Biostatistics (2023) **24**, 2, *pp.* 425–442
doi: 10.1093/biostatistics/kxab038
Advance Access publication on November 19, 2021

Bayesian finite mixture of regression analysis for cancer based on histopathological imaging–environment interactions

YUNJU IM, YUAN HUANG

Department of Biostatistics, Yale School of Public Health, 60 College ST, New Haven, CT, USA

AIXIN TAN

Department of Statistics and Actuarial Science, University of Iowa, 259 Schaeffer Hall, Iowa City, IA, USA

SHUANGGE MA*

Department of Biostatistics, Yale School of Public Health, 60 College ST, New Haven, CT, USA
shuangge.ma@yale.edu

SUMMARY

Cancer is a heterogeneous disease. Finite mixture of regression (FMR)—as an important heterogeneity analysis technique when an outcome variable is present—has been extensively employed in cancer research, revealing important differences in the associations between a cancer outcome/phenotype and covariates. Cancer FMR analysis has been based on clinical, demographic, and omics variables. A relatively recent and alternative source of data comes from histopathological images. Histopathological images have been long used for cancer diagnosis and staging. Recently, it has been shown that high-dimensional histopathological image features, which are extracted using automated digital image processing pipelines, are effective for modeling cancer outcomes/phenotypes. Histopathological imaging–environment interaction analysis has been further developed to expand the scope of cancer modeling and histopathological imaging-based analysis. Motivated by the significance of cancer FMR analysis and a still strong demand for more effective methods, in this article, we take the natural next step and conduct cancer FMR analysis based on models that incorporate low-dimensional clinical/demographic/environmental variables, high-dimensional imaging features, as well as their interactions. Complementary to many of the existing studies, we develop a Bayesian approach for accommodating high dimensionality, screening out noises, identifying signals, and respecting the "main effects, interactions" variable selection hierarchy. An effective computational algorithm is developed, and simulation shows advantageous performance of the proposed approach. The analysis of The Cancer Genome Atlas data on lung squamous cell cancer leads to interesting findings different from the alternative approaches.

Keywords: Bayesian; Cancer; Finite mixture of regression; Histopathological imaging–environment interaction.

*To whom correspondence should be addressed.

1. INTRODUCTION

Cancer is a highly heterogeneous disease. Patients with different subtypes of the same cancer or even with the same subtype can have different biomarkers, prognosis, and response to treatment patterns. Quantifying heterogeneity can assist better understanding cancer biology and delivering tailored treatment (Burrell *and others*, 2013; Baliu-Piqué *and others*, 2020). Most cancer heterogeneity analysis can be classified as unsupervised and supervised, both of which have led to important findings and complement but cannot replace each other. In this study, we conduct supervised heterogeneity analysis, where a response variable is present, subjects form subgroups, and different subgroups have different relationships between the response and covariates. Supervised analysis, in many cases, is "closer" to clinical practice. In supervised heterogeneity analysis, early studies have mostly analyzed low-dimensional clinical/demographic/environmental variables. The development of sequencing techniques has led to quite a few heterogeneity analyses based on high-dimensional omics variables, such as gene expressions and DNA mutations (Burrell *and others*, 2013; Kim and DeBerardinis, 2019; Morrison *and others*, 2014).

In cancer modeling, an alternative source of data comes from histopathological images—a "byproduct" of biopsy, which is ordered for most patients suspected of cancer. Brief information on extracting histopathological imaging features is provided in Figure 1 and described in detail in Section 4. Compared to omics data, histopathological imaging data enjoy much broader availability and higher cost-effectiveness. It contains rich information on tumors micro properties and the surrounding microenvironment. Here, it is noted that histopathological imaging data differ significantly from radiological image data. Radiological imaging data, such as those generated by computed tomography and magnetic resonance imaging, inform tumors macro properties such as location, density, and shape. Histopathological images have been traditionally used for definitive diagnosis and staging this is realized by pathologists examining such images under microscopes. Usually, only a small number of imaging features can be analyzed in such effort. More recently, automated digital image processing pipelines/software have been developed, which can extract high-dimensional features in a fast and objective way. A series of studies have shown that such features provide an alternative and effective way for modeling cancer outcomes (Echle *and others*, 2020; Chen *and others*, 2020*a*,*b*). In a recent study (Xu *and others*, 2019), inspired by gene–environment interaction analysis, histopathological imaging–environment interaction analysis is developed and shown to have sensible biological implications and satisfactory numerical performance. Our literature review suggests that most histopathological image-based analyses, including the interaction analysis, assume homogeneity. There are only a few supervised heterogeneity analysis. However, some studies, such as Belhomme *and others* (2015) and Luo *and others* (2017), are limited to main effects only (without interactions) and, quite often, low-dimensional features. He *and others* (2020) analyzes the main effects of high-dimensional imaging features using the penalized fusion technique and resorts to model averaging to achieve computational feasibility. In this study, we will take the natural next step and conduct supervised cancer heterogeneity analysis based on models that incorporate histopathological imaging–environment interactions. This strategy has been motivated by the significance of cancer heterogeneity analysis, still strong demand for more effective methods and analysis, unique advantages of histopathological imaging data, the promising performance of histopathological imaging–environment interaction analysis under homogeneity, and lack of heterogeneity analysis incorporating such interactions.

Literature on supervised heterogeneity analysis is vast. For comprehensive discussions, we refer to, for example, Schlattmann (2009). Among the existing techniques, finite mixture of regression (FMR) (McLachlan and Peel, 2000) has been a popular choice. Under FMR, it is assumed that subjects form subgroups, and different subgroups have different regression models for the response variable. For estimation, both frequentist and Bayesian techniques have been developed (Frühwirth-Schnatter *and others*, 2018). When the number of variables is large, additional developments are needed to accommodate high dimensionality, screen out noises, and achieve unique and reliable estimation. Under the frequentist framework,

this is often achieved using regularization, in particular penalization (Khalili and Chen, 2007; Städler *and others*, 2010). There have been equally successful developments under the Bayesian paradigm. For example, Gupta and Ibrahim (2007) and Lee *and others* (2016) propose FMR models that can identify covariates relevant for each subgroup. To determine the number of subgroups, Gupta and Ibrahim (2007) formulate a model comparison problem and compare the Bayes Factors for different subgroup numbers, and Lee *and others* (2016) adopt criteria such as the Akaike Information Criterion and Bayesian Information Criterion. From a Bayesian perspective, a fully Bayesian approach that treats the number of subgroups as a random variable may be preferable. A representative example is Liu *and others* (2015), which identifies subgroup-specific covariate effects with an unknown number of subgroups. Despite great successes, the aforementioned and other methods in the literature are not directly applicable to the proposed analysis that involves interactions. In particular, interaction analysis is uniquely challenged by the "main effects, interactions" hierarchy, which postulates that if an interaction term is identified as important, then the corresponding main effect(s) should be automatically identified (Bien *and others*, 2013). In the context of imaging–environment interaction analysis, this hierarchy amounts to a constraint on the interaction term and corresponding main imaging effect (Xu *and others*, 2019). With this hierarchy, "ordinary" high-dimensional techniques are not sufficient. Under the frequentist framework, for example, composite penalization and sparse group penalization have been developed to respect the hierarchy. One explanatory development under the Bayesian framework introduces a hierarchical prior that imposes a constraint to respect the hierarchy (Kim *and others*, 2018). Here, we note that this and other Bayesian developments are limited to the homogeneity case and not directly applicable to heterogeneity analysis.

The goal of this study is to develop an effective Bayesian FMR approach for supervised cancer heterogeneity analysis based on models that incorporate histopathological imaging–environment interactions. This study complements and advances from the existing literature in multiple important aspects. First, it conducts heterogeneity analysis based on imaging data, which complements the existing studies that are based on clinical, demographic, environmental, and omics variables. It may be particularly advantageous over the omics-based studies because of broader data availability and cost-effectiveness. It also advances from the existing image-based heterogeneity analysis by incorporating interactions, and from the imaging–environment interaction analysis by accommodating sample heterogeneity. Second, it tackles considerable technical challenges. More specifically, it advances from the existing Bayesian heterogeneity analysis by respecting the "main effects, interactions" hierarchy, and from the existing Bayesian interaction analysis by accommodating heterogeneity with an unknown number of subgroups. This is achieved by incorporating priors that respect the hierarchy on the subgroup-specific parameters and adopting a mixture model with a prior on the number of subgroups. It provides a competitive alternative to the penalization and other techniques. Third, this study provides a useful alternative for extracting information from The Cancer Genome Atlas (TCGA) and other cancer data, especially for lung cancer. With these advancements, it is warranted beyond the existing literature.

2. Methods

2.1. *Model*

Consider n independent subjects. For the ith subject, let y_i denote the response variable, $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ denote the p-dimensional vector of imaging (I) features and l-dimensional vector of environmental (E) variables, respectively. As in Xu *and others* (2019) and quite a few other studies, we take a loose definition and also include demographic, clinical, and some other low-dimensional variables in $\boldsymbol{w}$. To accommodate I–E interactions, we further denote $\boldsymbol{z}_{ij} = (x_{ij}, x_{ij}w_{i1}, \ldots, x_{ij}w_{il})^T$ for $j = 1, \ldots, p$ and $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}^T, \ldots, \boldsymbol{z}_{ip}^T)^T$ for $i = 1, \ldots, n$. Let $L = l + 1$. We note that $\boldsymbol{z}_{ij}$ is a L-dimensional vector that contains the main effect and all interaction terms related to the jth imaging feature. As such, quantifying the effects of the jth imaging

feature amounts to a two-step procedure: determining whether $\boldsymbol{z}_{ij}$ has an impact on the response variable at all, and, if yes, determining which components have an impact.

We consider a continuous response, make the Gaussian distribution assumption, and use linear regression to model its associations with covariates. Assume that there are K sample subgroups. Let $\boldsymbol{\beta}^* = (\beta^*_{10}, \ldots, \beta^*_{K0}, \boldsymbol{\beta}^{*T}_1, \ldots, \boldsymbol{\beta}^{*T}_K)^T$ denote the vector that contains all the subgroup-specific parameters. Here, for the dth subgroup, β^*_{d0} is the intercept. $\boldsymbol{\beta}^*_d = (\boldsymbol{\beta}^{*T}_{d1}, \ldots, \boldsymbol{\beta}^{*T}_{dp})^T$ is the vector of regression coefficients associated with all imaging features, where $\boldsymbol{\beta}^*_{dj} = (\beta^*_{dj1}, \ldots, \beta^*_{djL})^T$ represents the main effect and interactions of the jth imaging features.

To facilitate estimation, we introduce a latent subgroup membership for each subject. Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$ denote the vector of latent subgroup memberships for the n subjects. Its components take values in $\{1, \ldots, K\}$. Let $\boldsymbol{p} = (p_1, \ldots, p_K)^T$ denote the vector of the unknown subgroup proportions. The proposed model is:

$$\begin{gathered} y_i|\delta_i = d, \boldsymbol{\beta}^*, \boldsymbol{\eta}, \sigma^2 \sim \mathrm{N}\left(\beta^*_{d0} + \boldsymbol{z}_i^T \boldsymbol{\beta}^*_d + \boldsymbol{w}_i^T \boldsymbol{\eta},\ \sigma^2\right),\ i = 1, \ldots, n;\ d = 1, \ldots, K, \\ \delta_i \overset{\text{iid}}{\sim} \text{Multinomial}(1; p_1, \ldots, p_K),\ i = 1, \ldots, n. \end{gathered} \tag{2.1}$$

In Bayesian mixture modeling, priors are assigned to the latent subgroup memberships and subgroup-specific parameters. Further, we take a fully Bayesian approach, assign a prior on K, and flexibly drop the assumption of a known K. Such a mixture model has been referred to as a Mixture of Finite Mixture. Prior specifications for the above model are described in the next subsection. Here, we note that certain individual components of the model and prior specifications have roots in the existing literature. However, their combination to address supervised heterogeneity analysis built on high-dimensional interaction models is new and innovative.

2.2. *Prior specifications*

We first define the prior for the subgroup-specific parameter $\boldsymbol{\beta}^*$. Recall that $\boldsymbol{\beta}^*$ contains the subgroup-specific intercepts and regression coefficients associated with the imaging features. For the subgroup-specific intercepts $\beta^*_{d0}, d = 1, \ldots, K$, we assume the $N(\mu_0, \xi^2)$ prior. For the coefficients associated with the imaging features, we introduce sparsity to accommodate high dimensionality and distinguish between signals and noises. High-dimensional imaging features extracted using digital processing software also describe properties not related to cancer, making it necessary to conduct variable selection. This is also true for I–E interaction analysis (Xu *and others*, 2019). In our analysis, further complication is introduced by heterogeneity. Specifically, different subject subgroups may have different subsets of important variables associated with the response. To address these challenges, we impose spike and slab priors, which have been popular in the Bayesian variable selection literature, to the subgroup-specific parameters. Specifically, to allow for both imaging feature-level and within-imaging-feature-level sparsity, we first follow Xu and Ghosh (2015) and take a reparameterization:

$$\boldsymbol{\beta}^*_{dj} = V^{*\frac{1}{2}}_{dj} \boldsymbol{b}^*_{dj}, j = 1, \ldots, p; d = 1, \ldots, K,$$

where $\boldsymbol{b}^*_{dj} = (b^*_{dj1}, \ldots, b^*_{djL})^T$ and $V^{*\frac{1}{2}}_j = \text{diag}(\tau^*_{dj1}, \ldots, \tau^*_{djL}), \tau^*_{djl} \geq 0$ for $l = 1, \ldots, L$.

To achieve feature-level sparsity, we impose the multivariate spike and slab prior on each $\boldsymbol{b}^*_{dj}$:

$$\boldsymbol{b}^*_{dj} \overset{\text{ind}}{\sim} (1 - \pi_0)\mathrm{N}_L(\boldsymbol{0}, I_L) + \pi_0 \delta_0(\boldsymbol{b}^*_{dj}),\ j = 1, \ldots, p, \tag{2.2}$$

where I_L is the identity matrix, and δ_0 denotes the point mass at 0. This spike and slab prior, which is a mixture of a Normal distribution (slab part) and a point mass at $\mathbf{0}$ (spike part) with weights $1-\pi_0$ and π_0, respectively, assigns a nonzero probability to the coefficients being exactly zero. If $\boldsymbol{b}^*_{dj} = \mathbf{0}$, then all elements of $\boldsymbol{\beta}^*_{dj}$ are zero, suggesting that the jth imaging feature has no main effect or any interaction effect on the response in the dth subgroup.

To induce within-feature-level sparsity, we further impose the spike and slab prior on each τ^*_{djl} except for τ^*_{dj1}:

$$\begin{aligned} \tau^*_{dj1} &\overset{\text{ind}}{\sim} \mathrm{N}^+(0, s^2),\ j = 1, \ldots, p, \\ \tau^*_{djl} &\overset{\text{ind}}{\sim} (1-\pi_1)\mathrm{N}^+(\tau_m, s^2) + \pi_1 \delta_0(\tau^*_{djl}),\ j = 1, \ldots, p,\ l = 2, \ldots, L, \end{aligned} \tag{2.3}$$

where $\mathrm{N}^+(m, s^2)$ denotes the truncated normal distribution whose probability density function is proportional to that of $N(m, s^2)$ but truncated to be positive and then normalized. In our modeling, each τ^*_{djl} determines the magnitude of a regression coefficient for an effect associated with the jth imaging feature. When $\boldsymbol{\beta}^*_{dj} \neq \mathbf{0}$, the above prior specification ensures that $\tau^*_{dj1} \neq 0$. That is, the main effect is selected. For $l = 2, \ldots, L$, the spike and slab priors determine whether the individual I–E interactions have nonzero effects. It is noted that if at least one interaction is nonzero, then $\boldsymbol{\beta}^*_{dj} \neq \mathbf{0}$, leading to the nonzero main effect. This ensures that the "main effects, interactions" hierarchy is respected.

The feature-level sparsity and within-feature-level sparsity are controlled by the prior inclusion probabilities, π_0 and π_1 in (2.2) and (2.3), respectively. Fixing the values of these hyperparameters leads to more informative priors and poorer multiplicity control for a higher number of spurious covariates (Scott and Berger, 2010). This multiplicity problem can be handled by fully Bayesian models that assign hyperpriors to π_0 and π_1. Extending the idea of Ley and Steel (2009), we adopt conjugate beta hyperpriors $\pi_0 \sim \mathrm{Beta}(a_{\pi_0}, b_{\pi_0})$ and $\pi_1 \sim \mathrm{Beta}(a_{\pi_1}, b_{\pi_1})$. As for the choice of hyperparameters τ_m and s^2 (which determine the prior of the magnitude of nonzero coefficients τ^*'s), we resort to the theoretical conditions derived in Narisetty and He (2014, Section 2.1) for standard Bayesian variable selection models with spike and slab priors for a rough guideline. Briefly, the larger the number of covariates is relative to the size of each subgroup, and the more severe the multicollinearity of the design matrix, the larger the values of τ^*'s need to be to achieve variable selection consistency. Since the theoretical result bears no direct implication for the choice of these values in practice, the $\mathrm{N}^+(\tau_m, s^2)$ prior is assigned to give a significant probability to large values of τ^*'s, allowing their posterior values to be informed by the observed data. For most setups with moderate correlations among covariates, setting $\tau_m = 0$ and a vague prior on s^2 suffices. For setups with a large p (relative to n) and highly correlated covariates, setting the mode of the slab part of the prior, τ_m, away from the spike at zero helps identify more useful covariates. Following Xu and Ghosh (2015), a conjugate prior is assigned on s^2: $s^2 \sim$ Inverse Gamma$(1, \lambda)$. Instead of estimating λ using empirical Bayesian methods, we take the full Bayesian approach and assign λ a prior, $\lambda \sim \mathrm{Gamma}(a_\lambda, b_\lambda)$, where a_λ and b_λ are the shape and scale parameters, respectively. More discussions on prior choice and consistency are in Supplementary material available at *Biostatistics* online.

Given K, we assign the prior $\boldsymbol{p} \sim \mathrm{Dirichlet}_K(\alpha, \ldots, \alpha)$ to the subgroup proportions, where α is a constant independent of K. This is the prior adopted in Miller and Harrison (2018). On one hand, imposing the same precision α across different values of K is a restriction in terms of modeling. But this can be critical to efficient computing in Section 2.3, as K and $\boldsymbol{p}$ can then be marginalized out to avoid the transdimensional computing problems caused by the varying dimension of group-specific parameters for different values of K. On the other hand, given K, it is not as big a restriction to set all K parameters in the Dirichlet distribution to be α. This is because the prior on $(\boldsymbol{\beta}^{*T}_1, \ldots, \boldsymbol{\beta}^{*T}_K)^T$ is symmetric, and hence the distribution of $(\boldsymbol{\beta}^T_1, \ldots, \boldsymbol{\beta}^T_n)$ under any asymmetric distribution on $\boldsymbol{p}$ will be the same as if this asymmetric distribution is replaced by a symmetric version of it where the entries of $\boldsymbol{p}$ are randomly permuted. For the number

of subgroups K, we assume a Geometric distribution $K \sim \text{Geo(q)}$ where $p(K = k) = (1-q)^{(k-1)}q$, for $k = 1, 2, \ldots$, with some $q \in (0, 1)$.

For the remaining parameters, we assign $\boldsymbol{\eta} \sim \text{N}_l(\boldsymbol{\mu}_\eta, \sigma^2\Sigma_\eta)$ and $\sigma^2 \sim \text{Inverse Gamma}(a_0, b_0)$, where a_0 and b_0 are the shape and scale parameters, respectively. In practice, E variables are usually low-dimensional, manually selected, and important, and hence will not be subject to variable selection. In the above formulation, it is assumed that $\boldsymbol{\eta}$ is the same across subgroups. In our data analysis, all the samples have the same cancer type. That is, based on the demographic, clinical, and environmental variables, these samples have been concluded as sufficiently alike. As such, the goal is to see if more subtle data structures can be identified with the introduction of imaging features and their interactions with E variables. If needed, it is straightforward to design subgroup-specific $\boldsymbol{\eta}$.

Although this study emphasizes methodological development and applications, to provide a strong statistical basis, in Supplementary material available at *Biostatistics* online, we provide heuristic justifications on identifiability and consistency of the proposed model. Accordingly, we have incorporated such considerations when specifying priors in our numerical studies.

2.3. *Computation*

We develop Markov chain Monte Carlo (MCMC) algorithms to estimate the posterior distribution. Recall that $(\boldsymbol{\delta}^T, \boldsymbol{\beta}^{*T}, \boldsymbol{\eta}^T, \pi_0, \pi_1, s^2, \sigma^2, \lambda)^T$ is the vector of the latent subgroup memberships and parameters for the proposed model. Here, the vector of the latent subgroup memberships $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^T$ induces a partition $\mathcal{C}$ of $\{1, \ldots, n\}$. The goal is to develop a MCMC algorithm to explore the joint distribution over the space of $\mathcal{C}$ and the space of the other parameters. The joint distribution to be sampled from can be summarized as:

$$\begin{aligned} &p(\boldsymbol{\delta}, \boldsymbol{\beta}^*, \boldsymbol{\eta}, \pi_0, \pi_1, s^2, \sigma^2, \lambda | \boldsymbol{y}) \\ &\quad \propto f(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{\beta}^*, \boldsymbol{\eta}, \sigma^2) p(\boldsymbol{\beta}^*|\boldsymbol{\delta}, \pi_0, \pi_1, s^2) p(\boldsymbol{\delta}) p(\boldsymbol{\eta}|\sigma^2) p(\pi_0) p(\pi_1) p(s^2|\lambda) p(\lambda) p(\sigma^2). \end{aligned}$$

We develop a Metropolis-within-Gibbs sampler that updates the vector of subgroup memberships and subgroup-specific parameters, followed by the parameters that are not subgroup-specific.

To update the subgroup memberships, we adopt the strategy of introducing auxiliary variables (Miller and Harrison, 2018). Additional discussions are provided in Supplementary material available at *Biostatistics* online. Let $\boldsymbol{\delta}_{-i}$ denote $\boldsymbol{\delta}$ with its ith component removed, and t denote the number of unique values in $\boldsymbol{\delta}_{-i}$. Let $\boldsymbol{\beta}_a^* = (\boldsymbol{\beta}_{t+1}^{*T}, \ldots, \boldsymbol{\beta}_{t+m}^{*T})^T$ denote a set of m auxiliary variables that are identically and independently distributed from the prior specified in the previous subsection. Following Miller and Harrison (2018), the prior for $\boldsymbol{\delta}$ in (2.1) implies:

$$p(\delta_i = d | \boldsymbol{\delta}_{-i}, \boldsymbol{\beta}^*, \boldsymbol{\beta}_a^*) \propto \begin{cases} n_{d,-i} + \alpha & \text{if } d = \delta_k \text{ for some } k \neq i, \\ \frac{V_n(t+1)}{V_n(t)} \frac{\alpha}{m} & \text{if } d \neq \delta_k \text{ for all } k \neq i, \end{cases}$$

where $n_{d,-i} = |\{j \in \boldsymbol{\delta}_{-i} : \delta_j = d\}|$ denotes the size of the dth subgroup without the ith sample, $V_n(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\alpha k)^{(n)}} p_K(k)$, $a^{(b)} = a(a+1)\cdots(a+b-1)$, and $a_{(b)} = a(a-1)\cdots(a-b+1)$, with $a^{(0)} = a_{(0)} = 1$ by convention. Then the full conditional distribution for δ_i (conditional on the rest of the parameters) is:

$$\begin{aligned} p(\delta_i = d | \text{rest}) &\propto p(\delta_i = d | \boldsymbol{\delta}_{-i}, \boldsymbol{\beta}^*, \boldsymbol{\beta}_a^*) f(y_i | \boldsymbol{\beta}_d^*, \boldsymbol{\eta}, \sigma^2) \\ &\propto \begin{cases} (n_{d,-i} + \alpha) f(y_i | \boldsymbol{\beta}_d^*, \boldsymbol{\eta}, \sigma^2) & \text{if } d = \delta_k \text{ for some } k \neq i, \\ \frac{V_n(t+1)}{V_n(t)} \frac{\alpha}{m} f(y_i | \boldsymbol{\beta}_d^*, \boldsymbol{\eta}, \sigma^2) & \text{if } d \neq \delta_k \text{ for all } k \neq i. \end{cases} \end{aligned}$$

The subgroup-specific parameters are updated separately for each subgroup. To update the coefficients associated with the jth ($j = 1, \ldots, p$) imaging feature for the dth ($d = 1, \ldots, K$) subgroup, we first conduct the feature-level update, followed by the within-feature-level update. For the feature-level update, let $\boldsymbol{\beta}^*_{d(j)}$ denote $\boldsymbol{\beta}^*_d$ with its jth component removed, and $\boldsymbol{z}_{i(j)}$ denote $\boldsymbol{z}_i$ with its jth component removed. Conditional on the rest of the parameters, $\boldsymbol{b}^*_{dj}$ has a multivariate spike and slab distribution:

$$\boldsymbol{b}^*_{dj}|\text{rest} \sim (1 - g^b_j)N_L(\mu^b_{dj}, \sigma^2\Omega^b_{dj}) + g^b_j\delta_0(\boldsymbol{b}^*_{dj}),$$

where $\Omega^b_{dj} = \Big\{V^{*\frac{1}{2}}_{dj}\big(\sum_{i;\delta_i=d}\boldsymbol{z}_{ij}\boldsymbol{z}^T_{ij}\big)V^{*\frac{1}{2}}_{dj} + \sigma^2 I_L\Big\}^{-1}$ and $\mu^b_{dj} = \Omega^b_{dj}V^{*\frac{1}{2}}_{dj}\Big\{\sum_{i;\delta_i=d}\boldsymbol{z}_{ij}(y_i - \boldsymbol{w}^T_i\boldsymbol{\eta} - \beta^*_{d0} - \boldsymbol{z}^T_{i(j)}\boldsymbol{\beta}^*_{d(j)})\Big\}$. In the above, g^b_j represents the full conditional posterior probability of $\boldsymbol{b}^*_{dj}$ being equal to 0, which can be shown as:

$$g^b_j = p(\boldsymbol{b}^*_{dj} = 0|\text{rest}) = \frac{\pi_0}{\pi_0 + (1-\pi_0)|\sigma^2\Omega^b_{dj}|^{1/2}\exp(\frac{1}{2\sigma^2}\mu^{bT}_{dj}\Omega^b_{dj}\mu^b_{dj})}.$$

For the within-feature-level update, let $\boldsymbol{\beta}^*_{d(jl)}$ denote $\boldsymbol{\beta}^*_d$ with the lth element of the jth imaging variable group removed, and $\boldsymbol{z}_{i(jl)}$ denote $\boldsymbol{z}_i$ with the lth element of the jth imaging variable group removed. The conditional distribution of τ^*_{djl} for $l = 2, \ldots, L$ is a spike and slab distribution, while that of τ^*_{dj1} is just the slab part of the mixture distribution. That is,

$$\begin{aligned}
&\tau^*_{dj1}|\text{rest} \sim \text{N}^+(u_{dj1}, v^2_{dj1}),\\
&\tau^*_{djl}|\text{rest} \sim (1 - g^\tau_{djl})\text{N}^+(u_{djl}, v^2_{djl}) + g^\tau_{djl}\delta_0(\tau^*_{djl}),\ \text{for}\ l = 2, \ldots, L,
\end{aligned}$$

where for $l = 1, \ldots, L$, $v^2_{djl} = \big(\frac{1}{\sigma^2}b^{*2}_{djl}\sum_{i;\delta_i=d}z^2_{ijl} + \frac{1}{s^2}\big)^{-1}$, $u_{dj1} = v^2_{dj1}\Big\{\frac{1}{\sigma^2}b^*_{dj1}\sum_{i;\delta_i=d}z_{ij1}(y_i - \boldsymbol{w}^T_i\boldsymbol{\eta} - \beta^*_{d0} - \boldsymbol{z}^T_{i(j1)}\boldsymbol{\beta}^*_{d(j1)})\Big\}$, and for $l = 2, \ldots, L$, $u_{djl} = v^2_{djl}\Big\{\frac{1}{\sigma^2}b^*_{djl}\sum_{i;\delta_i=d}z_{ijl}(y_i - \boldsymbol{w}^T_i\boldsymbol{\eta} - \beta^*_{d0} - \boldsymbol{z}^T_{i(jl)}\boldsymbol{\beta}^*_{d(jl)}) + \frac{\tau_m}{s^2}\Big\}$. The weight g^τ_{djl} is the full conditional posterior probability of τ^*_{djl} being equal to zero, that is,

$$g^\tau_{djl} = p(\tau^*_{djl} = 0|\text{rest}) = \frac{\pi_1}{\pi_1 + (1-\pi_1)\Phi(\frac{\tau_m}{s})^{-1}(s^2)^{-\frac{1}{2}}(v^2_{djl})^{\frac{1}{2}}\exp(\frac{u^2_{djl}}{2v^2_{djl}})\Phi(\frac{u_{djl}}{v_{djl}})},$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

The full conditional distribution of the subgroup-specific intercept β^*_{d0} is given by:

$$\beta^*_{d0}|\text{rest} \sim N\left(\frac{\xi^2\sum_{i;\delta_i=d}(y_i - \boldsymbol{z}^T_i\boldsymbol{\beta}^*_d - \boldsymbol{w}^T_i\boldsymbol{\eta}) + \sigma^2\mu_0}{\xi^2 n_d + \sigma^2}, \frac{\sigma^2\xi^2}{\xi^2 n_d + \sigma^2}\right),$$

where $n_d = |\{j \in \boldsymbol{\delta} : \delta_j = d\}|$ denotes the size of the dth subgroup.

We next update the parameters that are not subgroup-specific. The full conditional distribution of $\boldsymbol{\eta}$ is:

$$\boldsymbol{\eta}|\text{rest} \sim \text{N}_l(\boldsymbol{m}_\eta, \sigma^2\Omega_\eta),$$

where $\Omega_\eta = \Big(\sum_{d\in\boldsymbol{\delta}}\sum_{i:\delta_i=d}\boldsymbol{w}_i\boldsymbol{w}^T_i + \Sigma^{-1}_\eta\Big)^{-1}$ and $\boldsymbol{m}_\eta = \Omega_\eta\Big\{\sum_{d\in\boldsymbol{\delta}}\sum_{i:\delta_i=d}\boldsymbol{w}_i(y_i - \beta^*_{d0} - \boldsymbol{z}^T_i\boldsymbol{\beta}^*_d) + \Sigma^{-1}_\eta\mu_\eta\Big\}$.

The posteriors of π_0 and π_1 conditional on all the other parameters are:

$$\pi_0|\text{rest} \sim \text{Beta}\left(\sum_{d\in\boldsymbol{\delta}}\sum_{j=1}^{p}\mathbb{1}(\boldsymbol{b}^*_{dj}=0)+a_{\pi_0}, \sum_{d\in\boldsymbol{\delta}}\sum_{j=1}^{p}\mathbb{1}(\boldsymbol{b}^*_{dj}\neq 0)+b_{\pi_0}\right),$$

$$\pi_1|\text{rest} \sim \text{Beta}\left(\sum_{d\in\boldsymbol{\delta}}\sum_{j=1}^{p}\sum_{l=2}^{L}\mathbb{1}(\tau^*_{djl}=0)+a_{\pi_1}, \sum_{d\in\boldsymbol{\delta}}\sum_{j=1}^{p}\sum_{l=2}^{L}\mathbb{1}(\tau^*_{djl}>0)+b_{\pi_1}\right).$$

The conditional distribution of s^2 is given by

$$p(s^2|\text{rest}) \propto \prod_{d\in\boldsymbol{\delta}}\left[\prod_{j=1}^{p}N^+(\tau^*_{dj1}|0,s^2)\prod_{l=2}^{L}\left\{(1-\pi_1)N^+(\tau^*_{djl}|\tau_m,s^2)+\pi_1\right\}\right]p(s^2|\lambda).$$

Since this is not a standard distribution, we resort to the random-walk Metropolis update with a Gaussian proposal distribution.

Lastly, the posterior distributions of λ and σ^2 are Gamma and Inverse Gamma, respectively:

$$\lambda|\text{rest} \sim \text{Gamma}(a_\lambda+1, \frac{b_\lambda s^2}{b_\lambda+s^2}), \quad \sigma^2|\text{rest} \sim \text{InverseGamma}(a_1, b_1),$$

where $a_1 = \frac{n+l}{2}+a_0$ and $b_1 = \frac{1}{2}\Big\{\sum_{d\in\boldsymbol{\delta}}\sum_{i;\delta_i=d}(y_i-\boldsymbol{w}_i^T\boldsymbol{\eta}-\beta^*_{d0}-\boldsymbol{z}_i^T\boldsymbol{\beta}^*_{dj})^2+(\boldsymbol{\eta}-\boldsymbol{\mu}_\eta)\Sigma_\eta^{-1}(\boldsymbol{\eta}-\boldsymbol{\mu}_\eta)\Big\}+b_0$.

2.4. *Inference based on MCMC samples*

It is well known that mixture models are not identifiable, due to the label-switching problem caused by the symmetry of parameters in the likelihood. Some inference goals are label-invariant and not affected by this problem. An example is the (marginal) posterior distribution of the number of subgroups, where its mode $\hat{K}$ is the commonly used point estimate of K.

Some other inference, like that for the group-specific estimate of the coefficient of a covariate, cannot be directly obtained from its marginal posterior, as every subgroup has exactly the same marginal posterior due to symmetry. This is why for Bayesian mixture models computed with MCMC, there is a rich literature on the postprocessing procedures. See Supplementary material available at *Biostatistics* online for more discussions on the nonidentifiability issue and, for example, Papastamoulis (2016) for a list of MCMC postprocessing algorithms. Our data analysis adopts Algorithm 5 of the aforementioned paper for postprocessing, which is proposed in Papastamoulis and Iliopoulos (2010) and Rodríguez and Walker (2014) and implemented using the R package `label.switching`.

Sometimes, it is of interest to obtain a point estimate of the subgrouping configuration. One solution is to assign each subject to the subgroup that it belongs to with the highest posterior probability. Given the relabeled MCMC samples, this is simply estimated by the subgroup that the subject belongs to the most often. For other alternatives that are based on decision theory, see for example, Wade and Ghahramani (2018).

Recall that one prominent feature of our model is that different variables may be selected for different subgroups. Within a subgroup, we can simply follow the standard approaches in the literature for Bayesian regression models with spike and slab priors. Specifically, in simulation, we consider the median probability model (MPM) that retains all covariates with marginal posterior inclusion probabilities (PIP) greater than 0.5. The MPM is known to have optimal prediction performance in certain setups, for example, when

the design matrix is orthogonal (Barbieri and Berger, 2004). However, it tends to select too few covariates in practice (Dey *and others*, 2008). One remedy is to choose a lower threshold, which is recommended in Narisetty and He (2014, Section. 2.2) and adopted in our real data analysis.

Finally, for inference of the group-specific coefficients in $\boldsymbol{\beta}_d^*$ from the dth subgroup, key quantities that measure the importance of the coefficients are their chances of being nonzero, which are the PIPs. In addition, one can inspect the marginal posterior densities of the coefficients and report summary statistics such as the posterior medians.

3. Simulation

We gauge the performance of the proposed approach and benchmark against alternatives using simulation. In what follows, we set the sample size $n = 200$ and number of subgroups $K = 2$. Response variables are independently generated from model (2.1) with $\sigma^2 = 1$. For the E variables, we set the number of variables $l = 5$ and consider both continuous and discrete types. More specifically, for the continuous E variables, we generate $\boldsymbol{w}_i$'s from a multivariate normal distribution with mean vector zero and covariance matrix that has an auto-regressive correlation structure with $\rho_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. For the discrete E variables, we first generate continuously distributed variables in the same way as above and then dichotomize at 0. The coefficients for all of the main E effects are generated from Uniform(0.8, 1.2). For the I variables, we generate $\boldsymbol{x}_i$'s from a multivariate normal distribution of dimension p with marginal means one and two different covariance structures. The first is the block-diagonal structure, reflecting that correlations are "local", where each block of size 5×5 has an auto-regressive structure with $\rho_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. The second is the banded correlation structure with $\rho_{ij} = \mathbb{1}_{(i=j)} + 0.33\mathbb{1}_{(|i-j|=1)}$. For the subgroup-specific regression coefficients associated with the I variables, $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$, in each subgroup, five of the main I effects and ten of the I–E interactions are set to be nonzero, and the rest are zero. The "main effects, interactions" hierarchy is satisfied. For each combination of the specifications on E variables (continuous, discrete) and covariance matrix of I features (block-diagonal, banded), we further consider the following four scenarios. [Scenario 1] The dimension of imaging features $p = 100$. The nonzero components of $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are independently generated from Uniform(-1.2, -0.8) and Uniform(0.8, 1.2), respectively. The subgroup memberships are generated from Multinomial$(1; 0.5, 0.5)$. [Scenario 2] Similar to the above, except that the nonzero components are generated from Uniform(-0.8, -0.5) and Uniform(0.5, 0.8), respectively, representing weaker signals. [Scenario 3] Similar to Scenario 1, except that the subgroup memberships are generated from Multinomial$(1; 0.3, 0.7)$. That is, the subgroups are imbalanced. [Scenario 4] Similar to Scenario 1, except that the I variables have a higher dimension with $p = 200$. Here, we note that the dimensions of covariates, especially p, have been chosen to be comparable to the data analyzed in the next section, and that, in principle, the proposed approach can be applied to settings with higher dimensions.

To implement the proposed method, we adopt $K \sim \text{Geo}(\text{q} = 0.1)$, $\alpha = 1$, and $(a_0, b_0, a_{\pi_0}, b_{\pi_0}, a_{\pi_1}, b_{\pi_1}, \tau_m, a_\lambda, b_\lambda, \mu_\eta, \Sigma_\eta) = (1, 1, 1, 1, 1, 1, 0, 1, 1, \mathbf{0}_5, 10I_5)$ for the prior distributions. For the number of auxiliary variables m in updating the subject subgroup memberships, we set $m = 10$. Computation is carried out by running the proposed sampler for 20 000 iterations, with the first half discarded as burn-in, and all inferences are based on the remaining MCMC samples. To initialize each Markov chain, all samples are assigned to the same subgroup, and the values of all the other parameters are randomly drawn from their priors. With multiple MCMC runs, we inspect the trace plots, compare across runs, and do not observe obvious signs of lack of convergence (sample plots are provided in Figures S.4 and S.5 of Supplementary material available at *Biostatistics* online). Further, for the label-invariant variables, Gelman and Rubins potential scale reduction factor (PSRF; Gelman and Rubin, 1992) is used for assessing convergence. The PSRF values are all below 1.1, indicating satisfactory convergence.

With the high dimensionality, complex data structure especially with the "main effects, interactions" hierarchy, and subgrouping structure, computation of the proposed method is inevitably more expensive than that in some existing studies. For one simulated dataset under Scenario 1, computation takes about 1.5 h on a desktop with standard configurations.

For comparison, we consider the following alternatives: (i) The Bayesian Sparse Group Selection with Spike and Slab Prior (BSGSS), which is developed in Xu and Ghosh (2015) and assumes homogeneity. With this benchmark approach, we can "re-establish" the importance of accounting for heterogeneity. (ii) The FMR Lasso (denoted as FMRLasso), which is developed in Städler *and others* (2010). This is one of the most popular heterogeneity analysis approaches for high-dimensional data. It assumes the FMR model and adopts Lasso for variable selection. It treats main effects and interactions in the same manner and may violate the variable selection hierarchy. Comparing with this approach can provide a benchmark for the proposed Bayesian estimation and also "re-establish" the importance of respecting the hierarchy. (iii) The FMR based on the Imputation-conditional consistency algorithm (denoted as ICC), which is developed in Li *and others* (2019). Under the FMR modeling, this approach conducts estimation using ICC, which is a general technique for handling missing data in high-dimensional settings. Conditioning on the assigned subgroup membership, the minimax concave penalty is applied for accommodating high dimensionality and conducting variable selection. It is noted that some alternative penalties (e.g., smoothly clipped absolute deviation) are expected to lead to similar performance. For the finite mixture models, it remains challenging to determine the number of subject subgroups. Here, we set their number of subgroups at the true $K = 2$ and note that this may generate favorable performance for the two FMR alternatives and is not feasible in real data analysis. The FMRLasso and ICC generate point estimates of the subgrouping configuration and subgroup-specific parameters $\boldsymbol{\beta}^*$. We acknowledge that there are other potentially applicable alternatives. The above three may be the most relevant and can be readily realized.

To gain more insight into the working characteristics of the proposed approach, in Figures S.1–S.3 of the Supplementary material available at *Biostatistics* online, for one representative simulation replicate, we present the true model parameters as well as estimation and variable selection performance of the proposed approach. Satisfactory performance is clearly observed. Specifically, the inclusion probabilities for the zero coefficients are very small, in particular, smaller than those for the nonzero coefficients, which leads to accurate variable selection. It is also observed that the colors of the estimates are close to those of the true values. Then, based on 100 simulation replicates, to more objectively evaluate the accuracy of estimating K, we report the mean and standard deviation (sd) of $\hat{K}$ for the proposed method. To evaluate the accuracy of subgrouping, we report the mean (sd) of the Adjusted Rand Index (ARI; Hubert and Arabie, 1985) for the proposed method, FMRLasso, and ICC. ARI yields a maximal value of 1 if the estimated and true subgroupings perfectly match, and can be negative if the two subgroupings are "less similar" than what is expected under random assignments. A higher ARI value indicates higher subgrouping similarity. In our evaluation, we compare the estimated subgrouping against the true. The proposed and alternative approaches all conduct variable selection. We evaluate variable selection accuracy using the true positive (TP) and false positive (FP) rates. Summary statistics are provided in Table 1 and Table S.1 of the Supplementary material available at *Biostatistics* online.

Across the whole spectrum of simulation, the proposed approach is observed to have competitive performance. For all simulation settings, it can almost perfectly identify K, which is often challenging in FMR and other heterogeneity analysis. It has superior subgrouping performance. Consider for example setup 1 and Scenario 1 (the upper-left corner of Table 1). The proposed approach has a mean Accuracy of 0.820, compared to 0.685 (FMRLasso) and 0.637 (ICC). It also has satisfactory variable selection performance. For the above particular setting, it is able to identify all TPs with almost no FPs for both subject subgroups. In comparison, FMRLasso has TP rates of 0.906 and 0.908 and FP rates of 0.021 and 0.011. Its lack of accuracy is at least partially attributable to the potential violation of the variable selection hierarchy. ICC has inferior performance with TP rates of 0.524 and 0.482. BSGSS has a TP rate of 0.926,

Table 1. *Simulation results for Scenarios 1 and 2, mean (sd) based on 100 replicates*

	Subgroup	Scenario 1				Scenario 2			
		Proposed	FMRLasso	ICC	BSGSS	Proposed	FMRLasso	ICC	BSGSS
Continuous E variables, a block-diagonal covariance matrix for I features									
$\hat{K}$		2.000 (0.00)				2.000 (0.00)			
ARI		0.820 (0.05)	0.685 (0.08)	0.637 (0.17)		0.722 (0.06)	0.591 (0.08)	0.561 (0.13)	
Main I effects									
TPR	1	1.000 (0.00)	0.906 (0.19)	0.524 (0.33)	0.926 (0.11)	1.000 (0.00)	0.768 (0.25)	0.422 (0.28)	0.908 (0.14)
	2	1.000 (0.00)	0.908 (0.16)	0.482 (0.33)		1.000 (0.00)	0.760 (0.23)	0.292 (0.27)	
FPR	1	0.000 (0.00)	0.021 (0.02)	0.001 (0.00)	0.780 (0.06)	0.002 (0.01)	0.016 (0.02)	0.003 (0.01)	0.679 (0.07)
	2	0.001 (0.00)	0.011 (0.02)	0.001 (0.00)		0.003 (0.01)	0.009 (0.01)	0.001 (0.00)	
I–E interactions									
TPR	1	1.000 (0.00)	0.924 (0.16)	0.691 (0.23)	0.926 (0.11)	0.998 (0.01)	0.823 (0.14)	0.543 (0.18)	0.908 (0.14)
	2	1.000 (0.00)	0.936 (0.11)	0.714 (0.23)		0.999 (0.01)	0.873 (0.14)	0.539 (0.18)	
FPR	1	0.006 (0.01)	0.053 (0.02)	0.020 (0.01)	0.784 (0.06)	0.022 (0.01)	0.052 (0.01)	0.025 (0.01)	0.704 (0.07)
	2	0.006 (0.01)	0.076 (0.02)	0.020 (0.01)		0.022 (0.01)	0.079 (0.01)	0.024 (0.01)	
Discrete E variables, a block-diagonal covariance matrix for I features									
$\hat{K}$		2.000 (0.00)				2.000 (0.00)			
ARI		0.816 (0.06)	0.700 (0.10)	0.638 (0.25)		0.721 (0.06)	0.618 (0.07)	0.582 (0.16)	
Main I effects									
TPR	1	1.000 (0.00)	0.932 (0.15)	0.636 (0.36)	0.934 (0.011)	0.998 (0.02)	0.802(0.21)	0.496 (0.29)	0.912 (0.13)
	2	1.000 (0.00)	0.886 (0.18)	0.548 (0.34)		1.000 (0.00)	0.736(0.20)	0.372 (0.27)	
FPR	1	0.001 (0.00)	0.020 (0.03)	0.003 (0.01)	0.811 (0.06)	0.003 (0.00)	0.017(0.02)	0.001 (0.01)	0.716 (0.08)
	2	0.001 (0.00)	0.009 (0.01)	0.002 (0.00)		0.002 (0.00)	0.010(0.02)	0.003 (0.06)	
I–E interactions									
TPR	1	0.999 (0.01)	0.948 (0.12)	0.701 (0.32)	0.934 (0.11)	0.997 (0.02)	0.855 (0.13)	0.566 (0.20)	0.912 (0.13)
	2	1.000 (0.00)	0.948 (0.12)	0.720 (0.30)		1.000 (0.00)	0.892 (0.12)	0.586 (0.20)	
FPR	1	0.027 (0.01)	0.039 (0.02)	0.019 (0.01)	0.815(0.06)	0.031 (0.00)	0.038 (0.01)	0.022 (0.01)	0.722 (0.08)
	2	0.026 (0.01)	0.063 (0.02)	0.020 (0.01)		0.031 (0.00)	0.065 (0.01)	0.022 (0.01)	
Continuous E variables, a banded covariance matrix for I features									
$\hat{K}$		2.000 (0.00)				2.000 (0.00)			
ARI		0.816 (0.05)	0.688 (0.08)	0.657 (0.15)		0.730 (0.06)	0.586 (0.11)	0.557(0.17)	
Main I effects									
TPR	1	1.000 (0.00)	0.932 (0.17)	0.510 (0.29)	0.932 (0.11)	0.996 (0.04)	0.742 (0.25)	0.398 (0.26)	0.864 (0.15)
	2	1.000 (0.00)	0.890 (0.15)	0.490 (0.30)		1.000 (0.00)	0.722 (0.25)	0.316 (0.25)	
FPR	1	0.001 (0.00)	0.023 (0.03)	0.001 (0.01)	0.761 (0.07)	0.002 (0.01)	0.025 (0.03)	0.003 (0.01)	0.669 (0.073)
	2	0.001 (0.00)	0.011 (0.02)	0.001 (0.00)		0.002 (0.00)	0.013 (0.02)	0.002 (0.01)	
I–E interactions									
TPR	1	1.000 (0.00)	0.929 (0.12)	0.688 (0.22)	0.932 (0.11)	0.985 (0.05)	0.797 (0.19)	0.528 (0.19)	0.864 (0.05)
	2	0.999 (0.01)	0.941 (0.09)	0.708 (0.22)		0.990 (0.04)	0.848 (0.17)	0.543 (0.22)	
FPR	1	0.006 (0.01)	0.054 (0.01)	0.021 (0.01)	0.766 (0.06)	0.019 (0.01)	0.056 (0.02)	0.024 (0.01)	0.675 (0.07)
	2	0.007 (0.01)	0.076 (0.01)	0.019 (0.01)		0.020 (0.01)	0.081 (0.01)	0.024 (0.01)	
Discrete E variables, a banded covariance matrix for I features									
$\hat{K}$		2.000 (0.00)				2.000 (0.00)			
ARI		0.811 (0.06)	0.699 (0.07)	0.661 (0.20)		0.714 (0.06)	0.610 (0.06)	0.563 (0.18)	
Main I effects									
TPR	1	1.000 (0.00)	0.948 (0.13)	0.674 (0.31)	0.932 (0.11)	1.000 (0.00)	0.832 (0.19)	0.496(0.29)	0.876(0.14)
	2	1.000 (0.00)	0.898 (0.15)	0.546 (0.30)		1.000 (0.00)	0.712 (0.21)	0.314(0.27)	
FPR	1	0.001 (0.00)	0.019 (0.02)	0.002 (0.00)	0.779 (0.06)	0.002 (0.01)	0.024 (0.03)	0.003(0.01)	0.689(0.08)
	2	0.002 (0.01)	0.010 (0.01)	0.002 (0.01)		0.001 (0.00)	0.012 (0.02)	0.003(0.01)	
I–E interactions									
TPR	1	1.000 (0.00)	0.949 (0.09)	0.726 (0.26)	0.932 (0.11)	1.000 (0.00)	0.852 (0.13)	0.550(0.21)	0.876(0.14)
	2	1.000 (0.00)	0.961 (0.08)	0.743 (0.24)		1.000 (0.00)	0.903 (0.09)	0.562(0.22)	
FPR	1	0.027 (0.01)	0.037 (0.01)	0.019 (0.01)	0.784 (0.06)	0.031 (0.00)	0.039 (0.02)	0.023(0.01)	0.695(0.07)
	2	0.027 (0.01)	0.062 (0.01)	0.018 (0.01)		0.031 (0.00)	0.067 (0.01)	0.023(0.01)	

$\hat{K}$, estimated number of subgroups; ARI, adjusted rand index; TPR, true positive rate; FPR, false positive rate.

however, an unsatisfactory FP rate of 0.780, which is caused by failing to account for heterogeneity. With partially dichotomized variables, the performance of the proposed approach may slightly deteriorate, which is as expected. We have also experimented with a few other simulation settings and made similar observations.

4. DATA ANALYSIS

We analyze TCGA data on lung squamous cell cancer (LUSC), a major subtype of nonsmall-cell lung cancer. For lung cancer, heterogeneity analysis, both supervised and unsupervised, has been extensively conducted. Such analysis can assist more accurately classifying disease and delivering more customized treatment. As referred to in Section 1, some of such analysis have been based on histopathological imaging data. Here, we further advance such analysis to incorporate I–E interactions. Data are downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/). The response variable of interest is FEV1 (forced expiratory volume in one second), which is a measure of the amount of air exhaled forcefully in 1 s. It is an important biomarker for lung capacity and has been associated with prognosis and other lung cancer outcomes. The histogram in Figure S.6 of the Supplementary material available at *Biostatistics* online shows that there may be two "peaks" around FEV1 = 65 and 77, respectively, suggesting that it may be sensible to assume a mixture distribution and examine heterogeneity. For E variables, we consider age, sex, smoking, and cancer stage, all of which have been associated with lung cancer outcomes and biomarkers. As such, variable selection is not of interest for these variables. Here we note that we have taken a "looser" definition of E variables, and that interaction analysis incorporating clinical and demographic variables has been strongly advocated in recent studies. In particular, these variables have been considered in the I–E interaction analysis under homogeneity (Xu *and others*, 2019). The imaging feature extraction pipeline is briefly sketched in Figure 1. Briefly, a histopathological slide (panel 2) obtained from biopsy is chopped into subimages (panel 3). Then 20 subimages are randomly selected (panel 4). These subimages are

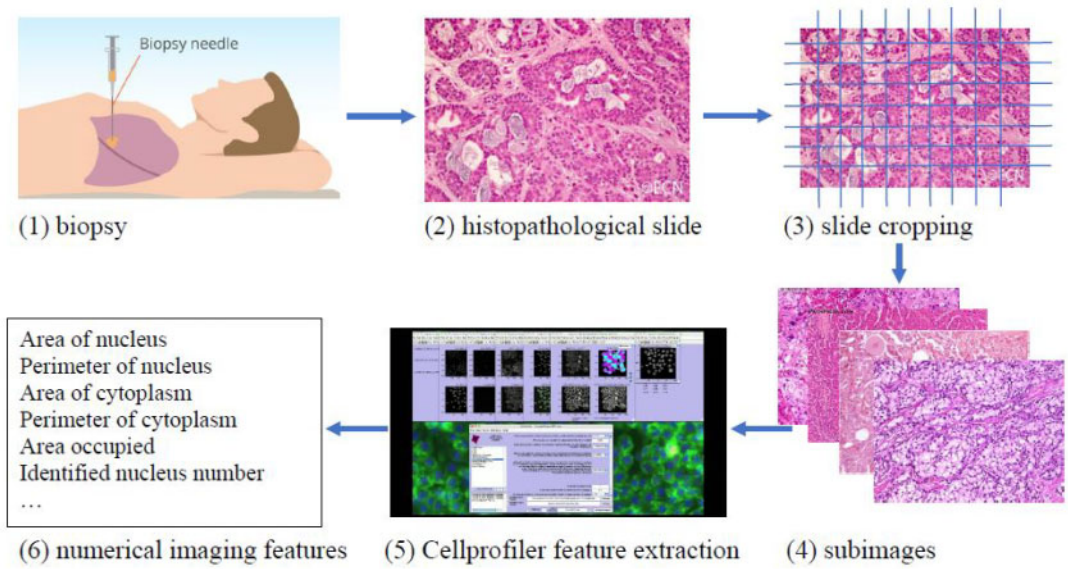

Fig. 1. Pipeline for extracting imaging features.

fed into Cellprofiler (https://cellprofiler.org/), a publicly available digital image processing software, for feature extraction (panel 5). Feature values are averaged over these 20 subimages, and irrelevant features (for example, time) are removed, leading to the final set of imaging features for analysis (panel 6). After removing subjects with missing response and E variable values, the final analyzed data contains 139 imaging features and the aforementioned four E variables on 164 subjects.

When implementing the proposed approach, we set $K \sim \text{Geo}(\text{q} = 0.5)$. As such, the prior encourages a relatively small number of subgroups by assigning 99% probability to $K \leq 7$, which is appropriate considering the limited sample size. We have tried other values of q, which lead to rather similar results. For π_0 and π_1, we assign the Beta(10, 4) and Beta(4, 4) priors, respectively, under which about 70% of the imaging features are noises, and 50% of the E variables interact with the imaging features. Such numbers may be higher than in the published literature, allowing for "sufficient room" for discovery. The rest of the hyperparameter values are set as $(a_0, b_0, \alpha, \tau_m, a_\lambda, b_\lambda, \mu_0, \xi^2, \mu_\eta, \Sigma_\eta) = (3, 1, 1, 10, 1, 1, 0, 1, \mathbf{0}_4, 10I_4)$, either to reflect vague beliefs on the prior distributions or to make the computation stable. We perform four independent MCMC runs, with 10 000 iterations for burn-in and 40 000 iterations after the burn-in. Every other iterations are collected to reduce storage cost. The four runs are carefully examined, and satisfactory convergence is observed. As a testament, in Figure S.8 of the Supplementary material available at *Biostatistics* online, for the estimated marginal inclusion probabilities (upper triangle) and estimates of the subgroup-specific regression coefficients (lower triangle), we show the pairwise comparisons between the four chains. The final results are based on pooling the outputs of the four runs.

The posterior distribution of the number of subgroups and its trace plot are shown in Figure S.7 of the Supplementary material available at *Biostatistics* online, which suggests that there are most likely two subgroups. Conditioning on $K = 2$, the MCMC draws are postprocessed to address label-switching (Papastamoulis, 2016, Algorithm 5). The two subgroups have respective sizes of 36 and 128 on average. In Figure 2 and Table 2, for each of the two subgroups, we present the ten most important main effects and their interactions, where importance is measured by PIP. We note that Figure 2 shows that most of the PIPs are considerably smaller than those observed in simulation, which is expected as a result of significantly weaker signals and more complicated correlation structures. Table 2 shows that the "main effects, interactions" hierarchy is respected, and that the two subgroups are significantly different in which set of covariates are the most influential for the response. More details are in Figure S.9 of the Supplementary material available at *Biostatistics* online, which shows the approximate posterior distribution of β^*_{djl} with the highest inclusion probabilities for subgroups 1 and 2.

Unlike in simulation, we do not know the true data generating model, and hence cannot directly evaluate subgrouping and variable selection performance. In addition, as discussed in Xu *and others* (2019) and references therein, high-dimensional imaging features extracted using digital image processing software do not have simple biological interpretations, making it impossible to "verify" the findings based on the selected variables. To fill this gap, we conduct a small real data-based simulation. In particular, the estimated subgrouping structures, top three imaging features for each subgroup, and their estimated main and interaction effects obtained above are taken as the true. Random errors are generated from a normal distribution as in simulation, and the response values are computed from the linear regression models. With 100 replicates, for subgrouping accuracy, the mean (sd) values are 0.793 (0.08). For the main I effects, the mean (sd) TP values are 0.927 (0.14) and 0.997 (0.03), and the mean (sd) FP values are 0.021 (0.01) and 0.006 (0.01). For the I–E interactions, the mean (sd) TP values are 0.993 (0.04) and 0.960 (0.10), and the mean (sd) FP values are 0.004 (0.00) and 0.008 (0.00). These results suggest that the subgrouping and identification findings in the above data analysis are reasonably credible.

Data are also analyzed using the alternatives, which lead to significantly different subgrouping, selection, and estimation results. As it is impossible to determine which set of results is more sensible, we conduct an indirect evaluation. Specifically, one subject is removed and form the testing data. The rest of the subjects form the training data and are analyzed using the proposed and alternative methods. The

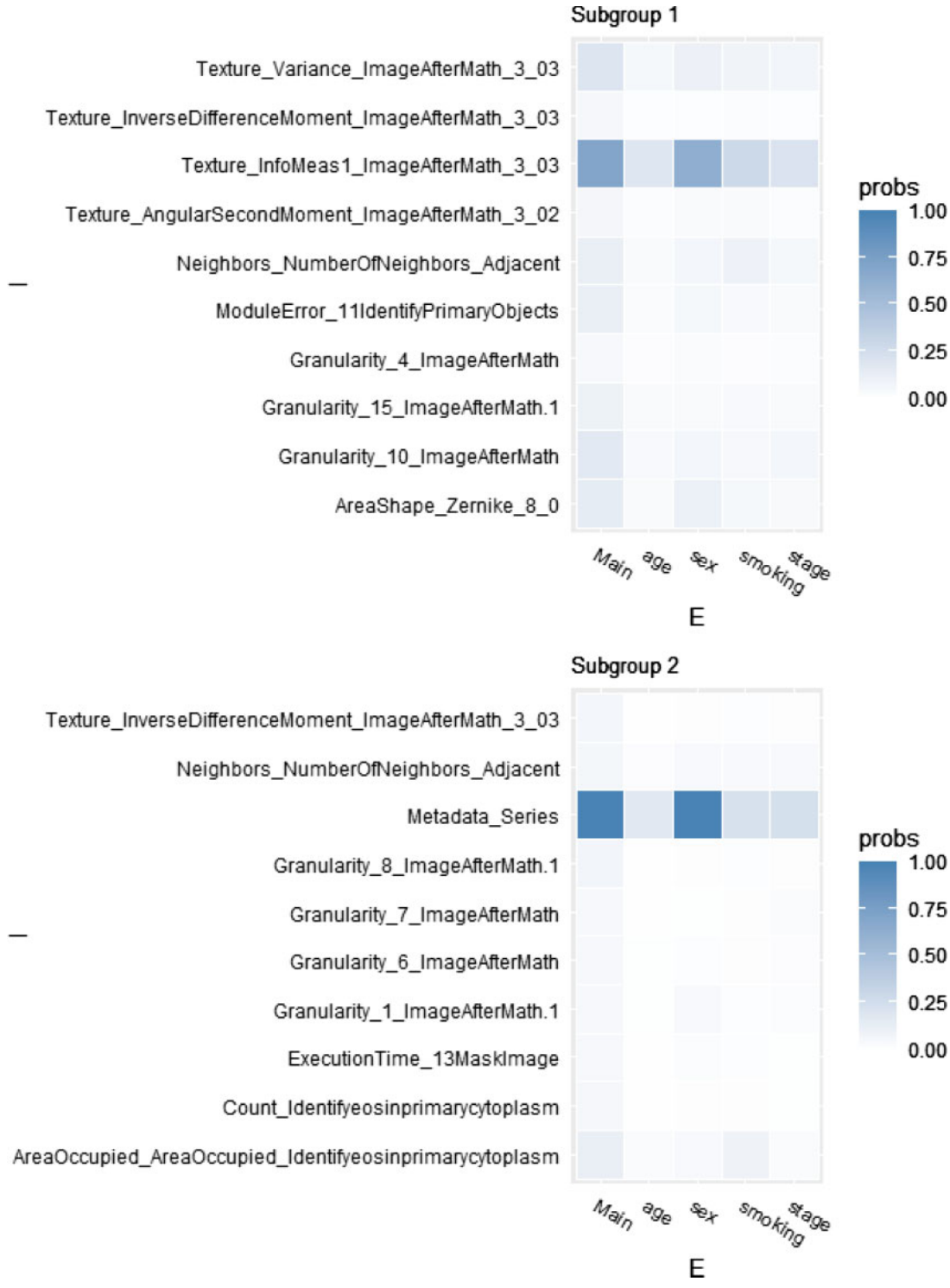

Fig. 2. Data analysis, PIPs for the top ten imaging features for the two subgroups separately.

Table 2. *Data analysis, top ten imaging features (for each subgroup), and their estimated main effects and interactions*

Image feature	Main effect	Interaction with age	sex	smoking	stage
Subgroup 1					
Texture_Variance_ImageAfterMath_3_03	37.751		−22.214		
Texture_InverseDifferenceMoment_ImageAfterMath_3_03	6.016				
Texture_InfoMeas1_ImageAfterMath_3_03	34.927	−2.519	−34.746	9.512	1.343
Texture_AngularSecondMoment_ImageAfterMath_3_02	−3.738				
Neighbors_NumberOfNeighbors_Adjacent	−8.971				
ModuleError_11IdentifyPrimaryObjects	15.694				
Granularity_4_ImageAfterMath	1.488				
Granularity_15_ImageAfterMath.1	−103.689				
Granularity_10_ImageAfterMath	−12.48				
AreaShape_Zernike_8_0	22.888		−30.9		
Subgroup 2					
Texture_InverseDifferenceMoment_ImageAfterMath_3_03	3.995				
Neighbors_NumberOfNeighbors_Adjacent	−0.958				
Metadata_Series	−100.833	1.262	100.984	1.417	0.111
Granularity_8_ImageAfterMath.1	5.579				
Granularity_7_ImageAfterMath	−1.818				
Granularity_6_ImageAfterMath	−9.533				
Granularity_1_ImageAfterMath.1	6.934				
ExecutionTime_13MaskImage	−11.893				
Count_Identifyeosinprimarycytoplasm	−4.64				
AreaOccupied_AreaOccupied_Identifyeosinprimarycytoplasm	0.84			9.394	

training data estimation is used for predicting the testing data. The "removal, estimation, and prediction" process is repeated across all the subjects. We note that it is not entirely clear which subgroup the removed subject belongs to and hence which model should be used. We use the subgroup/model that most of its subgroup members in the whole-data analysis belong to. The average squared roots of prediction MSE (mean squared error) values are 18.36 (proposed), 20.44 (FMRLasso), and 19.33 (ICC), which provides some support to the superiority of the proposed analysis.

5. Discussion

In this study, we have significantly expanded the scope of supervised cancer heterogeneity analysis by developing a Bayesian FMR approach that incorporates histopathological imaging features and, more importantly, their interactions with E variables. This study has also provided an alternative way for analyzing cancer studies and histopathological imaging data. As described above, the proposed approach also has multiple technical innovations, such as respecting the "main effects, interactions" hierarchy in Bayesian analysis and not specifying the number of subgroups. Simulation has shown its advantageous performance over the close competitors. In the analysis TCGA LUSC data, this study is the first to identify two subgroups based on imaging data. The data-based simulation and prediction evaluation can provide solid support to the credibility of our findings.

This study can be potentially extended in multiple ways. First, under the current model assumptions, the mixture probabilities do not depend on covariates. There are multiple ways to relax this assumption.

One is to specify a mixture of joint distributions of the response and covariates, and then the probabilities of a subject belonging to different subgroups will vary with the value of its covariates. Another possibility is to directly model the mixture probabilities, say, using logistic regression based on a subset of covariates. These extensions demand additional model assumptions and may incur higher computational cost. Further, it is of interest to consider alternative data types and models. With complex data structures and analysis objectives, computation is more expensive than in some existing studies and may further increase for larger data. It is of interest to develop more efficient computation. The adopted priors are the most popular in existing literature and also computationally simpler. Yet, it may be of interest to consider alternative priors. Comparing with additional alternatives in data analysis can help further establish the superiority of the proposed approach. Finally, weaker signals have been observed in data analysis, and only internal prediction evaluation has been conducted. Conducting experimental validation, although significant, is beyond our scope. It is of interest to search for powerful data to generate more definitive findings and conduct external prediction evaluation.

6. SOFTWARE

Software written in Julia, together with a brief readme file, is available at github.com/shuanggema/BHA-hdInt.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We thank the editors and reviewers for their careful review and insightful comments.
Conflict of Interest: None declared.

FUNDING

NSF (1916251) and NIH (CA204120).

REFERENCES

BALIU-PIQUÉ, M., PANDIELLA, A. AND OCANA, A. (2020). Breast cancer heterogeneity and response to novel therapeutics. *Cancers* **12**, 3271.

BARBIERI, M. M. AND BERGER, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics* **32**, 870–897.

BELHOMME, P., TORALBA, S., PLANCOULAINE, B., OGER, M., GURCAN, M. N. AND BOR-ANGELIER, C. (2015). Heterogeneity assessment of histological tissue sections in whole slide images. *Computerized Medical Imaging and Graphics* **42**, 51–55.

BIEN, J., TAYLOR, J. AND TIBSHIRANI, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics* **41**, 1111–1141.

BURRELL, R. A., MCGRANAHAN, N., BARTEK, J. AND SWANTON, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345.

CHEN, M., ZHANG, B., TOPATANA, W., CAO, J., ZHU, H., JUENGPANICH, S., MAO, Q., YU, H. AND CAI, X. (2020a). Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precision Oncology* **4**, 1–7.

CHEN, S., ZHANG, N., JIANG, L., GAO, F., SHAO, J., WANG, TAO, ZHANG, E., YU, H., WANG, X. AND ZHENG, J. (2020b). Clinical use of a machine learning histopathological image signature in diagnosis and survival prediction of clear cell renal cell carcinoma. *International Journal of Cancer* **148**, 780–790.

DEY, T., ISHWARAN, H. AND RAO, J. S. (2008). An in-depth look at highest posterior model selection. *Econometric Theory*, 377–403.

ECHLE, A., RINDTORFF, N. T., BRINKER, T. J., LUEDDE, T., PEARSON, A. T. AND KATHER, J. N. (2020). Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, 1–11.

FRÜHWIRTH-SCHNATTER, S., CELEUX, G. AND ROBERT, C. P. (2018). *Handbook of Mixture Analysis*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

GUPTA, M. AND IBRAHIM, J. G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association* **102**, 867–880.

HE, B., ZHONG, T., HUANG, J., LIU, Y., ZHANG, Q. AND MA, S. (2020). Histopathological imaging-based cancer heterogeneity analysis via penalized fusion with model averaging. *Biometrics*, doi: 10.1111/biom.13357.

HUBERT, L. AND ARABIE, P. (1985). Comparing partitions. *Journal of Classification* **2**, 193–218.

KHALILI, A. AND CHEN, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102**, 1025–1038.

KIM, J. AND DEBERARDINIS, R. J. (2019). Mechanisms and implications of metabolic heterogeneity in cancer. *Cell Metabolism* **30**, 434–446.

KIM, J., LIM, J., KIM, Y. AND JANG, W. (2018). Bayesian variable selection with strong heredity constraints. *Journal of the Korean Statistical Society* **47**, 314–329.

LEE, K.-J., CHEN, R.-B. AND WU, Y. N. (2016). Bayesian variable selection for finite mixture model of linear regressions. *Computational Statistics and Data Analysis* **95**, 1–16.

LEY, E. AND STEEL, M. F. J. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics* **24**, 651–674.

LI, Q., SHI, R. AND LIANG, F. (2019). Drug sensitivity prediction with high-dimensional mixture regression. *PLoS One* **14**, 1–18.

LIU, W., ZHANG, B., ZHANG, Z., TAO, J. AND BRANSCUM, A. J. (2015). Model selection in finite mixture of regression models: a Bayesian approach with innovative weighted g priors and reversible jump Markov chain Monte Carlo implementation. *Journal of Statistical Computation and Simulation* **85**, 2456–2478.

LUO, X., ZANG, X., YANG, L., HUANG, J., LIANG, F., RODRIGUEZ-CANALES, J., WISTUBA, I. I., GAZDAR, A., XIE, Y. AND XIAO, G. (2017). Comprehensive computational pathological image analysis predicts lung cancer prognosis. *Journal of Thoracic Oncology* **12**, 501–509.

MCLACHLAN, G. J. AND PEEL, D. (2000). *Finite Mixture Models*. Wiley-Interscience.

MILLER, J. W. AND HARRISON, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**, 340–356.

MORRISON, C. D., LIU, P., WOLOSZYNSKA-READ, A., ZHANG, J., LUO, W., QIN, M., BSHARA, W., CONROY, J. M., SABATINI, L., VEDELL, P. *and others*. (2014). Whole-genome sequencing identifies genomic heterogeneity at a nucleotide and chromosomal level in bladder cancer. *Proceedings of the National Academy of Sciences United States of America* **111**, E672–E681.

NARISETTY, N. N. AND HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics* **42**, 789–817.

PAPASTAMOULIS, P. (2016). label.switching: an r package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software* **69**, 1–24.

PAPASTAMOULIS, P. AND ILIOPOULOS, G. (2010). An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics* **19**, 313–331.

RODRÍGUEZ, C. E. AND WALKER, S. G. (2014). Label switching in Bayesian mixture models: deterministic relabeling strategies. *Journal of Computational and Graphical Statistics* **23**, 25–45.

SCHLATTMANN, P. (2009). *Medical Applications of Finite Mixture Models*. Statistics for Biology and Health. Berlin Heidelberg: Springer.

SCOTT, J. G. AND BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 2587–2619.

STÄDLER, N., BÜHLMANN, P. AND VAN DE GEER, S. (2010). l_1-penalization for mixture regression models. *Test* **19**, 209–256.

WADE, S. AND GHAHRAMANI, Z. (2018). Bayesian cluster analysis: point estimation and credible balls (with Discussion). *Bayesian Analysis* **1**3, 559–626.

XU, X. AND GHOSH, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis* **10**, 909–936.

XU, Y., ZHONG, T., WU, M. AND MA, S. (2019). Histopathological imaging-environment interactions in cancer modeling. *Cancers (Basel)* **11**, 579.

[*Received February 20, 2021; revised July 28, 2021; accepted for publication October 1, 2021*]