$$

where $\mathcal{N}(\cdot|m_0, \tau_0\Lambda^{-1})$ denotes a Normal with mean m_0 and τ_0 controls the precision of μ, and $\mathcal{W}_D(\cdot|s_0, S_0)$ represents a D-dimensional Wishart distribution with degree of freedom s_0 and scale matrix S_0. In this model, the density of y_i can also be viewed as following an infinite mixture of Normal distributions.

Without losing any generality, we suppose that the newly arrived observation y_i is assigned to the j^{th} component for $j = 1, \ldots, k_{i-1}+1$, where the new $(k_{i-1}+1)^{\text{th}}$ component is generated from the prior distributions given in Equation (5.2.9). Then, conditioned on previously updated component parameters $\mu_j \sim \mathcal{N}(\mu_j|m_j^{(i-1)}, \tau_j^{(i-1)}\Lambda_j^{-1})$ and $\Lambda_j \sim \mathcal{W}(\Lambda_j|s_j^{(i-1)}, S_j^{(i-1)})$, the posterior distribution of component parameters (μ_j, Λ_j) can be derived as

$$\begin{aligned}
&p(\mu_j, \Lambda_j|y_i, \gamma_i = j, \mathbf{y}^{(i-1)}, \gamma^{(i-1)}) \\
&\qquad \propto \mathcal{N}(y_i|\mu_j, \Lambda_j^{-1})\ \mathcal{N}(\mu_j|m_j^{(i-1)}, \tau_j^{(i-1)}\Lambda_j^{-1})\ \mathcal{W}(\Lambda_j|s_j^{(i-1)}, S_j^{(i-1)}).
\end{aligned} \tag{5.2.10}$$

If we examine the terms only dependent on parameter μ_j, the conditional posterior of μ_j can be derived as

$$\begin{aligned}
&p(\mu_j|y_i, \Lambda_j, m_j^{(i-1)}, \tau_j^{(i-1)}) \\
&\propto \exp\left\{-\frac{\tau_j^{(i-1)}+1}{2\tau_j^{(i-1)}}\left[\left(\mu_j - \frac{\tau_j^{(i-1)}y_i + m_j^{(i-1)}}{\tau_j^{(i-1)}+1}\right)^T \Lambda_j \left(\mu_j - \frac{\tau_j^{(i-1)}y_i + m_j^{(i-1)}}{\tau_j^{(i-1)}+1}\right)\right]\right\},
\end{aligned}$$

which is recognised as the kernel of a Normal distribution $\mathcal{N}(\mu_j|m_j^{(i)}, \tau_j^{(i)}\Lambda_j^{-1})$ with updating equations of mean $m_j^{(i)}$ and precision scale parameter $\tau_j^{(i)}$ defined as below,

$$
\begin{aligned}
m_j^{(i)} =& \frac{\tau_j^{(i-1)} y_i + m_j^{(i-1)}}{\tau_j^{(i-1)} + 1} \\
\tau_j^{(i)} =& \frac{\tau_j^{(i-1)}}{\tau_j^{(i-1)} + 1}.
\end{aligned}
\tag{5.2.11}
$$

We can derive the posterior distribution of Λ_j by gathering terms related to Λ_j, which is proportional to

$$
p(\Lambda_j | y_i, s_j^{(i-1)}, S_j^{(i-1)}) \propto \left| \frac{1}{\tau_j^{(i-1)} + 1} \right|^{\frac{1}{2}} |\Lambda_j|^{\frac{s_j^{(i-1)}+1-D-1}{2}}
\\
\exp\left\{ -\frac{1}{2} \mathrm{tr} \left[\left(\frac{(y_i - m_j^{(i-1)})(y_i - m_j^{(i-1)})^T}{\tau_j^{(i-1)} + 1} + S_j^{(i-1)^{-1}} \right) \Lambda_j \right] \right\}.
$$

This expression is the kernel of a Wishart distribution $\mathcal{W}\left(\Lambda_j | s_j^{(i)}, S_j^{(i)}\right)$, where the degree of freedom $s_j^{(i)}$ and scale matrix $S_j^{(i)}$ can be updated as,

$$
\begin{aligned}
s_j^{(i)} =& s_j^{(i-1)} + 1 \\
S_j^{(i)^{-1}} =& \frac{(y_i - m_j^{(i-1)})(y_i - m_j^{(i-1)})^T}{\tau_j^{(i-1)} + 1} + S_j^{(i-1)^{-1}}.
\end{aligned}
\tag{5.2.12}
$$

When $i = 1$, hyperparameters $(m_j^{(0)}, \tau_j^{(0)}, s_j^{(0)}, S_j^{(0)})$ in Equations (5.2.11) and (5.2.12) are set to be their initial values (m_0, τ_0, s_0, S_0) given by the prior distributions in Equation (5.2.9).

5.2.2 Optimal Permutation for Ordering

In order to reduce the ordering sensitivity mentioned in Section 5.1, we propose a sequential maximisation procedure to find the optimal permutation of the data, at the same time, to determine the corresponding allocations of the data. Then, the joint density of $\mathbf{y}$ and $\boldsymbol{\gamma}$ can be approximated by the determined permutation. The component parameters $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ can be updated following the determined optimal permutation of observations $\mathbf{y}$ and their allocations $\boldsymbol{\gamma}$.

Suppose observations $\{y_i\}_{i=1}^n$ are observed and their allocations are denoted by $\{\gamma_i\}_{i=1}^n$. We can decompose the joint density $f(\mathbf{y}, \boldsymbol{\gamma})$ as,

$$
\begin{aligned}
f(\mathbf{y}, \boldsymbol{\gamma}) =& f(y_1, y_2, \ldots, y_n, \gamma_1, \gamma_2, \ldots, \gamma_n), \\
=& f(y_1, \gamma_1) f(y_2, \gamma_2 | y_1, \gamma_1) \cdots f(y_n, \gamma_n | \mathbf{y}^{(n-1)}, \boldsymbol{\gamma}^{(n-1)}),
\end{aligned}
\tag{5.2.13}
$$

where $\mathbf{y}^{(i-1)} = \{y_l\}_{l=1}^{i-1}$ and $\boldsymbol{\gamma}^{(i-1)} = \{\gamma_l\}_{l=1}^{i-1}$ for $i = 1, \ldots, n$. Each $f(y_i, \gamma_i | \mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})$ is a joint marginal density for the current observation y_i and its corresponding allocation γ_i, based on previously selected observations $\mathbf{y}^{(i-1)}$ and their allocations $\boldsymbol{\gamma}^{(i-1)}$. An optimal permutation $\mathbf{y}_{\boldsymbol{\eta}} = (y_{\eta_1}, \ldots, y_{\eta_n})$ is determined by the proposed OPS algorithm, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$, and $y_{\eta_h} = y_i$ indicate that the h^{th} position in the optimal permutation is occupied by the observation y_i from original order. The OPS algorithm finds the optimal order through maximising the joint density $f(\mathbf{y}, \boldsymbol{\gamma})$, which is achieved by sequentially maximising each term $f(y_{\eta_h}, \gamma_h | \mathbf{y}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)})$ in Equation (5.2.13).

To maximise each marginal joint density $f(y_{\eta_h}, \gamma_h | \mathbf{y}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)})$, we only need to find a combination of $(y_{\eta_h} = y_i, \gamma_h = j)$ which gives the largest value from $y_i \in \{y_l\}_{l=1}^n$, $y_i \notin \{y_{\eta_h}\}_{h=1}^{h-1}$ and $\gamma_h = j$ for $j = 1, \ldots, k_{h-1}+1$. We can calculate the marginal joint density $f(y_{\eta_h}, \gamma_h | \mathbf{y}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)})$ by integrating out the component parameters (μ_j, Λ_j),

$$
\begin{aligned}
& f(y_{\eta_h}, \gamma_h | \mathbf{y}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)}) \\
=& \int_{\Lambda_j} \int_{\mu_j} f(y_{\eta_h}, \gamma_h | \mu_j, \Lambda_j)\ p(\mu_j | \Lambda_j, m_j^{(h-1)}, \tau_j^{(h-1)})\ p(\Lambda_j | s_j^{(h-1)}, S_j^{(h-1)})\ \mathrm{d}\mu_j\ \mathrm{d}\Lambda_j.
\end{aligned}
\tag{5.2.14}
$$

As it is mentioned in Section 5.2.1, the posterior distribution of mean parameter μ_j follows a Normal distribution $\mathcal{N}(\mu_j | m_j^{(h-1)}, \tau_j^{(h-1)} \Lambda_j^{-1})$, and precision matrix Λ_j follows a Wishart distribution $\mathcal{W}(\Lambda_j | s_j^{(h-1)}, S_j^{(h-1)})$. By integrating over the component parameters μ_j and Λ_j, the marginal joint density can be written as,

$$
\begin{aligned}
&f(y_{\eta_h}, \gamma_h | \mathbf{y}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)}) \\
=&f(y_{\eta_h} | \gamma_h, \mathbf{y}^{(h-1)})\ p(\gamma_h | \boldsymbol{\gamma}^{(h-1)}) \\
\propto&(\pi(s_j^{(h-1)} - D + 1))^{-\frac{D}{2}} \Gamma_D\left(\frac{s_j^{(h-1)}+1}{2}\right) \Gamma_D^{-1}\left(\frac{s_j^{(h-1)}}{2}\right) \left|\frac{S_j^{(h-1)}(s_j^{(h-1)} - D + 1)}{\tau_j^{(h-1)}+1}\right|^{\frac{1}{2}} \\
&\left|\frac{(y_{\eta_h} - m_j^{(h-1)})^T S_j^{(h-1)} (y_{\eta_h} - m_j^{(h-1)})(s_j^{(h-1)} - D + 1)}{(\tau_j^{(h-1)}+1)(s_j^{(h-1)} - D + 1)} + 1\right|^{-\frac{1}{2}(s_j^{(h-1)}+1)} p(\gamma_h | \boldsymbol{\gamma}^{(h-1)})
\end{aligned}
$$

which is proportional to a noncentral Student's t-distribution,

$$
\begin{aligned}
&f(y_{\eta_h}, \gamma_h | \mathbf{y}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)}) \\
&\quad\sim t_{s_j^{(h-1)} - D + 1}\left(y_{\eta_h} | m_j^{(h-1)}, \frac{S_j^{(h-1)}(s_j^{(h-1)} - D + 1)}{\tau_j^{(h-1)}+1}\right) p(\gamma_h | \boldsymbol{\gamma}^{(h-1)}).
\end{aligned}
\tag{5.2.15}
$$

Given the explicit form of $f(y_{\eta_h}, \gamma_h | \mathbf{y}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)})$ in Equation (5.2.15), it is not difficult to compute the joint density under all possible combination of observations and allocations $(y_{\eta_h} = y_i, \gamma_h = j)$ for the rest $(n - h + 1)$ unpicked observations and $k_{h-1} + 1$ number of mixing components. Once we have determined the pair of $(y_{\eta_h} = y_i, \gamma_h = j)$, the posterior distributions of component parameters $(\mu_{\gamma_h}, \Lambda_{\gamma_h})$ can be updated through Equations (5.2.11) and (5.2.12). Hence, the OPS updating procedure can be proceeded as described until all observations are updated.

Next, we go through the details of the proposed OPS algorithm. To simplify the notation, we use $f(y_{\eta_h}, \gamma_h)$ to denote the joint conditional density of (y_{η_h}, γ_h) given previously updated observations $\mathbf{y}^{(h-1)}$ and allocations $\boldsymbol{\gamma}^{(h-1)}$. In order to find the first observation in the optimal permutation, we calculate the marginal density for every observation y_i for $i = 1, \ldots, n$,

$$
f(y_{\eta_h} = y_i, \gamma_1 = 1) = \int f(y_{\eta_h} = y_i, \gamma_1 = 1 | \mu, \Lambda) \mathrm{d}\pi_0(\mu, \Lambda), \tag{5.2.16}
$$

where $\pi_0(\cdot)$ denotes the prior distribution stated in Equation (5.2.9), as none of observations has been proceeded yet. The first observation y_{η_1} is chosen to be the one which maximises the prior marginal distribution in Equation (5.2.16). At the same time, the allocation of y_{η_1} is automatically assigned to $\gamma_1 = 1$. The component parameters (μ_1, Λ_1) for the first component can then be updated via Equations (5.2.11) and (5.2.12).

When we consider the second position of $h = 2$, we need to find a pair of (y_{η_2}, γ_2) from the rest $(n-1)$ unpicked observations and $\gamma_2 = j$ for $j = 1, 2$, to maximise

$$f(y_{\eta_2} = y_i, \gamma_2 = j) = \int f(y_{\eta_2} = y_i, \gamma_2 = j, \mu_j, \Lambda_j) \mathrm{d}p(\mu_j, \Lambda_j | y_{\eta_1}, \gamma_1).$$

For $j = 1$, $p(\mu_j, \Lambda_j | y_{\eta_1}, \gamma_1)$ denotes the conditional posterior distributions of parameters (μ_1, Λ_1) for the first component. When $j = 2$, it means y_{η_2} is assigned to a new component, the calculation is the same as stated in Equation (5.2.16). Then, we can form a $(n-1) \times 2$ matrix values of $f(y_{\eta_2}, \gamma_2)$, where each row represents an observation and each column is a possible allocation for that observation. The pair of (y_{η_2}, γ_2) is determined by the pair with the largest value of $f(y_{\eta_2}, \gamma_2)$ in the matrix. Thus, based on the value of γ_2, we can update the component parameters $(\mu_{\gamma_2}, \Lambda_{\gamma_2})$ given observation y_{η_2}.

In general, in order to determine the pair $(y_{\eta_h} = y_i, \gamma_h = j)$ for a optimal position h, we calculate the marginal density $f(y_{\eta_h}, \gamma_h)$ of all combinations from the rest $(n-h+1)$ unpicked observations and all possible existing mixing component $j = 1, \ldots, k_{h-1}+1$. Then, we can form a $(n-h+1) \times (k_{h-1}+1)$ matrix as,

$$\begin{array}{c} \\ \text{obs. } 1 \\ \text{obs. } 2 \\ \vdots \\ \text{obs. } n-h+1 \end{array} \begin{array}{c} \begin{array}{ccc} \text{comp. } 1 & \cdots & \text{comp. } k_{h-1}+1 \end{array} \\ \begin{pmatrix} f(y_{\eta_h} = y_1, \gamma_h = 1) & \cdots & f(y_{\eta_h} = y_1, \gamma_h = k_{h-1}+1) \\ f(y_{\eta_h} = y_2, \gamma_h = 1) & \cdots & f(y_{\eta_h} = y_2, \gamma_h = k_{h-1}+1) \\ \vdots & \ddots & \vdots \\ f(y_{\eta_h} = y_{n-h+1}, \gamma_h = 1) & \cdots & f(y_{\eta_h} = y_{n-h+1}, \gamma_h = k_{h-1}+1) \end{pmatrix} \end{array} . \tag{5.2.17}$$

If the largest value is located at the i^{th} row and the j^{th} column in the matrix, we assign the observation y_i to occupy the h^{th} position in the optimal permutation, and allocate this y_i to the j^{th} component. Thus, component parameters (μ_j, Λ_j) are updated with hyper-parameters $(m_j^{(i)}, \tau_j^{(i)}, s_j^{(i)}, S_j^{(i)},)$, while parameters for other components $\{(\mu_l, \Lambda_l)\}_{\{l \neq j\}}$ remain unaffected. We keep applying this recursive procedure until $h = n$.

Once the recursion procedure ends, we obtain a finite mixture of Normal distributions with k_n components, where component parameters are updated to $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{y}^{(n)}, \boldsymbol{\gamma}^{(n)})$. Then, the predictive distribution of y_{n+1} given preciously updated $\mathbf{y}^{(n)}$ can be obtained by summarising over all possible values of allocation variable γ_{n+1} as,

$$f(y_{n+1}|\mathbf{y}^{(n)}) = \sum_{j=1}^{k_n+1} f(y_{n+1}, \gamma_{n+1} = j|\mathbf{y}^{(n)}, \boldsymbol{\gamma}^{(n)}). \tag{5.2.18}$$

The pseudo-code of the OPS algorithm is presented as follows.

Algorithm 2 Optimal Permutation Sequential Algorithm

Initialise prior values (m_0, τ_0, s_0, S_0) and set $k_1 = 1$.
for $i = 1, \ldots, n$ **do**
    Compute the joint marginal density in Equation (5.2.15).
end for
Find the observation y_i with the largest value and set $\eta_1 = i$.
for $h = 2, \ldots, n$ **do**
    for $i = 1, \ldots, (n - h + 1)$ **do**
        for $j = 1, \ldots, (k_{h-1} + 1)$ **do**
            Compute the joint marginal density in Equation (5.2.15).
        end for
    end for
end for
Find the location of the largest value in the matrix (5.2.17).
Set η_h equals to the row number of the largest location.
Set γ_h equals to the column number of the largest location.
if $\gamma_h > k_{h-1}$ **then**
    Create a new component with parameters updated from initial values.
else
    Update component parameters via Equations (5.2.11) and (5.2.12).
end if
Draw prediction via Equation (5.2.18).

5.2.3 Large Data Strategy of OPS

The proposed OPS algorithm introduced in Section 5.2.2 provides a fast sequential way to estimate the DPM models. In order to simultaneously choose y_{η_h} and its allocation γ_h, OPS finds the largest value of the joint marginal density $f(y_{\eta_h}, \gamma_h)$ for every possible combination of (y_{η_h}, γ_h). Hence, for each optimal permutation position h, it requires computing $(n-h+1)\times(k_{h-1}+1)$ number of joint densities as described in the matrix (5.2.17). The proposed OPS algorithm has a desirable advantage on the computational aspect for small datasets. However, when data size n increases, it is noticeable that the required computational cost increases as $\mathcal{O}(n^2)$. In this section, we propose a large data strategy for OPS (L-OPS) algorithm to deal with large datasets.

When implementing OPS, we have found that after selecting a certain number of observations, the variation of permutations for the rest observations does not affect the resulting estimate as much. This implies that the beginning of the selection procedure is critical for the mixture density estimation. Therefore, for a large dataset, we only run the OPS algorithm described in Section 5.2.2 upto a sufficient proportion of data, as long as it contains enough information to mimic the structure of the whole data. As for the rest observations, we only need to decide on the allocations with the natural order of observations and perform the sequential updating procedure until the end.

Considering observations $\mathbf{y} = \{y_i\}_{i=1}^{n}$ arrive with their natural ordering, we split $\mathbf{y}$ into two parts, for example, $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ where $\mathbf{y}_1 = \{y_i\}_{i=1}^{n_1}$ and $\mathbf{y}_2 = \{y_i\}_{i=n_1+1}^{n}$. We apply the proposed OPS method on the first part of data $\mathbf{y}_1 = \{y_i\}_{i=1}^{n_1}$, and obtain a mixture density estimate with parameters updated to $(\boldsymbol{\mu}^{(n_1)}, \boldsymbol{\Lambda}^{(n_1)})$. For observations in $\mathbf{y}_2$, we keep the original order and find their allocations based on the estimate from $\mathbf{y}_1$. For example, the allocation probability of the first observation y_{n_1+1} from the second part of data $\mathbf{y}_2$ to the j^{th}

component is given by

$$p(\gamma_{n_1+1} = j|\mathbf{y}^{(n_1+1)}, \boldsymbol{\gamma}^{(n_1)}) \propto f(y_{n_1+1}, \gamma_{n_1+1} = j|\mathbf{y}^{(n_1)}, \boldsymbol{\gamma}^{(n_1)}),$$

which is the same as what we used in OPS algorithm in Equation (5.2.4). According to the value of γ_{n_1+1}, the component parameters $(\mu_{\gamma_{n_1+1}}, \Lambda_{\gamma_{n_1+1}})$ can be updated straightforwardly following Equations (5.2.11) and (5.2.12). The L-OPS algorithm proceeds the sequential procedure until all observations in $\mathbf{y}_2$ are updated and the predictive distribution of y_{n+1} given $\mathbf{y}$ can be derived analytically.

Next, we briefly discuss how to determine the size of the $\mathbf{y}_1 = \{y_i\}_{i=1}^{n_1}$ when applying the L-OPS algorithm. We need $\mathbf{y}_1$ to contain enough observations in order to mimic the structure of the data $\mathbf{y}$. On the other hand, the more observations that we include in $\mathbf{y}_1$, the more computational effort is required for running OPS on $\mathbf{y}_1$. Since OPS has an obvious advantage when data size is small, we suggest to pick $n_1 = \min\{10\% \times n, 1000\}$, depending on the length of data. For example, if a dataset contains 5,000 observations, where $10\% \times 5000 < 1000$, then we choose $n_1 = 500$ as the size of $\mathbf{y}_1$. If the dataset has 20,000 observations, where $10\% \times 20000 > 1000$, then we can consider $n_1 = 1000$ to be large enough.

Algorithm 3 Large Data Strategy of Optimal Permutation Sequential Algorithm

```
Split dataset y = (y_1, y_2), where y_1 = {y_i}_{i=1}^{n_1} and y_2 = {y_i}_{i=n_1+1}^{n}
for i = 1, ..., n_1 do
    Run the optimal permutation sequential algorithm on y_1.
end for
Obtain the mixture density estimate with k_{n_1} mixing components.
for i = n_1 + 1, ..., n do
    for j = 1, ..., (k_{n_1} + 1) do
        Determine the allocation variable via Equation (5.2.15).
        Update corresponding component parameters by Equations (5.2.11) and (5.2.12).
    end for
end for
Draw prediction via Equation (5.2.18).
```

5.3 Results Analysis

In this section, we conduct simulation studies to evaluate the performances of the proposed OPS algorithm and its extension L-OPS algorithm. Both univariate and multivariate cases are considered. For each case, simulated data examples and a real data example are studied. For those data with small to median sample size up to 1000, we apply the OPS algorithm to estimate the mixture densities, and record the corresponding execution times. Meanwhile, the accuracy of density estimations is measured by the Kullback-Leibler divergence (KLD). For large sized data examples, we apply the L-OPS algorithm and evaluate its performance on estimation accuracy and time.

We compare the performance of the OPS algorithm with the SUGS algorithm proposed by Wang and Dunson [2011] and the VSUGS algorithm introduced by Zhang et al. [2014]. Following the suggestion in Wang and Dunson [2011], we run SUGS multiple times on the same data with random permutations and select the result with the order which achieves the largest pseudo marginal likelihood (PML). The VSUGS algorithm proposed by Zhang et al. [2014] is a generalisation for the SUGS algorithm, which follows the Variational Bayes (VB) framework with a fixed number of components. VSUGS uses a "soft" allocation procedure rather than the "hard" allocation applied in SUGS. They claimed that VSUGS gives a better fit of data whose components are close in some region. Similar to SUGS, we run VSUGS multiple times on the same data with random permutations and choose the outcome with the order which achieves the largest evidence lower bound (ELBO). For the real industrial dataset, we present the comparison of the number of failures obtained from the OPS algorithm with the slice sampler applied in Chapter 3 and the proposed hybrid Variational Bayes (HVB) approach in Chapter 4.

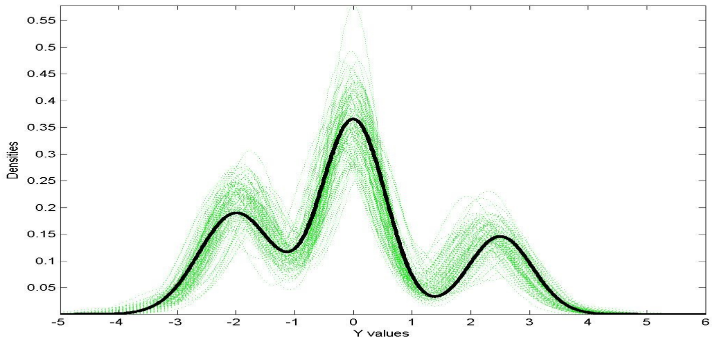
5.3.1 Simulated Univariate Data

We start the simulation study from estimating a mixture density of three univariate Normal distributions. The true density is given as

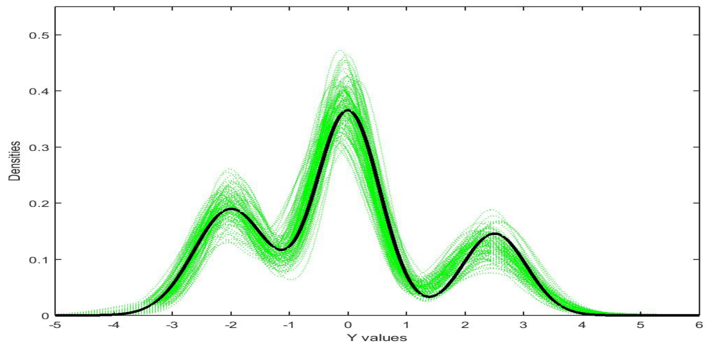
$$f(y) = 0.3\,\mathcal{N}(y|-2, 0.4) + 0.5\,\mathcal{N}(y|0, 0.3) + 0.25\,\mathcal{N}(y|2.5, 0.3), \tag{5.3.1}$$

which is the same mixture density used in Wang and Dunson [2011]. We randomly generate 100 datasets under the true density with different sample sizes $n = 100$, 200, 500 and 1000. The prior distributions of mean and precision parameters are fixed as $\mathcal{N}(\mu|0, 10\lambda^{-1})$ and $\mathcal{G}(\lambda|4, 1)$. Using the OPS algorithm outlined in Section 5.2, the predictive density is obtained by a single run. For each 100 datasets with the same length n, the 100 estimated predictive densities are plotted in Figure 5.1, which clearly shows the consistency of the OPS estimator for mixture density estimation. As the sample size increasing, the density estimate gets more and more concentrated to the true density, and the bandwidth of the 100 repetitions gets narrower.

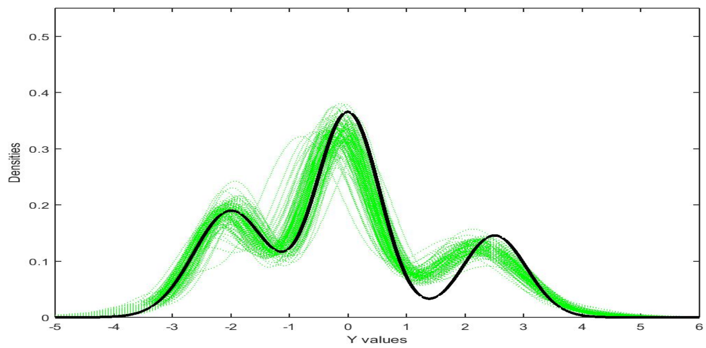
We calculate the KLD to measure how close the OPS estimates to the true density. To obtain a predictive estimate from the SUGS algorithm, we choose exactly the same prior initials as recommended in Wang and Dunson [2011], and run it for 30 times with random permutations of data. The best permutation is selected as the one with the largest pseudo marginal likelihood. Then the mixture density estimate is obtained according to the selected permutation. For the VSUGS algorithm, we fix the total number of components to be 10 upfront and repeat VSUGS for 30 times with random permutations. The best permutation is then chosen to be the one with the highest evidence lower bound. We summarise the averaged KLD values obtained from OPS for different sample sizes n in Table 5.1, along with the averaged KLD obtained from SUGS and VSUGS.



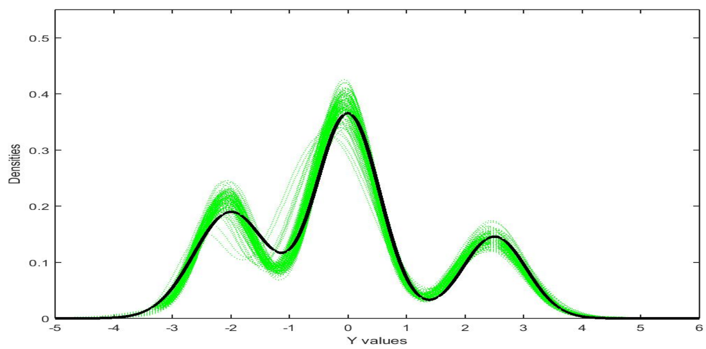
(a) n=100



(b) n=200



(c) n=500



(d) n=1000

Figure 5.1: OPS density estimates in univariate simulation for $n = 100$*, 200, 500 and 1000. The estimated densities (green, dotted) from 100 datasets and the true mixture density (black, solid).*

Table 5.1: Comparison of SUGS, VSUGS, and OPS in terms of the average of KLD from 100 datasets with different sample sizes $n = 100$*, 200, 500 and 1000.*

Sample size	SUGS	VSUGS	OPS
100	0.0443	0.0526	0.0398
200	0.0285	0.0314	0.0272
500	0.0412	0.0432	0.0439
1000	0.0063	0.0089	0.0139

According to Table 5.1, OPS provides consistent density estimates with various sample sizes. Comparing with to both SUGS and VSUGS, OPS can achieve more accurate density estimates when $n \leq 200$. For the data of size up to 500, OPS reaches a comparable level of KLD with both SUGS and VSUGS. When $n > 500$, OPS starts perform poorly.

Another important aspect we would like to evaluate is the computational time. The OPS algorithm is designed to remove order-dependent effect by finding the optimal permutation of data. On the other hand, SUGS and VSUGS repeat running for multiple time and pick the best density estimations by some criteria. For each sample size n, the execution times for those three methods are presented in Table 5.1.

Table 5.2: Comparison of SUGS, VSUGS, and OPS in terms of computational time (in seconds) from 100 datasets with different sample sizes $n = 100$*, 200, 500 and 1000.*

Sample size	SUGS	VSUGS	OPS
100	90	101	35
200	165	188	71
500	293	496	314
1000	833	818	1102

When $n \leq 200$, OPS clearly has an advantage on computation time, which is only a half and a third of the computational time required by SUGS and VSUGS respectively. However, when $n > 500$, the computational time of OPS increases dramatically. To run OPS algorithm, we need to compute the marginal density matrix (5.2.17) for each position h in the optimal permutation where $h = 1, \ldots, n$.

Therefore, as n increases, the required computational time for the marginal density matrix increases with $\mathcal{O}(n^2)$. In contrast, the computational time for SUGS and VSUGS are proportional to $\mathcal{O}(n)$.

In the end of this section, we examine the sensitivity of OPS algorithm about initial values of prior distributions. In order to generate densities estimation shown in Figure 5.1, we have chosen the initial value of m_0 as the sample mean,

$$m_0 = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

in the prior distribution $\mu \sim \mathcal{N}(\mu|m_0, \tau_0\Sigma)$. One concern of the OPS algorithm is that it always starts with the observation which achieves the maximum value of prior marginal, hence the OPS potentially can be sensitive of the initial values of prior distributions.

In order to test the sensitivity of initial values, we conduct two cases which the value of m_0 are chosen extremely. In both case, we randomly generate 100 observations which follow the same density of three Normal mixtures stated in Equation (5.3.1).

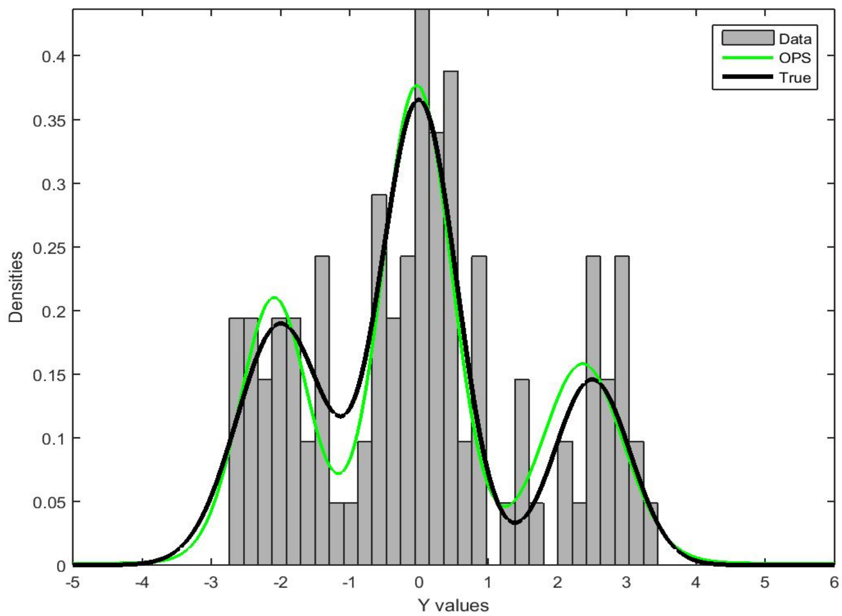


Figure 5.2: OPS density estimates for $n = 100$, where initial value of m_0 is chosen to be the maximum value of observations.

In the first case, the value of $m_0 = \max\{y_i\}_{i=1}^n$ is chosen to be the maximum value of all observations. The OPS algorithm successfully catchs three modes and the predictive distribution are shown in Figure 5.2.

In the second case, we generate another 100 observations and the value of $m_0 = \min\{y_i\}_{i=1}^n$ is chosen to be the minimum value of all observations. Similar density estimates are achieved by OPS algorithm, which is shown in Figure 5.3.

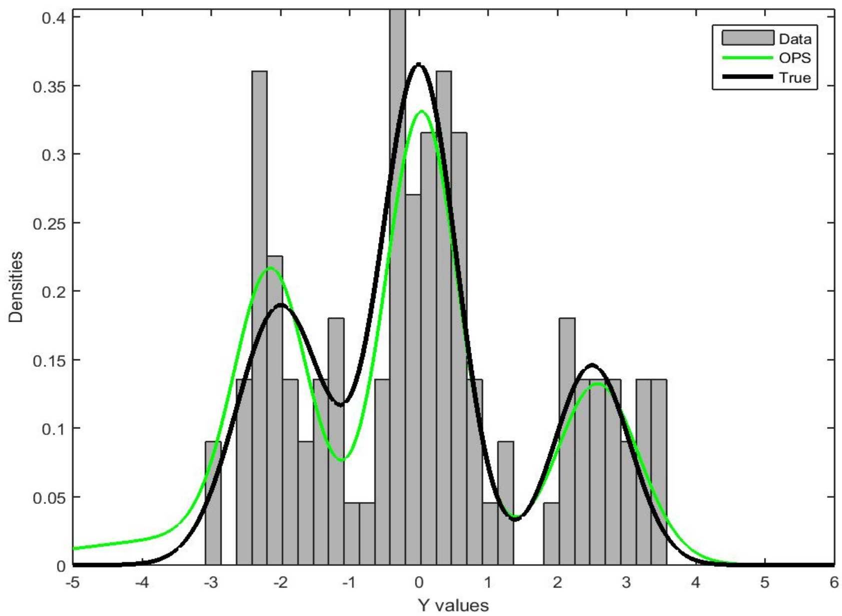


Figure 5.3: OPS density estimates for $n = 100$, where initial value of m_0 is chosen to be the minimum value of observations.

Furthermore, we random generate 100 observations from a bimodal density,

$$f(y) = 0.4\,\mathcal{N}(y|-2, 0.4) + 0.5\,\mathcal{N}(y|3, 0.1). \tag{5.3.2}$$

Then, the OPS algorithm is applied to estimate the mixture density where m_0 is chosen to be the sample mean. The estimated density is given in Figure 5.4. In this case, we conduct a situation that the sample mean initial value actually provides inappropriate starting point of the OPS algorithm. According to the density estimates, OPS results in a mixture of 2 Normal densities and successfully catches both separate modes on each side of the sample mean. According to the presented results of different choices of prior initials, we conclude that the OPS algorithm is not very sensitive about initial values.

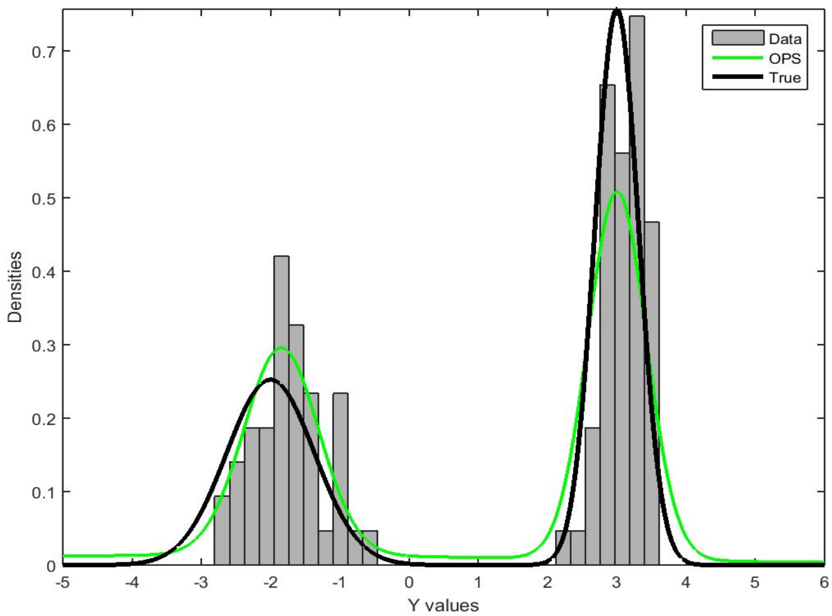


Figure 5.4: OPS density estimates for $n = 100$ with bimodal, where initial value of m_0 is chosen to be the sample mean.

Overall, the OPS algorithm provides reliable estimates of mixture densities with different sample sizes. OPS preforms especially well on small sized datasets according to accuracy of estimation and operational time. Meanwhile, the OPS algorithm does not show sensitivity about prior initials. It is noticeable that the computational time increases dramatically with the data size. When the data size $n > 500$, we recommend to use the L-OPS algorithm described in Section 5.3.3.

Large Data Strategy of OPS

Next, we present the performance of the large data strategy of OPS (L-OPS) algorithm on a large dataset with 5000 observations. As introduced in Section 5.2.3, we split the data into two parts $\mathbf{y}_1$ and $\mathbf{y}_2$, with the sizes denoted as n_1 and $(N - n_1)$. To study the affect of different splits, we will look at the splits with the first part data size $n_1 = 100$, 200, 500 and 1000, respectively. We record the KLD value and running time for each split in Table 5.3. We also record the

KLD value and running time of SUGS and VSUGS on the same dataset in Table 5.3. The dataset we simulated here is a (5000×1) vector.

Table 5.3: Comparison KLD and computational time obtained from SUGS, VSUGS and L-OPS with different sizes of $n_1 = 100$, 200, 500 and 1000.

	n_1 size	KLD	Time (sec)
	100	0.0210	2
	200	0.0256	7
L-OPS	500	0.0219	19
	1000	0.0113	23
	5000	0.0079	384
SUGS	5000	0.0039	136
VSUGS	5000	0.0020	43

The simulation results show that the performance of the L-OPS gets better with the KLD value decreases as the first part data size n_1 increases, while the computation time increases. For example, when $n_1 = 100$, the computational time of L-OPS is just 2 second CPU time.

Next, we visualise the curve fitting of density estimates obtained from L-OPS with different sizes of n_1.

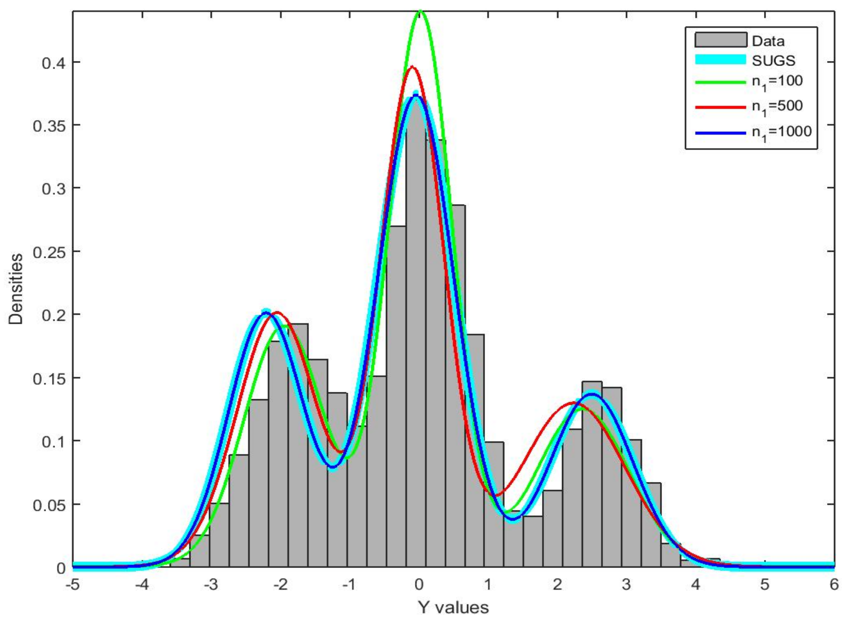


Figure 5.5: Large data strategy of OPS with different $n_1 = 100$, 500, 1000 obtained from OPS and SUGS.

When $n_1 = 100$, the resulting green curve is obviously not good fit. As n_1 increases to 500, the red curve fits the data better. Then, when $n_1 = 1000$, two estimated mixture curves obtained from L-OPS (blue) and SUGS (cyan) are overlapping to each other. According to Figure 5.5, L-OPS with $n_1 = 1000$ achieves a similar density estimate as the SUGS for a 5000×1 simulated dataset.

In summary, the large data strategy of OPS algorithm can be used as a fast alternative to estimate the mixture density with large sample size. The L-OPS algorithm execution time increases and KLD decreases as the size of n_1 increasing. It also can be viewed as a way to balance the accuracy of density estimation and time consumption by empirical selecting n_1.

5.3.2 Galaxy Data

The galaxy data is one of the commonly studied real data examples to assess various Bayesian methods. It is always used to assessing Bayesian methods for univariate density estimation and clustering. The data contains measured velocities of 82 galaxies for 6 well-separated conic sections of space. The prior distribution of mean μ follows a Normal distribution $\mathcal{N}(\mu|m_0, \tau_0\lambda)$ and the prior of precision λ follows a Gamma distribution $\mathcal{G}(\lambda|a_0, b_0)$, where initial values of hyperparameters are fixed as $\mu_0 = 0, \tau_0 = 10, a_0 = 20$, and $b_0 = 0.7$. Meanwhile, the default prior initials in SUGS are given by $\mu_0 = 0, \tau_0 = 10, a_0 = 1$, and $b_0 = 1$. then we repeatedly run it on 20 random permutations of the data as suggested in Wang and Dunson [2011]. The density estimates from OPS and SUGS are plotted in Figure 5.5.

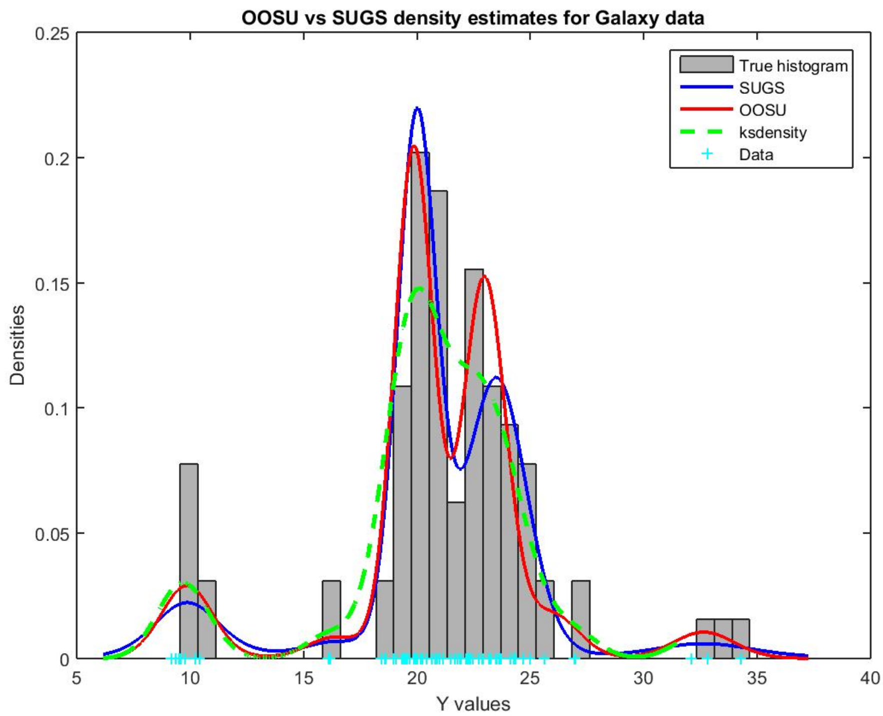


Figure 5.6: Predictive density estimate obtained from OPS (red, solid) and SUGS (blue, solid), kernel density estimate (green, dashed), histogram, and plots of galaxy data (+).

Here we also compare the the kernel density obtained using Matlab command called `ksdensity`. Clearly, the density estimated by `ksdensity` is over-smoothed, while OPS and SUGS provide better curve fittings. The estimated density from OPS in Figure 5.6 agrees closely.

Our method gave six distinctive clusters, which agrees with the fact that the true data come from 6 conic sections in space. We list the locations and weights of the six clusters obtained by OPS method in Table 5.4. We also list the locations and weights of five clusters obtained by SUGS to compare.

Table 5.4: Comparison of the estimated mixture densities obtained from OPS and SUGS on galaxy data.

	Comp.	1	2	3	4	5	6	Time
OPS	Locations	9.87	16.35	19.86	22.97	26.01	32.65	0.1877
	Weights	0.0854	0.0244	0.4390	0.3659	0.0488	0.0366	
SUGS	Locations	9.87	16.35	19.86	23.34	32.65	N/A	0.3915
	Weights	0.0854	0.0244	0.4390	0.4146	0.0366	N/A	

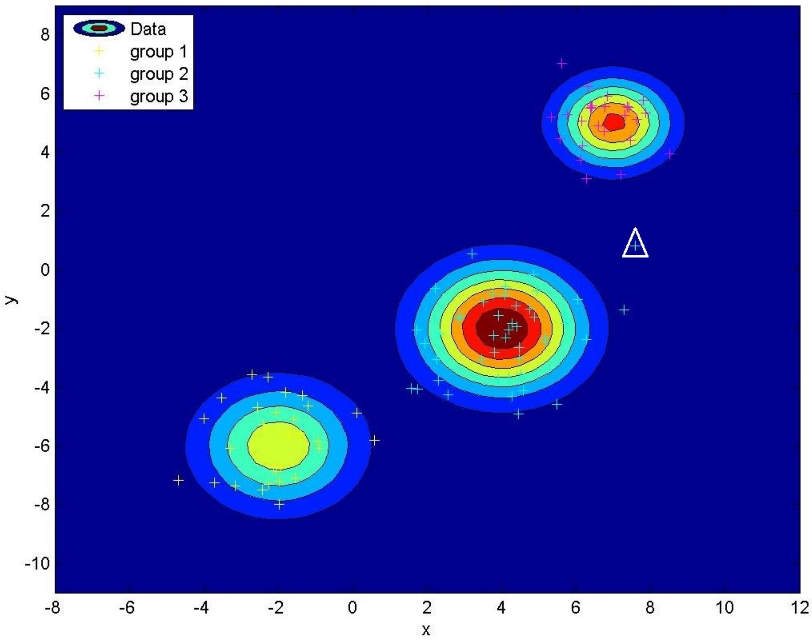
The component locations are sorted in an ascending order. According to Table 5.4, OPS and SUGS gave identical location estimates and corresponding weights of the first 3 components and the last one. The only difference is that OPS is able to distinguish component 4 and 5, and caught all 6 mixtures. Also, the runtime (in seconds) of our method is half of the SUGS time, which agrees with the findings in Table 5.2. Therefore, results presented above show the validation of the proposed OPS algorithm on univariate simulated datasets with various sizes and a real application on galaxy data. When dealing with datasets with small sample sizes, OPS has an obvious advantage on both accuracy and computational time.

5.3.3 Simulated Multivariate Data

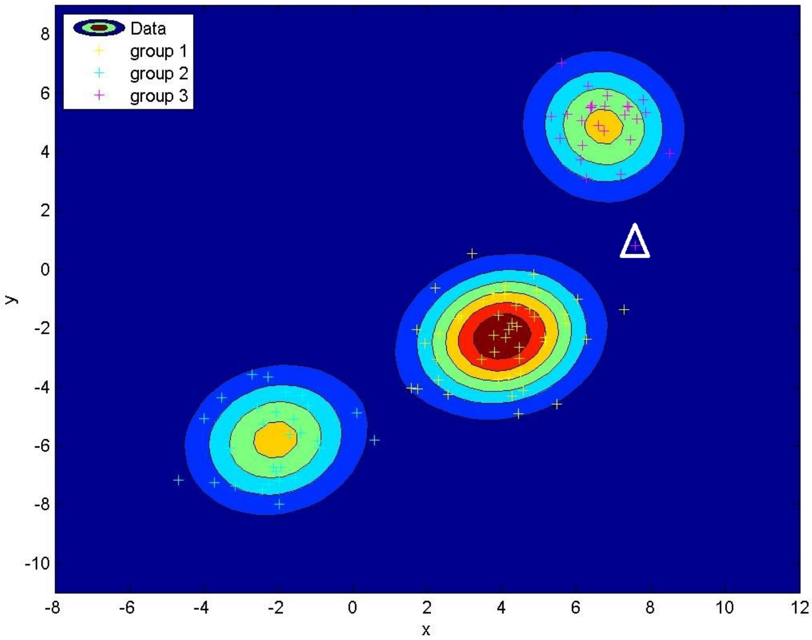
In this section, we first look at a bivariate simulation example. Mainly because it is not possible to visualise data in one plot when dimension of observations larger than 2. We simulate bivariate observations from the true density given as

$$f(\mathbf{y}) = 0.3\mathcal{N}\left(\begin{bmatrix}-2\\-6\end{bmatrix}, \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}\right) + 0.5\mathcal{N}\left(\begin{bmatrix}4\\-2\end{bmatrix}, \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}\right) + 0.2\mathcal{N}\left(\begin{bmatrix}7\\5\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right).$$

We sample 100 bivariate samples as data, and plot the true mixture density and the OPS estimate of the mixture density in Figure 5.4.



(a) True density



(b) OPS density estimate

Figure 5.7: Contour plot for a mixture of three bivariate Normal distributions with 100 observations, true density and the OPS density estimate.

Figure 5.7(a) represents the contour of true density with data denoted by '+' where different colours of '+' are used to distinguish observations generated from different Normal components. We also plot the OPS density estimate and the corresponding classification of data in Figure 5.7(b). If we compare these two plots closely, we can see that 99 data points are correctly allocated by OPS method except the marked by the white triangle, and OPS assigned it to the Normal component upper right. The OPS method accurately caught the three mixture modes in this example. Similar comparison results in Section 5.3.1 are achieved with other fast algorithms in multivariate cases.

5.3.4 Industrial Application

In Chapter 3, we proposed a Bayesian linear model with Dirichlet Process mixture prior (BL-DPMP) approach to reduce the number of measurements required for sensor characterisation. In order to obtain accurate estimates of the critical parameter β, the historical information with complete information is used as prior to compensate the missed information from the current sensor. The BL-DPMP approach takes the first $n = 240$ historical OLS estimates $\{\beta_i\}_{i=1}^{240}$ as observations, and then models $\{\beta_i\}_{i=1}^{n}$ by a DPM of Normal distributions described in Equation (3.3.1)

Instead of using the slice sampler to find the allocations and estimate the mixture density of β, we apply the OPS algorithm to the considered industrial problem. We record the allocations and weights obtained by OPS using the S1 dataset.

Table 5.5: Number of allocations and corresponding weight for each mixing Normal component obtained from OPS of the S1 historical dataset.

Comp.	No. of allo.	Weight	Comp.	No. of allo.	Weight
1	175	0.7292	2	48	0.2000
3	3	0.0125	4	3	0.00125
5	2	0.0083	6	2	0.0083
7	2	0.0083	8	1	0.0042
9	1	0.0042	10	1	0.0042
11	1	0.0042	12	1	0.0042

In Table 5.5, OPS gave 12 distinctive clusters of historical estimates $\{\beta_i\}_{i=1}^{240}$. We plot the histogram of $\{\beta_{i,2}\}_{i=1}^{240}$ in Figure 5.8, where each $\beta_{i,2}$ denotes the second dimension of β_i for $i = 1, \ldots, n$. The curve fitting obtained from OPS density estimate is also shown in Figure 5.8.

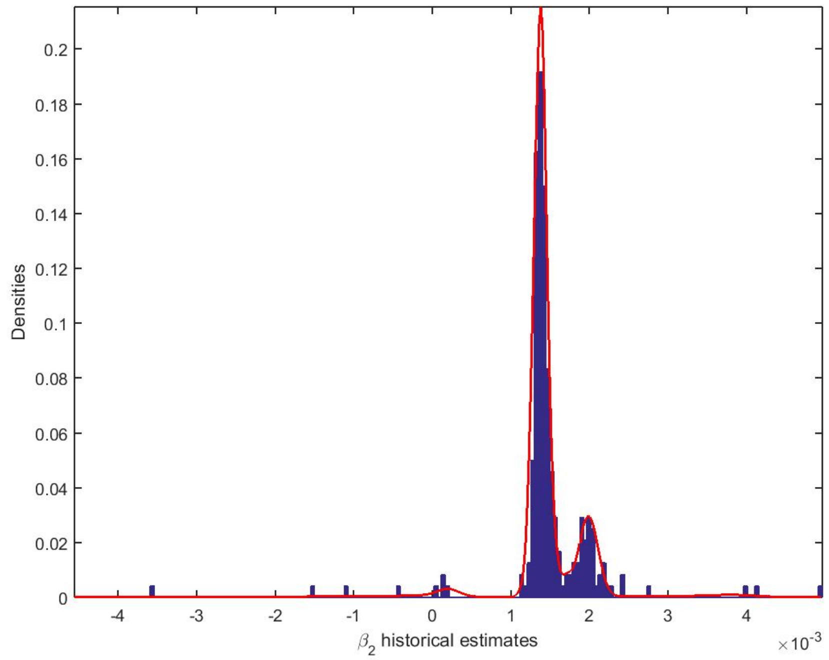


Figure 5.8: Histogram of the historical estimates $\{\beta_{i,2}\}_{i=1}^{240}$ and the curve fitting obtained from the BL-DPMP approach with OPS of the S1 type of sensors.

When a new sensor with reduced measurements (y_t^-, X_t^-) is available, the estimated mixture density of critical parameter β from OPS is used as prior choice for the regression coefficient β_t. The probability of selecting the most suitable Normal component is proportional to

$$\mathbb{P}(\gamma_t = j|\ y_t^-, X_t^-, \omega_j, \mu_j) \propto\ \omega_j\ \mathcal{N}(y_t^-|\ X_t^-\mu_j,\ \sigma^2\mathbb{I}_{18}).$$

where response variable y_t^- is a 18×1 vector, covariates X_t^- is a 18×5 matrix, γ_t denotes the allocation of the t^{th} sensor, ω_j and μ_j represent the weight and the mean for the j^{th} Normal component, the value of σ^2 is fixed to be extremely small, and $\mathbb{I}_{18}$ denotes a 18×18 identity matrix. Conditional on $\gamma_t = j$, the prior of β_t is assigned to be,

$$\pi(\beta_t) \sim \mathcal{N}(\beta_t|\ \mu_j, \Sigma_j).$$

Then, the posterior distribution of β_t can be estimated through a Bayesian linear model

$$p(\beta_t| \, y_t^-, X_t^-, \gamma_t = j, \mu_j, \Sigma_j) \propto \pi(\beta_t)\mathcal{N}\left(y_t^- | \, X_t^-\beta_t, \sigma^2\mathbb{I}_{18}\right).$$

We can calculate the characterisation error by $\varepsilon_t = y_t - X_t\beta_t$, where (y_t, X_t) represent the complete data with full measurements for the t^{st} sensor, and error term ε_t is a 32×1 vector.

According to the quality specifications stated in Equations (3.4.2), we calculate the absolute values of average errors for all temperature bands, and they must lie within a range of $[0, 0.3\%]$ to be a qualified product. In Table 5.6, we summarise the total numbers of failures from 655 sensors with reduced measurements obtained from the OPS, along with OLS approach with full measurements, previously described slice sampler in Chapter 3, hybrid Variational Bayes (HVB) in Chapter 4 with reduced measurements.

Table 5.6: Comparison performances obtained from OLS, Bayesian Linear approach, BL-DPMP approach with slice sampler, HVB and OPS on the incomplete dataset of the S1 type of sensors.

		No. of Failures (655)			Pass rate	Time (sec)
		0.1%	0.2%	0.3%		
OLS	complete	4	3	2	99.69%	0.04
MCMC	incomplete	162	24	8	98.78%	8486
HVB	incomplete	128	23	11	98.32%	225
OPS	incomplete	84	23	12	98.17%	4

The first row of Table 5.6 is the result achieved by the OLS method with complete S1 data, which can be viewed as a reference. Under the BL-DPMP approach, the number of pass products obtained by the slice sampler method (98.78%) is comparable to the OLS approach as the reference pass rate (99.69%). However, MCMC sampling method is computational expensive, which takes more than 2 hours to finish. The aim of the project is to reduce the characterisation time by taking a fewer number of measurements. Hence, we need a faster alternative to estimate the mixture density. For the HVB method proposed in Chapter 4, it achieves 98.32% pass rate and only takes about 3 minutes to finish. Moreover, the proposed OPS algorithm proposed in this chapter only takes 4 seconds. Meanwhile, the OPS algorithm gives 12 failures out of 655 sensors with reduced measurements, which equals to 98.17% pass rate. This pass rate is consider to be acceptable from the practical point of view.

To further illustrate the comparison among the proposed MCMC, HVB and OPS approaches in the BL-DPMP model, we present the estimated characterisation errors of one sensor as an example. The sensor is characterised by three approaches without having measurements at mid-range temperatures. The passing range is ±0.3% for average error of each temperature band.

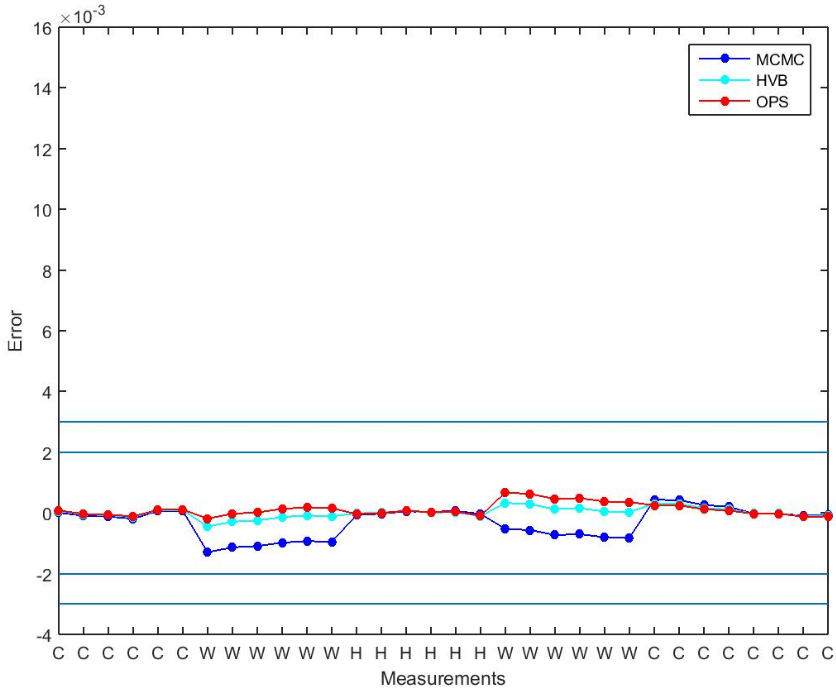


Figure 5.9: Errors comparison for one product among OLS, Bayesian linear, BL-DPMP approach with slice sampler, HVB and OPS of the S1 type of sensors.

Figure 5.9 shows that all three methods MCMC (blue), HVB (cyan) and OPS (red) successfully meet the company's quality specification. Moreover, errors obtained from all three methods lie within the ±0.1% error interval. Therefore, the OPS algorithm can deliver an accurate estimate of β with available historical information and reduced measurements. In conclude, the OPS algorithm provides an on-line solution for the considered industrial problem, which has the potential to be implemented on the characterisation line in the future.

5.4 Discussion

In this chapter, we proposed an optimal permutation sequential (OPS) algorithm to simultaneously find an optimal order of the data through a maximisation strategy and its corresponding allocations, and then we can then draw samples from weights and locations conditional on the allocations. The proposed OPS algorithm shows desirable advantages, in both terms of the high level of accuracy and the short computation time when the data size is small to median (up to 500 observations). For large datasets, the large data strategy of OPS (L-OPS) approach can be considered as an effective alternative to the original OPS algorithm. We examined the performance of OPS on both univariate case and multivariate case, and compared its performance with some other fast algorithms. We have also shown that the proposed OPS algorithm is not very sensitive of prior initials. A detailed discussion and suggestions are given in the analysis of results.

As shown in Section 5.3, the proposed OPS algorithm was slow when dealing with large datasets. The extended L-OPS algorithm is designed to further accelerate the processing speed. However, the accuracy of the mixture density estimation was not satisfactory. We attempt to further investigate the order-dependence of large data by testing different ways of selecting the size of n_1. For example, we can randomly select 10% of the data to form the first part of the data $\mathbf{y}_1$. In our proposed OPS approach, we used the joint marginal density to identify an optimal permutation of the data. We are also interested in other ways of finding a permutation of the data which that can lead to an accurate mixture density estimate. Future work is needed to examine other types of criteria in order to identify a suitable permutation of data without corresponding allocations.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we proposed a Bayesian linear model with Dirichlet process mixture prior (BL-DPMP), together with an optimal permutation sequential (OPS) algorithm as a solution to a real industrial enquiry. The problem under study is considered to be challenging, since the proposed solution intends to achieve three objectives at the same time. The first objective is to reduce the required number of measurements during the characterisation process. The second objective is to characterise each sensor via parameter estimation to achieve the high level of accuracy. The final objective is that the proposed statistical procedure should be able to perform an online characterisation.

In current practice, the ordinary least squares (OLS) method is applied to estimate the sensor parameter β with a certain number of measurements per device being necessary in order to achieve the required level of accuracy. The proposed Bayesian model reduces the number of measurements needed per device by incorporating a prior model constructed from the information of the previous related sensor estimates. We can then estimate the current sensor parameter β through the Bayesian theorem, by combining the prior distribution constructed from the historical dataset and the likelihood with measurements at the mid-range temperatures removed. Our intention is to use the historical information from previously characterised sensors to compensate the block missing information from the current sensor characterisation.

The natural of the data exhibited in Section 3.2 could suggest other ways of modelling historical estimates of parameter β. For instance, one choice could be a simple mixture of two t-distributions which can capture the outliers of historical β estimates shown in Figure 3.2. Here, we briefly state the reasons of not using such a mixture of two t-distributions. In general, a Student's t-distribution has heavier tails than a Normal distribution. Hence, the observations far away from the central area in Figure 3.2 can be considered as outliers and covered by t-distributions, instead of being individually modelled by Normal distributions. A mixture of two t-distributions might end with a more parsimonious model fitting. However, for the considered case here, our idea is to use the historical estimates of β provide information as prior to compensate the missed information from the likelihood of the current sensor. The estimated mixing components will be used as prior choices, and a mixture density with over-fitted number of components actually provide more options of the prior distribution for the coming sensor with reduced measurements. It is worth noting that those sensors with the outliers β estimates shown in Figure 3.2 are also passed the calibration. The locations of those outliers might be the locations for future modes. Therefore, we rather individually model the outliers by Normal distributions.

To capture the multi-modality shown in parameter estimates of a certain type of sensors, we proposed a Normal DPM model to capture the multi-modality of the regression coefficient β. The estimated mixture components are treated as the prior information of β and applied to a Bayesian linear regression model for the current sensor with reduced measurements. The proposed BL-DPMP approach was applied to the considered real industry dataset and outperformed both the OLS method and the Bayesian linear regression approach under the same condition. We also proposed a DPM regression model with data-driven prediction procedure to improve the prediction accuracy of a standard DPM regression model. As a result, the proposed DPMR-DDP approach significantly improved the prediction performance, although it still failed to achieve the satisfactory level of pass rate.

For both approaches proposed in Chapter 3, we applied the slice sampling scheme to estimate the mixture densities. This sampling scheme draws samples under exact posterior distribution, which requires expensive computational efforts. In order to reduce the computational time of the characterisation process, we proposed a hybrid Variational Bayes (HVB) algorithm which follows a greedy searching strategy. By exploring every possible model complexity, the HVB algorithm is highly likely to provide the global optimum solution. Moreover, at each step, we attempted to split every mixing component in two ways to provide a thorough description to the unknown multi-modality. The resulting performance of both the simulated data and the real industrial application suggests the reliability and capability of the proposed HVB algorithm on mixture density estimation problems. More importantly, HVB dramatically reduced the required computational effort.

The third objective of our project is to propose a statistical model which can be implemented in the characterisation line, where the model can be executed within seconds (online). In order to further accelerate the processing speed, a sequential update procedure is proposed to capture the structure of the historical information, where the observations are treated as arriving sequentially. The idea is to draw approximate independent and identically distributed samples from the posterior distribution of the allocation variables. Then, independent draws of component weights and locations can be taken from the posterior distribution conditional on the determined allocations. In this way, independent samples can be drawn from the predictive distribution. The proposed OPS algorithm approximates the true mixture density using a single run, which only runs for a few seconds. When dealing with small datasets, the proposed OPS has advantages on both the accuracy of the mixture density estimation and the computational time required. However, the proposed OPS algorithm faces a disadvantage when dealing with the large data size problem. To address this, we propose a large data strategy of OPS (L-OPS) algorithm. The simulation study shows the validity of the L-OPS algorithm when dealing with large datasets. By using the proposed BL-DPMP model together with the OPS algorithm, we provided an online solution to the real industrial case.

6.2 Future Work

It is worth noting that the industrial problem considered in this thesis comes from real life. The company made it apparent that a procedure of this nature will be implemented in the near future in the production line, which means the values of the sensor parameter will be estimated using the Bayesian method. However, there are a couple of concerns that need to be considered before implementing the Bayesian idea.

The first concern can be viewed as an information updating problem. When a new sensor with reduced measurements is characterised and passed, the parameter estimate of β can be used to refine the prior information for the next sensor and so on. For example, we use the historical estimates $\{\beta_i\}_{i=1}^n$ to construct a mixture model with k^* distinct Normal components with parameters $\{(\mu_j, \Sigma_j)\}_{j=1}^{k^*}$, where the mean μ_j follows $\mathcal{N}(\mu_j|m_j^{(n)}, \tau_j^{(n)}\Sigma_j)$ and the variance-covariance matrix Σ_j follows $\mathcal{W}(\Sigma_j^{-1}|s_j^{(n)}, S_j^{(n)})$. For the $(n+1)^{\text{th}}$ sensor characterisation, the estimated mixture model provides adequate prior information to the regression coefficient β_{n+1}. Following this, once the j^{th} Normal component is chosen, the prior distribution of β_{n+1} becomes $\pi(\beta_{n+1}) \sim \mathcal{N}(\beta_{n+1}|\mu_j, \Sigma_j)$. The posterior estimate $\hat{\beta}_{n+1}$ is calculated through a Bayesian linear regression model. The hyperparameters $(m_j^{(n+1)}, \tau_j^{(n+1)})$ of the mean μ_j used as the prior information for the next new sensor can be updated in the following way:

$$\begin{aligned} m_j^{(n+1)} =& \frac{\tau_j^{(n)} \hat{\beta}_{n+1} + m_j^{(n)}}{\tau_j^{(n)} + 1}, \\ \tau_j^{(n+1)} =& \frac{\tau_j^{(n)}}{\tau_j^{(n)} + 1}. \end{aligned}$$

Similarly, the hyperparameters $(s_j^{(n+1)}, S_j^{(n+1)})$ of the variance-covariance matrix Σ_j can be updated by

$$\begin{aligned} s_j^{(n+1)} =& s_j^{(n)} + 1, \\ S_j^{(n+1)} =& \frac{(\hat{\beta}_{n+1} - m_j^{(n)})(\hat{\beta}_{n+1} - m_j^{(n)})^T}{\tau_j^{(n)} + 1} + S_j^{(n)^{-1}}. \end{aligned}$$

In this way, the information of the $(n+1)^{\text{th}}$ sensor is updated to the mixture model, and the updated mixture model will be used as the prior information for the next sensor characterisation.

The second concern is to investigate the details of implementing the proposed procedure to different types of sensors. To implement the proposed procedure means that the characterisation line needs to be re-set. However, the company manufactures a large number of types of sensors with different amount of historical information available. For some types of sensors, we may find that there is not enough available historical information to construct our priors. Hence, this lead to the motivation to identify which types of sensors can be characterised by the proposed procedure with reduced measurements. For each identified type of sensors, we need to determine an appropriate size of the historical sample, in order to construct the prior information.

From the work developed throughout the Section 3.3.2, we found that the DPM regression model is unlikely to guarantee the high accuracy when estimating of the mixture density; this is important as high precision is the dominant concern. In order to estimate the relationship between response variables $\{y_i\}_{i=1}^n$ and covariates $\{X_i\}_{i=1}^n$ from the historical data, the DPM regression framework introduced by Müller et al. [1996] is thought for granted to be the model should be used in the considered situation. However, as shown in Table 3.4, the standard DPM regression model was the only method which broke down under the complete data condition. For the considered situation, the DPM regression model cannot achieve the satisfactory level of accuracy. We will further investigate this problem and propose a feasible solution by adding a refined procedure to improve the prediction precision of the mixture density estimation.

Bibliography

C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, 2:1152–1174, 1974.

H. Attias. Inferring parameters and sturcutre of latent variable models by variational bayes. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

D. J. Bayliss and S. P. Downes. An evaluation of a paar dma 5000 density meter. *Technical Report, National Physical Laboratory*, 2003.

M. Beal and Z. Ghahramani. The variational bayesian em algorithm for incomplete data: With application to scoring graphical model structures. *Bayesian Statistics 7 (Eds: J. M. Bernardo, M. Bayarri, J. Berger, A. David, D. Heckerman, A. F. M. Smith and M. West)*, 7:453–464, 2003.

C. M. Biship. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

D. Blackwell and J. B. MacQueen. Ferguson distributions via plya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

C. A. Bush and S. N. MacEachern. A semiparametric bayesian model for randomised block designs. *Biometrika*, 83:275–285, 1996.

C. Constantinopoulos and A. Likas. Unsupervised learning of gaussian mixtures based on variational component splitting. *IEEE Transactions on Neural Networks*, 8(3):745–755, 2007.

A. Corduneanu and C. M. Bishop. Variational bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA, 2001.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

D. B. Dunson and J. H. Park. Kernel stick-breaking processes. *Biometrika*, 95: 307–323, 2008.

M. D. Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89:268–277, 1994.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:1301–1312, 1995.

C. Eves, X. Wang, and S. G. Walker. Bayesian information for sensors. *Quality and Reliability Engineering International*, 31(8):1717–1724, 2015.

T. Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 1990.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6, 1984.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, United States of America, 1996.

J. E. Griffin and M. Steel. Order-based dependent dirichlet processes. *Journal of the American Statistical Association*, 10:179–194, 2006.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57, 1970.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian Nonparametrics*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK, 2010.

H. Ishwaran and L. James. Gibbs sampling methods for stick-breaking prior. *Journal of American Statistical Association*, 96:161–173, 2001.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and computing*, 21:93–105, 2011.

J. S. Liu. Nonparametric hierarchical bayes via sequential imputations. *The Annals of Statistics*, 24:911–930, 1996.

A. Y. Lo. On a class of bayesian nonparametric estimates i. density estimates. *Annals of Statistics*, 12:351–357, 1984.

S. MacEachern. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics - Simulation and Computation*, 23:727–741, 1994.

C. A. McGrory and D. M. Titterington. Variational approximations in bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, 51(11):5352–5367, 2007.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21, 1953.

P. Müller, A. Erkanli, and M. West. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83, 1996.

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

M. A. Newton. On a nonparametric recursive estimator of the mixing distribution. *Sankhya, Ser. A*, 64:306–322, 2002.

M. A. Newton and Y. Zhang. A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, 86:15–26, 1999.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica.*, 4: 639–650, 1994.

B. Shahbaba and R. Neal. Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2006.

D. M. Titterington. Bayesian methods for neural networks and related models. *Statistical Science*, pages 128–139, 2004.

N. Ueda and Z. Ghahramani. Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, 15(10):1223–1241, 2002.

N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. Smem algorithm for mixture models. *Neural computation*, 12(9):2109–2128, 2000.

S. Wade, D. B. Dunson, S. Petrone, and L. Trippa. Improving prediction from dirichlet process mixtures via enrichment. *Journal of Machine Learning Research*, 15:1041–1071, 2014.

S. G. Walker. Sampling the dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36:45–54, 2007.

B. Wang and D. M. Titterington. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

B. Wang and M. Titterington. Inadequacy of interval estimates corresponding to variational bayesian approximations. *Journal of the American Statistical Association*, 82:373–380, 2005.

L. Wang and D. B. Dunson. Fast bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20:196–216, 2011.

X. Wang and S. G. Walker. An optimal data ordering scheme for dirichlet process mixture models. *Computational Statistics and Data Analysis*, 112:42–52, 2017.

B. Wu, C. A. McGrory, and A. N. Pettitt. A new variational bayesian algorithm with application to human mobility pattern modeling. *Statistics and Computing*, 22(1):185–203, 2012.

X. Zhang, D. J. Nott, C. Yau, and A. Jasra. A sequential algorithm for fast fitting of dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 23:1143–1162, 2014.