



# Bayesian approaches to variable selection in mixture models with application to disease clustering

Zihang Lu<sup>a</sup> and Wendy Lou<sup>b</sup>

<sup>a</sup>Department of Public Health Sciences, Queen's University, Kingston, Ontario, Canada; <sup>b</sup>Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

## ABSTRACT

In biomedical research, cluster analysis is often performed to identify patient subgroups based on patients' characteristics or traits. In the model-based clustering for identifying patient subgroups, mixture models have played a fundamental role in modeling. While there is an increasing interest in using mixture modeling for identifying patient subgroups, little work has been done in selecting the predictors that are associated with the class assignment. In this study, we develop and compare two approaches to perform variable selection in the context of a mixture model to identify important predictors that are associated with the class assignment. These two approaches are the one-step approach and the stepwise approach. The former refers to an approach in which clustering and variable selection are performed simultaneously in one overall model, whereas the latter refers to an approach in which clustering and variable selection are performed in two sequential steps. We considered both shrinkage prior and spike-and-slab prior to select the importance of variables. Markov chain Monte Carlo algorithms are developed to estimate the posterior distribution of the model parameters. Practical applications and simulation studies are carried out to evaluate the clustering and variable selection performance of the proposed models.

## ARTICLE HISTORY

Received 20 October 2020  
Accepted 9 October 2021

## KEYWORDS


Bayesian growth mixture model; clustering; latent class; non-linear growth trajectories; variable selection

## 1. Introduction

In biomedical research, cluster analysis based on patients' characteristics or traits is often performed to identify disease subtypes. Disaggregating disease heterogeneity can help better understand the underlying biological mechanisms, which is a key building block for better disease management strategies, novel treatments and precision medicine.

Of many existing statistical methods, the finite mixture model (FMM) is a popular tool for modeling population heterogeneity. This model refers to modeling with categorical latent variables that represent subgroups of the population that is unobserved and need to infer from the data [40]. When clustering is the main interest, FMM is a powerful tool and is usually referred to as model-based clustering. Application of model-based clustering can be found in many different areas of research, for example, in the analysis of gene

**CONTACT** Zihang Lu  [zihang.lu@queensu.ca](mailto:zihang.lu@queensu.ca)

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2021.1994529>

expression data [60], brain magnetic resonance image data [61], adolescent alcohol use [11] and identifying asthma phenotypes [52]. A comprehensive review of this topic can be found in the discussion [21] and the recent review of model-based clustering [41]. An R package *FlexMix* has also been developed to perform finite mixture regression analysis [23,32].

While it is important to correctly identify clusters, for many cluster analyses, identifying clinical factors that are associated with the class assignment is also of great interest. These factors are crucial since they can help clinicians to make informed treatment decisions. However, it is often hardly known which variables are associated with class membership. Including all variables into the model without a selection will result in a large and complex model, which is also often difficult to interpret. Therefore, it is desirable to identify the appropriate subset of the variables in a mixture model to obtain a parsimonious model that simultaneously achieves consistent variable selection and optimal classification. Within the mixture model framework, several methods have been proposed to perform variable selection. Following the categorization in [18], the existing variable selection methods in the mixture model can be roughly divided into penalization approaches, in which variable selection is performed by using a penalized log-likelihood approach [28,59], model selection approaches, in which variable selection is considered as a model selection problem [12,17,19] and Bayesian approaches, in which variable selection is conducted by making inference about the posterior distribution via sampling strategies such as Markov Chain Monte Carlo (MCMC) [51,61]. A common feature of these approaches is that the variable selection procedure is applied in component-specific distribution (see  $f_k(\cdot)$  defined in Equation (2) in Section 3) to identify important variables or measurements, while assuming the probabilities of mixture components (or mixture weights) (see  $\pi_{ik}(\cdot)$  defined in Equation (3) in Section 3) do not depend on any covariates. Nevertheless, in clinical practice researchers may also be interested in which covariates predict the class assignment. Therefore, allowing variable selection in the context of the mixture model to determine the variables affecting the probabilities of the mixture components would significantly increase the flexibility of the model.

In the current study, we develop two approaches to perform variable selection in the context of the mixture model to identify important predictors that are associated with the class assignment. These two approaches are referred to as the one-step approach and the stepwise approach. The former refers to an approach in which clustering and variable selection are performed simultaneously in one overall model, whereas the latter refers to an approach in which clustering and variable selection are performed in two sequential steps. Shrinkage and spike-and-slab priors are adapted to these proposed approaches for selecting important variables. MCMC algorithms are developed to estimate the posterior distributions of parameters of interest. Real data and simulated data are used to compare the performance of these two approaches under different scenarios. The remainder of this paper is organized as follow: in Section 2, we describe two motivating studies and datasets to be considered in the current study. In Section 3, we first review the FMM and then describe the proposed model for both cross-sectional and longitudinal data. Variable selection methods based on shrinkage and spike-and-slab priors are also introduced in this section. In Section 4, we discuss the Bayesian inference for the proposed models. In Section 5, we apply the proposed models to analyze two motivating datasets for discovering disease phenotypes and

their risk factors, and in Section 6, we perform a simulation study to compare the clustering and variable selection performance of these models under different scenarios. Finally, in Section 7, we discuss and conclude our findings.

## 2. Motivating examples

Our methodological development is motivated by two clinical studies, namely the Primary Biliary Cirrhosis (PBC) study and the Childhood Asthma Management Program (CAMP).

### 2.1. Primary biliary cirrhosis study

The first example is a well-known study, the Mayo Clinic trial of the liver Primary Biliary Cirrhosis (PBC). PBC is a fatal chronic liver disease with an unknown cause. A randomized placebo-controlled trial was conducted between 1974 and 1984 to study the effect of the drug D-penicillamine on the treatment of PBC. In this 10-year interval, the trial recruited 424 PBC patients who met the eligibility criteria [14]. This dataset can be found in [16] and is also available in R *survival* package. The trial collected several variables, for example, patients' age survival status (0=alive, 1=liver transplant, 2=dead), drug (1=D-penicillamine, 2=placebo), presence of ascites (0=no, 1=yes), hepatomegaly (0=no, 1=yes), spiders (0=no, 1=yes) and edema (0=no, 1=yes), as well as other indices such as serum bilirubin and cholesterol level. A common problem from the clinical practice is how to make use of these markers of disease progression to identify subgroups (i.e. clusters) of PBC patients with similar characteristics, and what are the predictors of these group assignments. These subgroups may provide an important indication of patients' prognostic.

### 2.2. Childhood asthma management program

It is widely accepted that asthma is not a single disease, but a distinct disease caused by different underlying biological mechanisms. Discovering different phenotypes of asthma and factors associated with these phenotypes can help better understand the underlying biological mechanisms, which is a key building block for better asthma management strategies, novel treatments and precision medicine.

The Childhood Asthma Management Program (CAMP) is a triple-blinded randomized clinical trial originally designed to evaluate whether treatment with either an inhaled corticosteroid (budesonide) or an inhaled noncorticosteroid drug (nedocromil) safely produces an improvement in lung growth when compared with treatment for symptoms only. The primary outcome of CAMP is lung function measured by forced expiratory volume in 1 s (FEV1). Participants recruited to the trial were also followed by three phases of observational follow-up lasting 13 years. Both trial and observational follow-up included an annual prebronchodilator and postbronchodilator spirometry test (during which FEV1 data were collected) as part of the protocol. At baseline, 1041 children were randomly assigned to receive 200  $\mu\text{g}$  of budesonide (311 children), 8 mg of nedocromil (312 children), or placebo (418 children) twice daily. These participants were from 5 to 12 years of age with mild-to-moderate asthma and were treated for four to six years. The trial found that the anti-inflammatory medications did not have a better long-term effect than placebo

on lung function growth [7]. The trial also assessed differences between treatment groups regarding airway responsiveness, morbidity and physical growth etc. More details about the design of this study can be found in [6].

Since lung function has been associated with asthma in childhood [35] and adulthood [44], we are interested in understanding the growth and decline in lung function in these patients with childhood asthma. These distinct lung function patterns may reveal links between asthma and subsequent chronic airflow obstruction and aid in developing optimal personalized treatment strategies. Moreover, identifying determinants of the growth and decline in lung function would facilitate early disease diagnosis and prevent it from progressing to a more severe stage.

### 3. Methodology

In this section, we begin by first reviewing the finite mixture model, then describing the one-step and stepwise approaches for cluster analysis. We also describe the Gaussian mixture model for both cross-section and longitudinal data, as well as variable selection priors.

#### 3.1. Review of the finite mixture model

The model presented here is intended to be in sufficient generality to allow for more complex data structures. Let  $\mathbf{y}_i$  denote the data for  $N$  subject ( $i = 1, \dots, N$ ), where  $\mathbf{y}_i$  can be a single data point or multi-dimensional vector. An FMM with  $K$  components (i.e. classes) can be written as  $f(\mathbf{y}_i) = \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i)$ , for  $k = 1, \dots, K$ , where  $f_k(\mathbf{y}_i)$  are densities of cluster  $k$  and  $\pi_k$  are the probabilities of mixture components (or mixture weights) that sum to one, that is,  $0 \leq \pi_k \leq 1$  and  $\sum_{k=1}^K \pi_k = 1$ . It is not uncommon that  $f_k$  is assumed to have a parametric form, i.e.  $f_k(\mathbf{y}_i) = f_k(\mathbf{y}_i; \Phi_k)$ , where  $\Phi_k$  is a set of parameters that characterize the density  $f_k(\cdot)$ . Therefore, the FMM can be written as

$$f(\mathbf{y}_i; \Phi) = \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i; \Phi_k), \quad (1)$$

where  $\Phi = (\pi', \Phi_1', \dots, \Phi_K')$ . Depending on the types of data, one can specify a different probability model for  $f_k$ . For example, to model continuous cross-sectional data  $f_k$  may be a Gaussian density defined by mean and variance, whereas to model longitudinal data  $f_k$  may be a multivariate Gaussian density defined by mean vector and variance-covariance matrix.

#### 3.2. One-step approach

A more general formulation of the FMM is to allow  $\pi_k$  depend on some covariates  $\mathbf{x}_i$  and the class-specific density function depend on some covariates  $\mathbf{u}_i$ , that is,

$$f(\mathbf{y}_i; \mathbf{x}_i, \mathbf{u}_i, \Phi) = \sum_{k=1}^K \pi_{ik}(\mathbf{x}_i, \beta_k) f_k(\mathbf{y}_i; \mathbf{u}_i, \theta_k), \quad (2)$$

where  $\Phi = (\pi', \Phi'_1, \dots, \Phi'_K)'$ , and  $\Phi_k = (\beta'_k, \theta'_k)'$  denotes all parameters to be estimated for the  $k^{th}$  component.  $\theta_k$  is a set of parameters characterized the distribution  $f_k$ .  $\pi_{ik}(\mathbf{x}_i, \beta_k)$  is the probability of subject  $i$  belonging to class  $k$ , depending on the  $p$  dimensional covariates  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  and its associated parameters  $\beta_k = (\beta_{k1}, \dots, \beta_{kp})'$ . Specifically,  $\pi_{ik}(\mathbf{x}_i, \beta_k)$  can be modeled using a multinomial logistic regression

$$\pi_{ik}(\mathbf{x}_i, \beta_k) = \frac{\exp\{\beta_{k0} + \mathbf{x}'_i \beta_k\}}{\sum_{l=1}^K \exp\{\beta_{l0} + \mathbf{x}'_i \beta_l\}}. \tag{3}$$

We set  $\beta_{10} = 0$  (i.e. the intercept of the first class) and  $\beta_1 = \mathbf{0}$  (i.e. the  $p$ -dimensional coefficients of the first class) for identifiability purpose. The coefficients  $\beta_k$  (for  $k > 1$ ) can be interpreted in terms of change in log-odds relative to the first category. We refer to this model as the one-step approach in which predictors are included in the FMM.

### 3.3. Stepwise approach

Alternatively, one can also determine the variables that are associated with the class membership using an unconditional approach, also known as the stepwise approach. In the stepwise approach, the classification and prediction are separated steps such that class predictors do not affect the classification results. This avoids the scenarios in which including different covariates could lead to a different classification.

Specifically, the stepwise approach to incorporate predictors is consist of two steps. In the first step, we identify the latent classes (i.e. most likely class membership) via an unconditional model without including any predictors, i.e.  $f(y_i) = \sum_{k=1}^K \pi_k f_k(y_i)$ , where  $\pi_k$  does not depend on covariates and is identical to all subjects in class  $k$ . And in the second step, the class membership variable derived from the model is used as the outcome variable and regressed on the latent class predictors, which can be achieved by using a multinomial logistic regression as in (3).

### 3.4. Gaussian mixture model for cross-sectional data

To specify  $f_k$ , here we consider two common models based on Gaussian distribution. The first model can be used to model continuous cross-sectional data, whereas the second model can be used to model continuous longitudinal data.

For cross-sectional data, let  $y_i = y_i$  denote a single data point of subject  $i$ . Let  $u_i = u_i$  denote the time when the measurement is collected. The Gaussian mixture regression model can be written as

$$f_k(y_i; u_i, \theta_k) = N(\mu_k, \sigma_k^2), \tag{4}$$

where  $N(\cdot, \cdot)$  is a normal distribution and to model the component mean with the time covariate  $u_i$ , one can further specify a regression model  $\mu_k = U'_i \boldsymbol{\gamma}_k$ , where  $U_i$  is a  $q \times 1$  design matrix of  $u_i$ . In such case,  $\theta_k = (\boldsymbol{\gamma}_k, \sigma_k^2)$ . For example, if the component mean  $\mu_k$  is assumed to have a quadratic relationship with  $u_i$ , one can specify  $U_i = (1, u_i, u_i^2)'$  and  $\boldsymbol{\gamma}_k = (\gamma_{k0}, \gamma_{k1}, \gamma_{k2})'$ , which leads to  $\mu_k = \gamma_{k0} + \gamma_{k1} u_i + \gamma_{k2} u_i^2$ . To facilitate the Gibbs sampling of the model parameters, we used the following standard independent conjugate priors,  $\boldsymbol{\gamma}_k \sim \text{MVN}(\mathbf{0}, \mathbf{V}_{k0})$  and  $\sigma_k^2 \sim \text{IG}(a_{k0}, b_{k0})$ , where  $\mathbf{V}_{k0}$  is the variance–covariance

matrix of a  $q$  dimensional multivariate normal distribution. IG denotes the inverse gamma distribution with parameters  $a_{k0}$  and  $b_{k0}$ .

### 3.5. Gaussian mixture model for longitudinal data

For longitudinal data, let  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$  denote the data for subject  $i$ , where  $n_i$  denote the number of measurements for subject  $i$ . Let  $\mathbf{u}_i = (u_{i1}, \dots, u_{in_i})'$  denote the time when these measurements are collected. The growth mixture model for class  $k$  can be written as

$$f_k(\mathbf{y}_i, \boldsymbol{\vartheta}_{ik}; \mathbf{u}_i, \boldsymbol{\theta}_k) = \text{MVN}(\mathbf{U}'_i \boldsymbol{\gamma}_k + \mathbf{Z}'_i \boldsymbol{\vartheta}_{ik}, \sigma_k^2 \mathbf{I}_{n_i \times n_i}), \tag{5}$$

where  $\boldsymbol{\theta}_k = (\boldsymbol{\gamma}_k, \Sigma_k, \sigma_k^2)$  and  $\boldsymbol{\vartheta}_{ik} \sim \text{MVN}(\mathbf{0}, \Sigma_k)$ .  $\mathbf{U}_i$  denotes the  $q_1 \times n_i$  design matrix of the fixed effect and  $\mathbf{Z}_i$ , which is a subset of the columns of  $\mathbf{U}_i$ , denotes the  $q_2 \times n_i$  ( $q_2 < q_1$ ) design matrix of random effect.  $\boldsymbol{\vartheta}_{ik}$  is the corresponding random effect coefficient. Under this model, the subject  $i$  that belongs to class  $k$  has a mean trajectory  $\mathbf{U}'_i \boldsymbol{\gamma}_k$  and the variance–covariance matrix  $\mathbf{Z}_i \Sigma_k \mathbf{Z}'_i + \sigma_k^2 \mathbf{I}_{n_i \times n_i}$ . We used the following standard independent conjugate priors,  $\boldsymbol{\gamma}_k \sim \text{MVN}(\mathbf{0}, \mathbf{V}_{k0})$ ,  $\sigma_k^2 \sim \text{IG}(a_{k0}, b_{k0})$ ,  $\Sigma_k^{-1} \sim \text{Wishart}(r_{k0}, (r_{k0} \mathbf{R}_{k0})^{-1})$ , where the prior for the Wishart distribution is parametrized such that the mean is  $\mathbf{R}_{k0}^{-1}$ . In the special case where  $\Sigma_k$  is a diagonal matrix, i.e.  $\Sigma_k = \phi_k^2 \mathbf{I}_{q_2}$ , an inverse gamma prior is used, i.e.  $\phi_k^2 \sim \text{IG}(a_{k0}^*, b_{k0}^*)$ .

### 3.6. Variable selection priors

For the mixture model specified in (2) and (3), the number of coefficients  $\boldsymbol{\beta}$  needed to be estimated grows as the number of class  $K$  and the covariate dimension  $p$  increase. From a practical perspective, it is crucial to identify only the variables of importance in order to obtain an interpretable and parsimonious model ; therefore, it is necessary to place restrictions on the estimation of  $\boldsymbol{\beta}$  in order to obtain a robust final model.

Bayesian variable selection methods have received increasing attention and a variety of MCMC methods have been proposed for identifying important variables. These methods fall within the concept of Bayesian modeling average (BMA) in which parameter estimates uncertainty and model uncertainty are simultaneously achieved [24,37,58]. A popular class of the method for variable selection is imposing a shrinkage prior on the regression coefficients  $\boldsymbol{\beta}$  to cause a ‘shrinkage’ of the parameter estimation to lie around the origin. Examples of shrinkage priors include Bayesian lasso (also known as Laplace prior) [46,56], Horseshoe prior [9,10], Dirichlet–Laplace prior [4] and the modified Bayesian lasso method [38]. However, shrinkage prior itself usually would not lead to variable selection, and hard shrinkage (HS) rules (e.g. coefficients  $> 1$  standard deviation (SD)) may be applied to achieve this purpose. A review of these methods can be found in [37]. In our current implementation, we considered a popular shrinkage prior, the Horseshoe prior, which possesses strong theoretical guarantees for estimation, prediction and variable selection [9,10].

#### 3.6.1. Horseshoe prior

Since its proposal, the Horseshoe prior has become one of the most popular methods for shrinkage due to its outstanding performance and computational advantages. The

Horseshoe prior is a global-local shrinkage procedure in which the local shrinkage for the coefficient is determined by a hyper-parameter  $\epsilon_k$  and the overall level of shrinkage is determined by a hyper-parameter  $\xi_{kj}$ . While the Horseshoe prior does not have a closed-form density function, it can be written as a scale mixture of normals,

$$\beta_{kj} | \epsilon_k, \xi_{kj} \sim N(0, \epsilon_k^2 \xi_{kj}^2), \quad \epsilon_k \sim C^+(0, 1), \quad \xi_{kj} \sim C^+(0, 1), \tag{6}$$

where  $C^+$  denotes half-Cauchy distribution. Unlike Bayesian lasso which imposes shrinkage effect uniformly across all coefficients, the Horseshoe uses half-Cauchy distributions over the global parameter ( $\xi_{kj}$ ) and local hyper-parameter ( $\epsilon_k$ ) which results in strong shrinkage over weak coefficients whereas almost no shrinkage over the large coefficients. Such property has been proven to be useful to discriminate between true effects and noise. To efficiently sample from the posterior distribution, Makalic and Schmidt [39] proposed a simple sampler for regression models the Horseshoe prior.

Another category of variable selection method is the spike-and-slab prior (also known as discrete mixtures) which places a mixture prior of a point mass at  $\beta = 0$  and a continuous at  $\beta \neq 0$ . The spike-and-slab prior directly estimates the variable inclusion probabilities and thereby provides a direct measurement of variable importance. Examples of spike-and-slab priors include stochastic search variable selection (SSVS) [22], Gibbs variable selection [8,13] and Kuo and Mallick (KM) prior based on unconditional distribution [31]. A review of these methods can be found in [45]. In our current study, we considered a commonly used prior, i.e. SSVS prior.

### 3.6.2. Stochastic Search Variable Selection

In SSVS, selecting a subset of important predictors is equivalent to setting the associated  $\beta_{kj}$  of those non-selected variables to zero. With this idea, the SSVS places a normal mixture prior on  $\beta_{kj}$

$$\beta_{kj} | \delta_{kj} \sim (1 - \delta_{kj})N(0, \tau_{kj}^2) + \delta_{kj}N(0, c_{kj}^2 \tau_{kj}^2), \quad \delta_{kj} \sim \text{Bernoulli}(p_{kj}). \tag{7}$$

Therefore,  $\beta_{kj} \sim N(0, \tau_{kj}^2)$ , if  $\delta_{kj} = 0$ , and  $\beta_{kj} \sim N(0, c_{kj}^2 \tau_{kj}^2)$ , if  $\delta_{kj} = 1$ . The idea here is to set  $\tau_{kj}$  very small, such that those  $\beta_{kj}$  for which  $\delta_{kj} = 0$  will tend to be clustered around 0 (leading to the spike), and to set  $c_{kj}$  very large, such that for those  $\beta_{kj}$  for which  $\delta_{kj} = 1$  will tend to be dispersed (leading to the slab). To facilitate the posterior computation, the SSVS can also be represented as a multivariate normal prior,  $\boldsymbol{\beta}_k | \boldsymbol{\delta}_k \sim \text{MVN}(\mathbf{0}, \mathbf{D}_k \boldsymbol{\Gamma}_k \mathbf{D}_k)$ , where  $\boldsymbol{\Gamma}_k$  is the prior correlation matrix,  $\mathbf{D}_k = \text{diag}((a_{k1} \tau_{k1}), \dots, (a_{kp} \tau_{kp}))$ , with  $a_{kj} = 1$  if  $\delta_{kj} = 0$  and  $a_{kj} = c_{kj}$  if  $\delta_{kj} = 1$ .

It is recognized that SSVS is closely connected to the Horseshoe prior. To see this connection, one can re-parametrize the SSVS in (7) as  $\beta_{kj} | \delta_{kj} \sim (1 - \delta_{kj})N(0, s_{kj}^2) + \delta_{kj}N(0, \tilde{s}_{kj}^2)$ , where  $s_{kj}^2$  and  $\tilde{s}_{kj}^2$  denote the hyper-variances for the spike and slab distributions. Setting  $\tilde{s}_{kj} = 0$  induces a degenerate distribution at the origin and under such case the SSVS can be written analogous to (6) as  $\beta_{kj} | \delta_{kj} \sim N(0, \delta_{kj}^2 \tilde{s}_{kj}^2)$ . Here  $\delta_{kj}$  and  $\tilde{s}_{kj}$  have taken the role of  $\epsilon_k$  and  $\xi_{kj}$  in (6), but instead of giving continuous distribution for  $\delta_{kj}$  it only allows to take two values (i.e.  $\delta_{kj} = 0, 1$ ).

## 4. Bayesian inference

Based on the prior distributions described previously, in this section, we sketch the ideas of posterior computation via Gibbs sampling, followed by the clustering procedure and approaches to determine the number of clusters.

### 4.1. Posterior computation

To facilitate the posterior computation via Gibbs sampling, we considered a data-augmentation approach which is based on a new class of Polya-Gamma (PG) distribution (a subset of the class of infinite convolutions of gamma distributions), which allows fully Bayesian inference in models with binomial or multinomial likelihoods through an efficient Gibbs sampler.

A challenge of using an MCMC algorithm (including the Gibbs sampler) is the non-identifiability of the classes (or components). This issue arises because the mixture model is invariant under permutation of the indices of the classes, i.e. the parameters  $\Phi_1$  in class 1 cannot be distinguished from parameters  $\Phi_2$  in class 2 because they are exchangeable in the sense that the likelihood function will be invariant. As a result, the marginal posterior distributions of the parameters will be identical for each mixture component. This phenomenon is also known as the ‘label switching’ problem. In the current study, we applied a popular post-processing algorithm to reorder the labels based on Kullback–Leibler divergence [55]. The convergence of MCMC can be checked using Geweke statistics as well as visual inspection of the trace plots. The algorithm for posterior updates of parameters is provided in the Supplementary Materials.

The optimal classification for subject  $i$  ( $i = 1, \dots, N$ ) can be determined based on the marginal posterior component probabilities that subject  $i$  is assigned to class  $k$ , which is defined as

$$P(z_{ik} = 1 | \mathbf{y}_i) = \int P(z_{ik} = 1 | \mathbf{y}_i, \Phi) P(\Phi | \mathbf{Y}) d\Phi$$

$$\approx \sum_{s=1}^S \frac{\pi_{ik}^{(s)} P(z_{ik}^{(s)} = 1 | \mathbf{y}_i, \Phi_k^{(s)})}{\sum_{l=1}^K \pi_{il}^{(s)} P(z_{il}^{(s)} = 1 | \mathbf{y}_i, \Phi_l^{(s)})},$$

where  $\mathbf{Y}$  denotes all data and  $\Phi$  denotes all the parameters.  $S$  is the total number of MCMC iterations.  $z_{ik}^{(s)}$ ,  $\Phi_k^{(s)}$  and  $\pi_{i,k}^{(s)}$  denote the values at the  $s$ th iteration. Subject  $i$  can then be assigned to the class with the largest  $P(z_{ik} = 1 | \mathbf{y}_i)$ , for  $k = 1, \dots, K$ .

### 4.2. Determine the number of clusters

Although there is no widely accepted method for determining the number of clusters  $K$  in the mixture model, there have been many methods proposed. The common practice assumes the number of clusters  $K$  is fixed and finding the best  $K$  can be viewed as a model selection process. In such cases, model selection criteria can be implemented to compare models with different values of  $K$ . Statistical information criteria are commonly used, such as Akaike’s Information Criterion (AIC) [1] and Bayesian Information Criterion (BIC) [50]. Alternatively, the Lo–Mendell–Rubin (LMR) [33] method or bootstrap likelihood



ratio test (BLRT) [40] have also been used to determine  $K$  in finite mixture models. Nylund *et al.* [43] compared these indices for selecting the number of clusters in latent class models, factor mixture models and growth mixture models, and found that BIC outperformed the AIC, and the bootstrap likelihood ratio test provided the most consistent indicator of clusters.

Many model selection approaches were also proposed within the Bayesian framework. For example, the Bayes factor (BF) [27] is a commonly used index that is based on integrated likelihood, and can be applied when there are more than two candidate models and can be used for comparing non-nested models. For comparing two models,  $M_1$  and  $M_2$ , the BF is defined as the ratio of the two integrated likelihood, i.e.  $B_{12} = \frac{P(y|M_1)}{P(y|M_2)}$ . While BF provides the researcher with a directly interpretable number that quantifies the evidence provided by the data, it requires integration over all parameter space and therefore is computationally intensive, particularly when there are many parameters of interest. Alternatively, BIC can be used and previous studies suggest that it provides good performance in model-based clustering context [20]. The BIC is defined as  $BIC_k = -2 \log P(Y | \Phi_k, M_k) + v_k \log(N)$ , where  $\log P(Y | \Phi_k, M_k)$  is the log-likelihood under model  $M_k$ ,  $Y$  denotes all the data,  $N$  is the sample size and  $v_k$  is the number of parameters. The model with the smallest BIC is usually preferred. The log BF can be approximated by BIC with  $2 \ln(B_{10}) \approx 2(\Delta BIC)$ , where  $B_{10}$  is the BF comparing a model with  $k+1$  classes to a model with  $k$  classes, and  $\Delta BIC$  is the changes between these two models [26,42].

Apart from the above methods, many other approaches are also proposed, such as Deviance Information Criterion (DIC) [53], ‘no small clusters’ criterion [49], the approximate weight of evidence (AWE) [2], integrated classification likelihood (ICL) [5], etc.

## 5. Practical application

In this section, we apply both one-step and stepwise approaches with variable selection priors to the PBC and CAMP datasets introduced in Section 2.

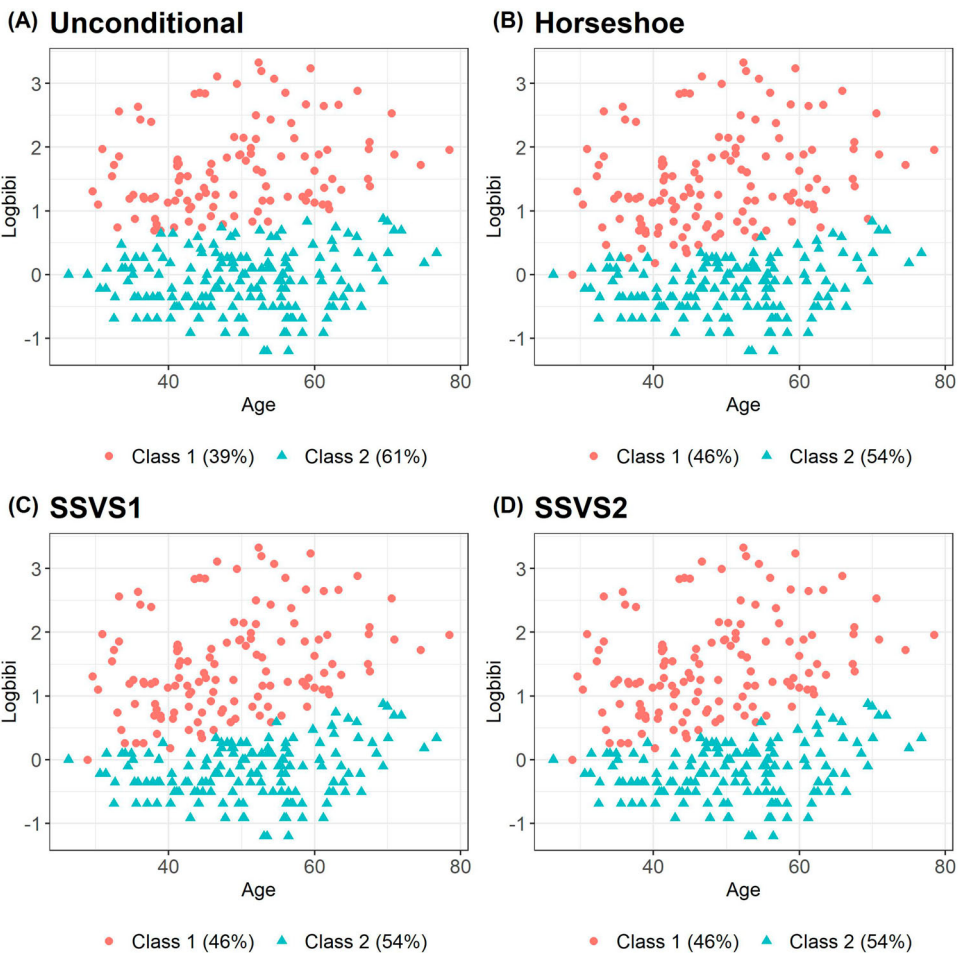
### 5.1. Primary biliary cirrhosis data

For the PBC data, the primary interest is to identify subgroups of patients with similar serum Bilirubin levels and predictors that are associated with these group assignments. Bilirubin is an orange-yellow substance made during the normal breakdown of red blood cells and higher than normal levels of bilirubin may indicate an increased risk of liver problems. Bilirubin was converted to a logarithmic scale (logbili) prior to modeling. Eight predictors were considered in the current analysis, namely treatment (trt, D-penicillamine vs. placebo), edema (edema), alkaline phosphatase (alk.phos), serum cholesterol (chol), serum albumin (albumin), triglycerides (trig), standardized blood clotting time (protime), histologic stage of disease (stage, stage 1 or 2 vs. stage 3 or 4). Here, we initially considered the 312 subjects who participated in the randomized trial. We then removed 30 (9.6%) subjects who had incomplete covariates data. Therefore, the data included in our final analysis consists of 282 subjects. Age was centered and all predictors were standardized such that the means were 0 and variances were 1.

We chose the hyper-parameters of prior distributions to provide weakly information to the parameters of interest. For priors of the growth trajectory parameters, we used  $\mathbf{V}_{k0} = 1000\mathbf{I}_2$ ,  $a_{k0} = 3$  and  $b_{k0} = 0.01$ . We used  $\lambda_1^2 = \dots = \lambda_K^2 = \lambda^2 \sim \text{gamma}(\iota_{a0} = 0.01, \iota_{b0} = 0.01)$ , where  $\iota_{a0}$  and  $\iota_{b0}$  are the shape and rate parameters, respectively. This induces a non-informative prior on the tuning parameter. For the hyper-parameters in SSVS, based on a preliminary analysis, we set  $c_{kj}^2 = 100$  and  $\tau_{kj}^2 = 0.01$  (SSVS1) and  $c_{kj}^2 = 100$  and  $\tau_{kj}^2 = 0.04$  (SSVS2) for all  $k$  and  $j$ . These values induce a hyper-variance of 1 for slab and 0.01 for spike in SSVS1, and a hyper-variance of 4 for slab and 0.04 for spike in SSVS2. For each model, we ran the MCMC with 6000 iterations, discarded the first 1000 iterations as burn-in and kept every 5th iterations. The final chain includes 1000 samples.

The BIC for one- to six-class models are 832.9, 803.8, 822.8, 845.2, 867.5 and 879.4, respectively. And the  $2 \ln(B_{10})$  comparing models of 2-class vs. 1-class, 3-class vs. 2-class, 4-class vs. 3-class, 5-class vs. 4-class, 6-class vs. 5-class are 58.2,  $-38$ ,  $-44.8$ ,  $-44.6$ , and  $-23.8$ , respectively. According to Kass and Raftery recommendation,  $2 \ln(B_{10}) > 10$  is considered as very strong evidence against a simpler model [27]. Therefore, the two-class model provided the best fit to the current data compared to all other models. Therefore, we set  $K = 2$  and refit the one-step approach with predictors and stepwise approaches with predictors based on different variable selection priors described in Section 3. The clustering results for different models are shown in Figure 1 and parameter estimates with 95% credible interval (CR) are shown in Table E1. Overall there were clear patterns of two groups, with Class 1 indicating patients with high Bilirubin and Class 2 indicating patients with low Bilirubin. However, the class proportions were different between the conditional models (i.e. with predictors) and the unconditional model (i.e. without predictors). Specifically, the unconditional model (Figure 1(A)) assigned 39% of patients to Class 1 and 61% to Class 2, whereas conditional models with Horseshoe (Figure 1(B)), SSVS1 (Figure 1(C)) and SSVS2 (Figure 1(D)) priors consistently assigned 46% to Class 1 and 54% to Class 2. This indicates models including covariates changed the patients' posterior class probability, which resulted in higher uncertainty compared to the unconditional model (Figure E1). This difference between conditional and unconditional models resulted in different class proportions.

The covariate effects based on different models were shown in Figure 2. In general, within the same category of models (i.e. one-step or stepwise approaches), all models provided similar coefficient estimation (Figure 2(A)), which was expected given the small number of predictors considered in this analysis. Based on the HS rule, the one-step model with Horseshoe prior identified six important predictors (i.e. albumin, chol, edema, protime, stage, trig), whereas based on the inclusion probability ( $> 0.5$ ), SSVS1 and SSVS2 identified five (i.e. albumin, chol, protime, stage, trig) and two predictors (i.e. chol, trig), respectively (Figure 2(B)). These predictors were shown to be associated with the patients' class membership. For example, a patient with a higher albumin value was more likely to be assigned to Class 1 when compared to Class 2. It is interesting to observe that the estimated coefficients from the conditional models (i.e. Horseshoe-onestep, SSVS1-onestep, SSVS2-onestep) moved towards the origin as compared to unconditional models (i.e. Horseshoe-stepwise, SSVS1-stepwise, SSVS2-stepwise). This attenuation was due to the increased number of patients in Class 1 in the conditional models. For the stepwise models, Horseshoe identified the same set of predictors as those in one-step models. The

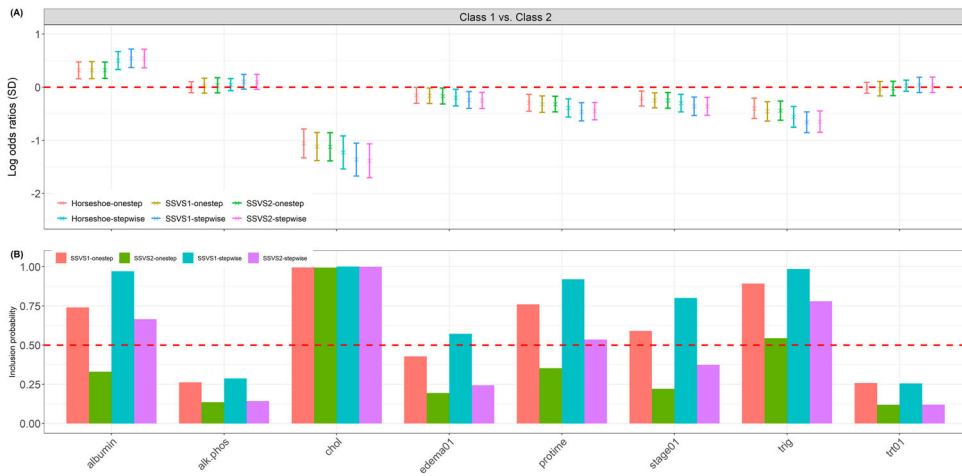


**Figure 1.** Observed trajectory patterns of two-class models based on different variable selection priors in PBC data. ‘Unconditional’ refers to model without including predictors. SSVS1:  $c^2 = 100$ ,  $\tau^2 = 0.01$ ; SSVS2:  $c^2 = 100$ ,  $\tau^2 = 0.04$ .

SSVS1-stepwise model also identified the same set of predictors as the shrinkage priors, whereas the SSVS2-stepwise only kept four predictors (albumin, chol, protime, trig) in the model.

### 5.2. Childhood asthma management program data

For the CAMP data, the primary interest is to characterize the FEV1 trajectory at the population level and the individual level, as well as to identify significant predictors of abnormal FEV1 longitudinal patterns. In the current analysis, we included 657 participants who contributed to a total of 15,138 measurements. On average, each participant had 23 measurements. Of these participants, there are 450 (68.5%) white and 391 (60%) male. The mean (SD) of age is 8.97 (2.10). For predictors of the class membership, here we considered 12 baseline covariates and their first-order interaction (66 combinations) as



**Figure 2.** Covariate effects of two-class models in PBC data. (A) Log odds ratios estimated from different models. (B) Inclusion probabilities estimated from different SSVS models. SSVS1:  $c^2 = 100$ ,  $\tau^2 = 0.01$ ; SSVS2:  $c^2 = 100$ ,  $\tau^2 = 0.04$ .

candidate predictors, which induce a total of 78 covariates. These 12 baseline covariates are race (white vs. other), any\_pets (any pets, yes vs. no), agehome (age of current home, years), ast.age (age when asthma confirmed by MD), whitecell (white blood cell count), hemoglobin, gas.stove (gas cooking stove, range or oven), wood.stove (yes vs. no), hosp.ast (child ever in hospital for asthma, yes vs. no), mother.ast (mother has asthma, yes vs. no). To model the FEV1 patterns, we considered a quadratic function with random intercept and slope.

We chose similar hyper-parameters for the growth trajectory model as the previous example. Specifically, we used  $V_{k0} = 1000I_2$ ,  $a_{k0} = 3$  and  $b_{k0} = 0.01$ ,  $r = 4$  and  $R_k = 3I_2$ . For hyper-parameter of Lasso prior, here we used  $\lambda_1^2 = \dots = \lambda_K^2 = \lambda^2 \sim \text{gamma}(\iota_{a0}, \iota_{b0})$ , where  $\iota_{a0} = 10$  and  $\iota_{b0} = 1/\hat{\lambda}^2$ , respectively. The  $\hat{\lambda}$  denotes the maximum likelihood estimate of  $\lambda$ . This yields a gamma distribution with a mean of 10 times of the  $\hat{\lambda}$ . For hyper-parameters of SSVS, we considered four different settings. In SSVS1, we set  $c_{kj}^2 = 100$  and  $\tau_{kj}^2 = 0.01$  which yielded a hyper-variance of 1 for slab and 0.01 for spike. In SSVS2, we set  $c_{kj}^2 = 100$  and  $\tau_{kj}^2 = 0.1$  which yielded a hyper-variance of 10 for slab and 0.1 for spike. In SSVS3, we set  $c_{kj}^2 = 100$  and  $\tau_{kj}^2 = 0.3$  which yielded a hyper-variance of 30 for slab and 0.3 for spike. In SSVS4, we set  $c_{kj}^2 = 100$  and  $\tau_{kj}^2 = 0.5$  which yielded a hyper-variance of 50 for slab and 0.5 for spike. For each model, we ran the MCMC with 60,000 iterations, discarded the first 10,000 iterations as burn-in and kept every 50th iterations as burn-in and kept every. The final chain of each model includes 1000 samples.

The BIC for one- to six-class models are 9556, 8206, 8028, 7936, 7938 and 7932, respectively. While a six-class model yielded the lowest BIC, the  $2 \ln(B_{10})$  of comparing the six-class model to a four-class model was 7.6, which did not show a significant improvement in terms of model fitting. In contrast, the  $2 \ln(B_{10})$  comparing a four-class model to a three-class model was 184, which showed a significant improvement ( $> 10$ ) and justified

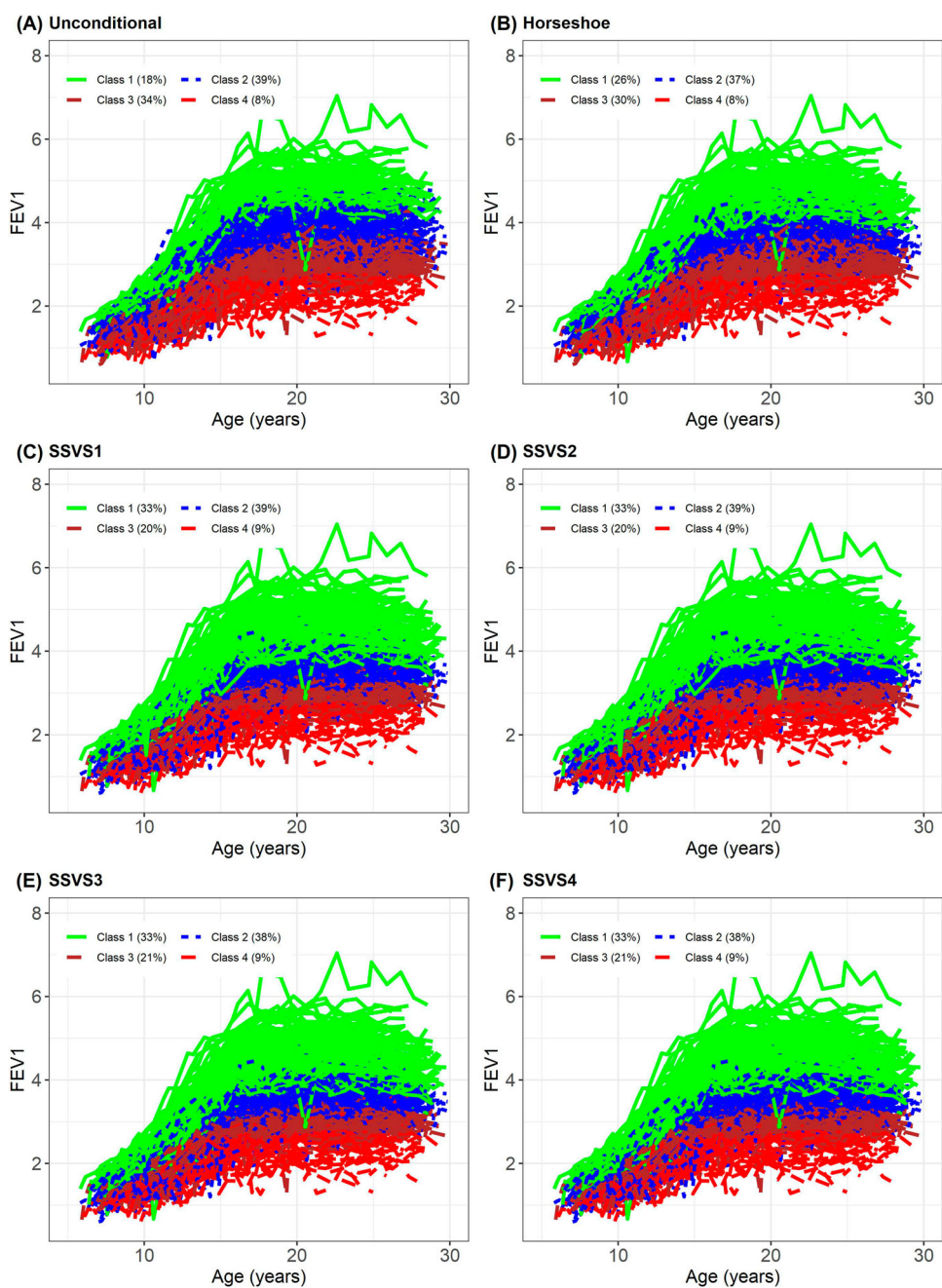
using a more complicated model. This evidence indicated that a four-class model provided a good fit to the current data.

The observed trajectory patterns of the four-class model based on different variable selection priors are shown in Figure 3 and the parameter estimates with 95% CR are shown in Table E2. The four FEV1 growth patterns were similar across different models. Class 1 represented normal FEV1 patterns and patients from this class had the best lung function compared to patients from all other classes. Class 2 and Class 3 represented mild and moderate reduced FEV1 patterns, while Class 4 represented severe reduced FEV1 patterns, which suggested that patients from this group had the worst lung function compared to the patients from all other classes. However, the proportion of patients assigned to different classes differ across different models. Specifically, the proportion of Class 2 (ranged 37% to 39%) and Class 4 (8% to 9%) were relatively stable across different models whereas Class 1 (ranged 18% to 33%) and Class 3 (20% to 34%) were substantially different among different models. The difference in class proportion can be partially explained by the predictors kept by each model. Similar to the previous example, the models including covariates changed the patients' posterior class probability, which also resulted in different class proportions (Figure E2). Consider Class 1 (with the best FEV1 patterns) as the reference category, the predictors selected by each model were shown in Figure 4. Unlike the previous example when only eight predictors were considered, in the current example different models kept a different number of variables and there is a lack of consistency. Take the comparison between Class 2 vs. Class 1 for example, Horseshoe prior identified only one predictor (i.e. bPREFEV) in either the one-step or stepwise approach. For SSVS prior, we observed that the larger the hyper-variance (for either spike or slab), the smaller number of predictors were kept in the model. For example, SSVS1 resulted in a model with the largest number of predictors in either the one-step or stepwise approach. In contrast, SSVS4 yielded the smallest model. Similar results were observed for the comparison of Class 3 vs. Class 1 and Class 4 vs. Class 1.

To compare the behavior of different models, we provided the estimated posterior distribution of a selected predictor (for variable bPREFEV) in Figure E3. It was evident that the posterior density of the Horseshoe model was tighter (i.e. with smaller SD) compared to other models. And the estimate shifted away from zero when higher class (i.e. the class with lower FEV1 patterns) is compared to Class 1. This is expected given Class 1 was the group with the highest mean FEV1 over time and the higher the class number, the lower the FEV1 and, therefore, the larger the difference (Figure 3).

## 6. Simulation study

To further investigate the performance of variable selection in both one-step and stepwise approaches, in this section, we conducted a small simulation study. We only considered the longitudinal setting in this simulation. We generated the data from mixture models with  $K$  classes using quadratic mean functions with random intercept and slope. We considered  $K = 2, 3, 4$  which are commonly seen in practice. The trajectory of each class was generated to mimic the FEV1 trajectories from the CAMP. Specifically, we considered two scenarios (Figure E4), representing high separation (Scenario 1) among classes and low separation among classes (Scenario 2). The total sample size was set at  $N = 600$ , and the number of



**Figure 3.** Observed trajectory patterns of four-class models based on different variable selection priors in CAMP data. 'Unconditional' refers to model without including predictors. SSVS1:  $c^2 = 100$ ,  $\tau^2 = 0.01$ ; SSVS2:  $c^2 = 100$ ,  $\tau^2 = 0.1$ ; SSVS3:  $c^2 = 100$ ,  $\tau^2 = 0.3$ ; SSVS4:  $c^2 = 100$ ,  $\tau^2 = 0.5$ .



**Figure 4.** Variable selected by different variable selection approaches. SSVS1:  $c^2 = 100$ ,  $\tau^2 = 0.01$ ; SSVS2:  $c^2 = 100$ ,  $\tau^2 = 0.1$ ; SSVS3:  $c^2 = 100$ ,  $\tau^2 = 0.3$ ; SSVS4:  $c^2 = 100$ ,  $\tau^2 = 0.5$ .

measurements per individual  $n_i$  was set to have an equal probability of being 1 to  $n_{max} = 10$ , using a uniform distribution.

For subject-level covariates (e.g. baseline covariates), we generated a  $p = 20$  dimensional covariates  $\mathbf{X}$  from a multivariate normal distribution  $MVN(\mathbf{0}, \mathbf{\Lambda})$ , where  $\mathbf{\Lambda}$  has exchangeable correlation structure with  $\rho = 0.5$ , representing moderate correlation between predictors. The regression coefficients were set to be sparse to represent common scenarios often seen in practice. For identifiability purpose, we set  $\beta_{10} = 0$  and  $\beta_1 = \mathbf{0}$ , and therefore, the first class was considered as a reference category. For  $K = 2$ , we set  $\beta_{20} = \log(2)$  and  $\beta_2 = (\log(2), \log(5), \log(5), \log(8), \log(8), 0, \dots, 0)$ , with  $p-5$  zeros. For  $K = 3$ , we set  $\beta_{20} = \log(2)$ ,  $\beta_{30} = -\log(2)$ ,  $\beta_2 = (\log(2), \log(5), \log(5), \log(8), \log(8), 0, \dots, 0)$ ,  $\beta_3 = (0, \dots, 0, \log(2), \log(5), \log(5), \log(8), \log(8))$  with  $p-5$  zeros for both  $\beta_2$  and  $\beta_3$ . And for  $K = 4$ , we set  $\beta_{20} = \log(2)$ ,  $\beta_{30} = -\log(1.2)$ ,  $\beta_{40} = -\log(2)$ ,  $\beta_2 = (\log(2), \log(5), \log(5), \log(8), \log(8), 0, \dots, 0)$ ,

$\beta_3 = (0, \dots, 0, \log(2), \log(5), \log(5), \log(8), \log(8), 0, \dots, 0)$ ,  $\beta_4 = (0, \dots, 0, \log(2), \log(5), \log(5), \log(8), \log(8), \log(8))$  with  $p-5$  zeros for  $\beta_2$ ,  $\beta_3$  and  $\beta_4$ , respectively.

To evaluate the performance of different models, we used several evaluation metrics. To evaluate the clustering performance, we considered the correct classification rate (cRate) and adjusted rand index (aRand). Both of these two indices measure the agreement between the estimated class membership and the true class membership, and the higher cRate or aRand indicates a better agreement. To evaluate the variable selection performance, we used the true positive rate (TPR), true negative rate (TNR), accuracy (ACY) and root mean square error (RMSE). These metrics were calculated over all the  $p(K-1)$  estimated coefficients from each model to represent the overall model performance in clustering and parameter estimation. For each setting in each scenario, we simulated 25 datasets, and both one-step and stepwise approaches with Horseshoe or SSVS priors were applied to these datasets to estimate the class and identify predictors of the class membership, assuming the number of classes  $K$  is known. The mean and SD of these metrics were reported.

The simulation results for Scenario 1 (high separation) and Scenario 2 (low separation) are provided in Tables 1 and 2, respectively. Specifically, in Scenario 1 and when  $K = 2$ , all approaches performed reasonably well in identifying the true class membership. This is expected given the high separation between classes. However, it was interesting that one-step approaches yielded slightly lower cRates and aRands compared to stepwise approaches, while within either the one-step or stepwise approaches, the clustering performance was similar among each other. On the other hand, one-step approaches with Horseshoe and SSVS1 yielded slightly better variable selection performance in terms of accuracy, compared to their stepwise model counterpart. For example, Horseshoe-onestep yielded higher accuracy compared to Horseshoe-stepwise (0.97 vs. 0.94). Of note, the one-step or stepwise approach with Horseshoe prior correctly identified all non-zero coefficients in all datasets, with a mean (SD) TPR of 1 (0). Similar results were observed for  $K = 3$  and  $K = 4$ . In particular, the Horseshoe priors yielded higher TPR and ACY compared to SSVS priors (i.e. SSVS1 and SSVS2) for either a one-step or stepwise approach. Of note, Horseshoe prior maintained high variable selection accuracy in both one-step and stepwise approaches. However, the one-step model with Horseshoe prior generally yielded better variable selection performance compared to the stepwise model with Horseshoe prior in all settings and scenarios we considered. Similar results were found in Scenario 2.

## 7. Discussion

In this study, we developed and compared two approaches for variable selection in the context of the mixture model to determine the variables affecting the probabilities of the mixture components: the one-step approach and the stepwise approach. Horseshoe prior and SSVS prior were used within these two approaches to select important variables. We also developed MCMC algorithms based on Gibbs sampling to estimate the posterior distribution of model parameters. The proposed models were applied to two clinical datasets with the goal of finding disease phenotypes and their predictors. A simulation study was carried out to investigate the clustering and variable selection performance under different settings and scenarios.



**Table 1.** Simulation results for Scenario 1 (high separation).

	cRate	aRand	TPR	TNR	ACY	RMSE
$K = 2, N = 600$						
Horseshoe-onestep	0.99 (0)	0.97 (0.01)	1 (0)	0.95 (0.05)	0.97 (0.04)	0.28 (0.05)
SSVS1-onestep	0.99 (0)	0.97 (0.01)	0.98 (0.06)	0.68 (0.17)	0.75 (0.13)	0.27 (0.04)
SSVS2-onestep	0.99 (0)	0.97 (0.01)	0.83 (0.07)	1 (0.01)	0.96 (0.02)	0.27 (0.04)
Horseshoe-stepwise	1 (0)	1 (0.01)	1 (0)	0.91 (0.08)	0.94 (0.06)	0.19 (0.06)
SSVS1-stepwise	1 (0)	1 (0.01)	0.99 (0.04)	0.49 (0.15)	0.62 (0.11)	0.34 (0.11)
SSVS2-stepwise	1 (0)	1 (0.01)	0.93 (0.1)	0.99 (0.03)	0.97 (0.03)	0.33 (0.11)
$K = 3, N = 600$						
Horseshoe-onestep	0.85 (0.03)	0.65 (0.05)	1 (0)	0.84 (0.05)	0.88 (0.04)	0.94 (0.04)
SSVS1-onestep	0.84 (0.03)	0.64 (0.05)	0.94 (0.05)	0.6 (0.18)	0.69 (0.12)	0.97 (0.05)
SSVS2-onestep	0.84 (0.03)	0.64 (0.05)	0.39 (0.13)	0.95 (0.05)	0.81 (0.05)	0.96 (0.05)
Horseshoe-stepwise	0.9 (0.02)	0.75 (0.04)	1 (0)	0.71 (0.09)	0.79 (0.07)	1.02 (0.05)
SSVS1-stepwise	0.9 (0.02)	0.75 (0.04)	0.81 (0.11)	0.53 (0.1)	0.6 (0.06)	1.08 (0.06)
SSVS2-stepwise	0.9 (0.02)	0.75 (0.04)	0.42 (0.14)	0.85 (0.04)	0.74 (0.04)	1.07 (0.06)
$K = 4, N = 600$						
Horseshoe-onestep	0.85 (0.02)	0.7 (0.03)	1 (0)	0.76 (0.06)	0.82 (0.04)	0.9 (0.1)
SSVS1-onestep	0.85 (0.02)	0.69 (0.03)	0.96 (0.04)	0.38 (0.16)	0.53 (0.12)	0.98 (0.12)
SSVS2-onestep	0.85 (0.01)	0.69 (0.03)	0.59 (0.07)	0.82 (0.06)	0.76 (0.05)	0.96 (0.11)
Horseshoe-stepwise	0.86 (0.02)	0.73 (0.03)	1 (0)	0.64 (0.07)	0.73 (0.05)	0.93 (0.05)
SSVS1-stepwise	0.86 (0.02)	0.73 (0.03)	0.78 (0.08)	0.37 (0.06)	0.48 (0.05)	1.03 (0.07)
SSVS2-stepwise	0.86 (0.02)	0.73 (0.03)	0.33 (0.07)	0.73 (0.05)	0.63 (0.04)	1.01 (0.07)

Notes: Results are presented as mean (SD) over all simulated datasets. cRate: correct classification rate; aRand: adjusted Rand index; TPR: true positive rate; TNR: true negative rate; ACY: accuracy; RMSE: root mean square error. SSVS1:  $c^2 = 100, \tau^2 = 0.01$ ; SSVS2:  $c^2 = 100, \tau^2 = 0.1$ .

In our practical applications, we considered BIC and BF to determine the number of clusters in the proposed mixture models. While these indices have been widely used in many model-based cluster analyses, they can be infeasible when there are a larger number of candidate models needed to fit. Alternatively, one can treat  $K$  as a random variable, and the inference of  $K$  is considered as part of the modeling process. In this regard, Richardson and Green proposed a reversible jump MCMC method, which allows the sampler jumps between parameter subspaces of different dimensionality [48]. Stephens proposed a birth-and-death process, in which the MCMC sampler allows the number of components to vary by allowing new components to be ‘born’ and existing components to ‘die’ [54]. Moreover, Bayesian non-parametric methods such as the Dirichlet process mixture [15] can also be employed.

A previous study by Vermunt suggested two improved stepwise approaches to account for the uncertainty of the class assignment estimated from a latent class model [57]. In our current study, we also compared the one-step and stepwise approaches but instead focus on estimating the covariate effects and selecting the variables that are associated with the class assignment. It is worth noting that even if covariate selection works well with both one-step and stepwise approaches, some covariate effects are expected to be downward biased given the penalization effect from either the shrinkage or spike-and-slab prior, which is also reflected in the estimates of RMSE in our simulation study (Tables 1 and 2). Moreover, for the stepwise approach, if the uncertainty of the class membership is not taken into account, a systematic underestimation of covariate effects may occur [57].

The number of predictors considered in the current study is not large as compared to many medical studies which may have hundreds or thousands of predictors, such as genetics studies. The definition of large here is more relevant to the computational scale in which

**Table 2.** Simulation results for Scenario 2 (low separation).

	cRate	aRand	TPR	TNR	ACY	RMSE
$K = 2, N = 600$						
Horseshoe-onestep	0.88 (0.02)	0.57 (0.07)	1 (0)	0.97 (0.04)	0.98 (0.03)	0.71 (0.08)
SSVS1-onestep	0.88 (0.02)	0.56 (0.07)	0.88 (0.1)	0.93 (0.06)	0.92 (0.05)	0.69 (0.09)
SSVS2-onestep	0.87 (0.02)	0.56 (0.06)	0.06 (0.11)	1 (0)	0.76 (0.03)	0.69 (0.09)
Horseshoe-stepwise	0.91 (0.01)	0.66 (0.04)	1 (0)	0.95 (0.05)	0.96 (0.04)	0.62 (0.17)
SSVS1-stepwise	0.91 (0.01)	0.66 (0.04)	0.93 (0.1)	0.83 (0.08)	0.86 (0.06)	0.6 (0.19)
SSVS2-stepwise	0.91 (0.01)	0.66 (0.04)	0.41 (0.19)	1 (0)	0.85 (0.05)	0.6 (0.19)
$K = 3, N = 600$						
Horseshoe-onestep	0.81 (0.04)	0.54 (0.06)	1 (0)	0.94 (0.04)	0.95 (0.03)	0.73 (0.08)
SSVS1-onestep	0.81 (0.04)	0.54 (0.06)	0.86 (0.1)	0.71 (0.09)	0.74 (0.05)	0.73 (0.09)
SSVS2-onestep	0.8 (0.04)	0.54 (0.06)	0.16 (0.14)	1 (0.01)	0.79 (0.03)	0.73 (0.09)
Horseshoe-stepwise	0.81 (0.05)	0.56 (0.06)	1 (0)	0.87 (0.09)	0.9 (0.07)	0.73 (0.16)
SSVS1-stepwise	0.81 (0.05)	0.56 (0.06)	0.7 (0.13)	0.63 (0.08)	0.65 (0.06)	0.73 (0.19)
SSVS2-stepwise	0.81 (0.05)	0.56 (0.06)	0.2 (0.12)	0.95 (0.04)	0.77 (0.04)	0.73 (0.18)
$K = 4, N = 600$						
Horseshoe-onestep	0.77 (0.03)	0.55 (0.04)	1 (0)	0.85 (0.06)	0.89 (0.04)	0.68 (0.08)
SSVS1-onestep	0.76 (0.03)	0.55 (0.04)	0.88 (0.08)	0.51 (0.1)	0.6 (0.08)	0.69 (0.09)
SSVS2-onestep	0.76 (0.03)	0.54 (0.04)	0.3 (0.12)	0.97 (0.03)	0.8 (0.04)	0.68 (0.09)
Horseshoe-stepwise	0.78 (0.03)	0.56 (0.04)	1 (0)	0.81 (0.09)	0.85 (0.07)	0.67 (0.13)
SSVS1-stepwise	0.78 (0.03)	0.56 (0.04)	0.69 (0.13)	0.49 (0.08)	0.54 (0.05)	0.68 (0.16)
SSVS2-stepwise	0.78 (0.03)	0.56 (0.04)	0.21 (0.07)	0.87 (0.06)	0.71 (0.04)	0.68 (0.16)

Notes: Results are presented as mean (SD) over all simulated dataset. cRate: correct classification rate; aRand: adjusted Rand index; TPR: true positive rate; TNR: true negative rate; ACY: accuracy; RMSE: root mean square error. SSVS1:  $c^2 = 100, \tau^2 = 0.01$ ; SSVS2:  $c^2 = 100, \tau^2 = 0.1$ .

estimating  $2^{p(K-1)}$  models would be infeasible when  $p$  and/or  $K$  increase. Therefore, the variable selection provides a convenient tool to identify important predictors and to facilitate better interpretation. Furthermore, the literature has not reached a consensus in terms of whether adding covariates can improve the class recovery or not (i.e. subjects are classified more accurately). Huang *et al.* [25] found that whether covariates were included in the model and which covariates were included could have an impact on the class assignment, which highlights how the inclusion of covariates and the decision of what covariates to include can dramatically influence the nature of the latent class variable. A previous study also showed that deciding the number of latent classes without predictors of latent class (i.e. via an unconditional model), and including the latent class predictors into the model subsequently lead to good estimates for all model parameters [29]. Similarly, our study finds that the one-step approach with variable selection priors resulted in larger uncertainty in the posterior class probability compared to the unconditional model in which no predictors were included (Figures E1 & E2) and, therefore, resulted in worse clustering performance.

The implementation of SSVS in the current study considers several sets of hyperparameters in both Practical Applications and Simulation Studies. It is recognized that SSVS is sensitive to the specification of its hyperparameters  $c^2$  and  $\tau^2$ , as observed in our analyses, and these hyperparameters are rarely known in practice. Moreover, tuning these hyperparameters is a difficult process in many applications. While our implementation considers different specifications of these hyperparameters, careful selection and tuning these parameters by a more extensive search may lead to refined results. In contrast, the Horseshoe prior is free of tuning parameters and was found to have desirable variable selection performance in our practical applications and simulation study. However, the Horseshoe prior has also

been criticized for not providing sufficient shrinkage to large coefficients, which is undesirable when the parameters are weakly identified. Its variants such as regularized Horseshoe [47] that allows users to specify a minimum level of regularization to the largest values can be considered in the future study to refine the model.

The current proposed model refers to a basic model where  $f_k$  is a Gaussian distribution, and classes (or components) are used to model unobserved heterogeneity. More complicated models can be considered in future studies. In this regard, variable selection priors can be adapted to mixture models such as latent class analysis [3], shape invariant mixture model [36], Bayesian consensus clustering [34] and Bayesian model for multivariate continuous and discrete longitudinal data [30].

## Acknowledgments

The authors thank the reviewers for their comments and suggestions, which substantially improve the quality of the manuscript.

## Data availability statement

The PBC data are available in R *survival* package and also in Appendix D of Fleming and Harrington [16]. The CAMP data are not publicly available as there is currently no ethical approval to share data. However, these data can be requested at the National Heart, Lung, and Blood Institute (NHLBI) website (<https://biolincc.nhlbi.nih.gov/studies/camp/>).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This research was supported by the Research Initiation Grant (Queen's University) to the first author.

## References

- [1] J.E. Cavanaugh and A.A. Neath, *The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements*, Wiley Interdiscip. Rev. Comput. Stat. 11 (2019), pp. e1460
- [2] J.D. Banfield and A.E. Raftery, *Model-based Gaussian and non-Gaussian clustering*, Biometrics 49 (1993), pp. 803–821.
- [3] F. Bartolucci, G.E. Montanari, and S. Pandolfi, *Latent ignorability and item selection for nursing home case-mix evaluation*, J. Classification 35 (2018), pp. 172–193.
- [4] A. Bhattacharya, D. Pati, N.S. Pillai, and D.B. Dunson, *Dirichlet–Laplace priors for optimal shrinkage*, J. Am. Stat. Assoc. 110 (2015), pp. 1479–1490.
- [5] C. Biernacki, G. Celeux, and G. Govaert, *Assessing a mixture model for clustering with the integrated completed likelihood*, IEEE. Trans. Pattern Anal. Mach. Intell. 22 (2000), pp. 719–725.
- [6] CAMP Research Group, *The childhood asthma management program (CAMP): Design, rationale, and methods*, Control. Clin. Trials 20 (1999), pp. 91–120.
- [7] CAMP Research Group, *Long-term effects of budesonide or nedocromil in children with asthma*, N. Engl. J. Med. 343 (2000), pp. 1054–1063.
- [8] B.P. Carlin and S. Chib, *Bayesian model choice via Markov chain Monte Carlo methods*, J. R. Stat. Soc. B (Methodol.) 9 (1995), pp. 473–484.
- [9] C.M. Carvalho, N.G. Polson, and J.G. Scott, *The horseshoe estimator for sparse signals*, Biometrika. 97 (2010 Jun 1), pp. 465–480.

- [10] C.M. Carvalho, N.G. Polson, and J.G. Scott, *The horseshoe estimator for sparse signals*, *Biometrika* 97 (2010), pp. 465–480.
- [11] C.R. Colder, R.T. Campbell, E. Ruel, J.L. Richardson, and B.R. Flay, *A finite mixture model of growth trajectories of adolescent alcohol use: Predictors and consequences*, *J. Consult. Clin. Psychol.* 70 (2002), pp. 976–985.
- [12] N. Dean and A.E. Raftery, *Latent class analysis variable selection*, *Ann. Inst. Statist. Math.* 62 (2010), pp. 11.
- [13] P. Dellaportas, J.J. Forster, and I. Ntzoufras, *Bayesian variable selection using the Gibbs sampler*, *Biostatistics* 5 (2000), pp. 273–286.
- [14] E.R. Dickson, P.M. Grambsch, T.R. Fleming, L.D. Fisher, and A. Langworthy, *Prognosis in primary biliary cirrhosis: Model for decision making*, *Hepatology* 10 (1989), pp. 1–7.
- [15] M.D. Escobar, *Estimating normal means with a Dirichlet process prior*, *J. Am. Statist. Assoc.* 89 (1994), pp. 268–277.
- [16] T. Fleming and D. Harrington, *Counting Processes and Survival Analysis*, John Wiley & Sons, Inc., New York, 1991.
- [17] A. Flynt and N. Dean, *Growth mixture modeling with measurement selection*, *J. Classification* 36 (2019), pp. 3–25.
- [18] M. Fop and T.B. Murphy, *Variable selection methods for model-based clustering*, *Stat. Surv.* 12 (2018), pp. 18–65.
- [19] M. Fop, K.M. Smart, and T.B. Murphy, *Variable selection for latent class analysis with application to low back pain diagnosis*, *Ann. Appl. Stat.* 11 (2017), pp. 2080–2110.
- [20] C. Fraley and A.E. Raftery, *How many clusters? Which clustering method? Answers via model-based cluster analysis*, *Comput. J.* 41 (1998), pp. 578–588.
- [21] C. Fraley and A.E. Raftery, *Model-based clustering, discriminant analysis, and density estimation*, *J. Am. Statist. Assoc.* 97 (2002), pp. 611–631.
- [22] E.I. George and R.E. McCulloch, *Variable selection via Gibbs sampling*, *J. Am. Statist. Assoc.* 88 (1993), pp. 881–889.
- [23] B. Grun and F. Leisch, *FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters* *J. Stat. Softw.* 28 (2008), pp. 1–35. <https://doi.org/10.18637/jss.v028.i04>.
- [24] J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky, *Bayesian model averaging: a tutorial*, *Stat. Sci.* 0 (1999), pp. 382–401.
- [25] D. Huang, M.-L. Brecht, M. Hara, and Y.-I. Hser, *Influences of covariates on growth mixture modeling*, *J. Drug Issues* 40 (2010), pp. 173–194.
- [26] B.L. Jones, D.S. Nagin, and K. Roeder, *A SAS procedure based on mixture models for estimating developmental trajectories*, *Sociol. Methods Res.* 29 (2001), pp. 374–393.
- [27] R.E. Kass and A.E. Raftery, *Bayes factors*, *J. Am. Statist. Assoc.* 90 (1995), pp. 773–795.
- [28] A. Khalili and J. Chen, *Variable selection in finite mixture of regression models*, *J. Am. Statist. Assoc.* 102 (2007), pp. 1025–1038.
- [29] M. Kim, J. Vermunt, Z. Bakk, T. Jaki, and M.L. Van Horn, *Modeling predictors of latent classes in regression mixture models*, *Struct. Equ. Model.* 23 (2016), pp. 601–614.
- [30] A. Komárek and L. Komárková, *Clustering for multivariate continuous and discrete longitudinal data*, *Ann. Appl. Stat.* 7 (2013), pp. 177–200.
- [31] L. Kuo and B. Mallick, *Variable selection for regression models*, *Sankhyā: The Indian Journal of Statistics, Series B* (1998), pp. 65–81.
- [32] F. Leisch, *FlexMix: A general framework for finite mixture models and latent glass regression in R*, 2004.
- [33] Y. Lo, N.R. Mendell, and D.B. Rubin, *Testing the number of components in a normal mixture*, *Biometrika* 88 (2001), pp. 767–778.
- [34] E.F. Lock and D.B. Dunson, *Bayesian consensus clustering*, *Bioinformatics* 29 (2013), pp. 2610–2616.
- [35] Z. Lu, R.E. Foong, K. Kowalik, T.J. Moraes, A. Boyce, A. Dubeau, S. Balkovec, P.M. Gustafsson, A.B. Becker, P.J. Mandhane, S.E. Turvey, W. Lou, F. Ratjen, M. Sears, and P. Subbarao, *Ventilation inhomogeneity in infants with recurrent wheezing*, *Thorax* 73 (2018), pp. 936–941.

- [36] Z. Lu and W. Lou, *Shape invariant mixture model for clustering non-linear longitudinal growth trajectories*, *Stat. Methods Med. Res.* 28 (2019), pp. 3769–3784.
- [37] Z. Lu and W. Lou, *Bayesian approaches to variable selection: A comparative study from practical perspectives*, *Int. J. Biostat.* (2021).
- [38] A. Lykou and I. Ntzoufras, *On Bayesian lasso variable selection and the specification of the shrinkage parameter*, *Stat. Comput.* 23 (2013), pp. 361–390.
- [39] E. Makalic and D.F. Schmidt, *A simple sampler for the horseshoe estimator*, *IEEE Signal Process. Lett.* 23 (2015), pp. 179–182.
- [40] G. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, 2004.
- [41] P.D. McNicholas, *Model-based clustering*, *J. Classification* 33 (2016), pp. 331–373.
- [42] D. Nagin, *Group-based Modeling of Development*, Harvard University Press, 2005.
- [43] K.L. Nylund, T. Asparouhov, and B.O. Muthén, *Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study*, *Struct. Equ. Model.* 14 (2007), pp. 535–569.
- [44] P.M. O’Byrne, S. Pedersen, C.J. Lamm, W.C. Tan, and W.W. Busse, *Severe exacerbations and decline in lung function in asthma*, *Am. J. Respir. Crit. Care Med.* 179 (2009), pp. 19–24.
- [45] R.B. O’Hara and M.J. Sillanpää, *A review of Bayesian variable selection methods: What, how and which*, *Bayesian Anal.* 4 (2009), pp. 85–117.
- [46] T. Park and G. Casella, *The Bayesian Lasso*, *J. Am. Statist. Assoc.* 103 (2008), pp. 681–686.
- [47] J. Piironen and A. Vehtari, *Sparsity information and regularization in the horseshoe and other shrinkage priors*, *Electron. J. Stat.* 11 (2017), pp. 5018–5051.
- [48] S. Richardson and P.J. Green, *On Bayesian analysis of mixtures with an unknown number of components (with discussion)*, *J. R. Stat. Soc. B (Stat. Methodol.)* 59 (1997), pp. 731–792.
- [49] J. Rousseau and K. Mengersen, *Asymptotic behaviour of the posterior distribution in overfitted mixture models*, *J. R. Stat. Soc. B (Stat. Methodol.)* 73 (2011), pp. 689–710.
- [50] G. Schwarz, *Estimating the dimension of a model*, *Ann. Stat.* 6 (1978), pp. 461–464.
- [51] C. Silvestre, M.G.M.S. Cardoso, and M. Figueiredo, *Feature selection for clustering categorical data with an embedded modelling approach*, *Expert Syst.* 32 (2015), pp. 444–453.
- [52] V. Siroux, X. Basagaña, A. Boudier, I. Pin, J. Garcia-Aymerich, A. Vesin, R. Slama, D. Jarvis, J.M. Anto, F. Kauffmann, and J. Sunyer, *Identifying adult asthma phenotypes using a clustering approach*, *Eur. Respir. J.* 38 (2011), pp. 310–317.
- [53] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van Der Linde, *Bayesian measures of model complexity and fit*, *J. R. Stat. Soc. B (Stat. Methodol.)* 64 (2002), pp. 583–639.
- [54] M. Stephens, *Bayesian analysis of mixture models with an unknown number of components: An alternative to reversible jump methods*, *Ann. Stat.* 28 (2000), pp. 40–74.
- [55] M. Stephens, *Dealing with label switching in mixture models*, *J. R. Stat. Soc. B (Stat. Methodol.)* 62 (2000), pp. 795–809.
- [56] R. Tibshirani, *Regression shrinkage and selection via the lasso*, *J. R. Stat. Soc. B (Methodol.)* 58 (1996), pp. 267–288.
- [57] J.K. Vermunt, *Latent class modeling with covariates: Two improved three-step approaches*, *Polit. Anal.* 18 (2010), pp. 450–469.
- [58] L. Wasserman, *Bayesian model selection and model averaging*, *J. Math. Psychol.* 44 (2000), pp. 92–107.
- [59] B. Wu, *Sparse cluster analysis of large-scale discrete variables with application to single nucleotide polymorphism data*, *J. Appl. Stat.* 40 (2013), pp. 358–367.
- [60] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzzo, *Model-based clustering and data transformations for gene expression data*, *Bioinformatics* 17 (2001), pp. 977–987.
- [61] Y. Zhang, M. Brady, and S. Smith, *Segmentation of brain mr images through a hidden Markov random field model and the expectation-maximization algorithm*, *IEEE Trans. Med. Imaging* 20 (2001), pp. 45–57.