Biostatistics (2021), **0**, 0, *pp.* 1–14
doi:10.1093/biostatistics/output

Supplementary Materials for "Bayesian Finite Mixture of Regression Analysis for Cancer based on Histopathological Imaging-Environment Interactions" by

YUNJU IM[1], YUAN HUANG[1], AIXIN TAN[2], SHUANGGE MA[1*]

[1] *Department of Biostatistics, Yale School of Public Health, New Haven 06520, USA*

[2] *Department of Statistics and Actuarial Science, University of Iowa, Iowa City, 52242, USA*

This file contains additional discussions and numerical results referenced in the main text.

1. Identifiability and Consistency of the proposed Bayesian model

1.1 *Identifiability*

Parameters in a finite mixture model are not identifiable due to label switching (Redner and Walker, 1984). Nevertheless, there is usually a form of "local identifiability" guaranteeing the existence of subsets of the parameter space within which the parameters are identifiable. See for example Kim and Lindsay (2015). Here we provide heuristic discussions on identifiability in the context of our model.

With respect to the likelihood function defined in (2.1) in Section 2 of the article, with $\boldsymbol{\delta}$ integrated out, the space of the vector of parameters $\theta = (\boldsymbol{\beta}^{*T}, \boldsymbol{\eta}^T, \sigma^2, K, \boldsymbol{p}^T)^T$, say Θ, is unidentifiable if there exists $\theta, \theta' \in \Theta$, $\theta \neq \theta'$ such that

$$f(\boldsymbol{y}|\boldsymbol{X}, \theta) = f(\boldsymbol{y}|\boldsymbol{X}, \theta') \text{ for all } (\boldsymbol{y}, \boldsymbol{X}), \tag{S.1}$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ and $\boldsymbol{X} = (\boldsymbol{z}_1, \boldsymbol{w}_1, \cdots, \boldsymbol{z}_n, \boldsymbol{w}_n)^T$ denote the response and covariates for all subjects in the sample, respectively, and

$$f(\boldsymbol{y}|\boldsymbol{X},\theta) = \Pi_{i=1}^{n}\left[\sum_{j=1}^{K} p_k f(y|z,w,\boldsymbol{\beta}_j^*,\boldsymbol{\eta},\sigma)\right]. \tag{S.2}$$

Consider the parameter space $\Theta = \{\theta : K > 0; p_1 \cdots p_K > 0; (\beta^*_{d0}, \boldsymbol{\beta}^{*T}_d) \neq (\beta^*_{d'0}, {\boldsymbol{\beta}^*_{d'}}^T)$ for any $d \neq d'\}$, where the restrictions exclude trivial or equivalent mixing components. Then Θ is still unidentifiable under label switching, because, for example, $\theta = (\boldsymbol{\beta}^{*T} = (\boldsymbol{\beta}_1^{*T}, \boldsymbol{\beta}_2^{*T}), \boldsymbol{\eta}^T, \sigma^2, K = 2, \boldsymbol{p}^T = (\frac{1}{2}, \frac{1}{2}))^T$ and a relabelling of its two equally weighted mixing components, $\theta' = (\boldsymbol{\beta}^{*T} = (\boldsymbol{\beta}_2^{*T}, \boldsymbol{\beta}_1^{*T}), \boldsymbol{\eta}^T, \sigma^2, K = 2, \boldsymbol{p}^T = (\frac{1}{2}, \frac{1}{2}))^T$ satisfy equation (S.1). For the corresponding Bayesian model, if the priors are chosen to be invariant under label switching, then the posterior conditioning on any given K has $K!$ symmetric regions. This symmetry makes some standard Bayesian inference (such as calculating marginal posterior means) meaningless for group-specific parameters. There is a rich literature on how to tackle this problem. In this study, we use the R package *label.switching* for the post processing of MCMC when needed.

In the regression setup, for the likelihood function (2.1) in Section 2 of this article, we state below another possible source of nonidentifiability that we refer to as the partial switching nonidentifiability, which can be avoided under certain conditions, or otherwise makes little trouble in practice by considering a restricted parameter space. As an illustration, consider the case of $K = 2$, $\boldsymbol{X}$ is of dimension 1, and there are exactly two distinct values observed for $\boldsymbol{X}$, namely, $\{\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\}$. Then for any $\boldsymbol{\beta}^* = (\beta^*_{10}, \beta^*_{20}, \beta^*_1, \beta^*_2)^T$ written in short as $(\alpha_1, \alpha_2, \beta_1, \beta_2)^T$, define $\boldsymbol{\beta}^{*\prime}$ as the solution to the linear system:

$$\begin{bmatrix} 1 & \boldsymbol{X}^{(1)} & 0 & 0 \\ 1 & \boldsymbol{X}^{(2)} & 0 & 0 \\ 0 & 0 & 1 & \boldsymbol{X}^{(1)} \\ 0 & 0 & 1 & \boldsymbol{X}^{(2)} \end{bmatrix} \begin{bmatrix} \alpha_1' \\ \beta_1' \\ \alpha_2' \\ \beta_2' \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & \boldsymbol{X}^{(1)} \\ 1 & \boldsymbol{X}^{(2)} & 0 & 0 \\ 1 & \boldsymbol{X}^{(1)} & 0 & 0 \\ 0 & 0 & 1 & \boldsymbol{X}^{(2)} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix}. \tag{S.3}$$

Then $\boldsymbol{\beta}^{*\prime} \neq \boldsymbol{\beta}^*$ as long as $(\alpha_1, \beta_1) \neq (\alpha_2, \beta_2)$. Equation (S.3) suggests that, at $\boldsymbol{X}^{(1)}$, the mean responses of subgroups 1 and 2 under θ agree with those of subgroups 2 and 1 under θ', respectively (hence the term "partial switching"); while at $\boldsymbol{X}^{(2)}$, the mean responses of subgroups 1 and 2 under θ agree with those of subgroups 1 and 2 under θ', respectively. Eventually, when the label is integrated out in (S.2), the set of means of y in the mixing components is the same under $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^{*\prime}$. Let θ and θ' have group-specific regression coefficients equal to $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$, respectively, with common values for the other components. Then (S.1) holds for the marginal likelihood in (S.2). One can see this cannot happen if there are three or more distinct values observed for $\boldsymbol{X}$. More generally, for q-dimensional $\boldsymbol{X}$, the aforementioned partial switching nonidentifiability in the regression coefficients will not occur as long as the number of distinct values of $\boldsymbol{X}$ in each dimension is three or more. Now consider regression setups where some covariates are dichotomous. Let S denote the set of indices of dichotomous covariates, and A denote the set of parameter values such that the vectors of group-wise coefficients for all non-dichotomous covariates perfectly coincide in some subgroups, that is,

$$A = \{\theta : \boldsymbol{\beta}^*_{dl} = \boldsymbol{\beta}^*_{d'l} \text{ for all } l \notin S, \text{ for some } d \neq d', d, d' \in \{1, \cdots, K\}\}.$$

Then the mixture model (S.2) will not suffer from partial switching nonidentifiability on the restricted parameter space Θ/A. Therefore, as long as A receives zero prior probability, there will be no practical problem caused by this type of nonidentifiability. And the aforementioned methods for handling the standard label switching nonidentifiability suffices.

1.2 *Consistency*

Broadly speaking, the Doob's Theorem (Doob, 1949; Schwartz, 2008; Miller, 2018) says that, a fixed Bayesian model with proper priors is posterior consistent under very general conditions. Specifically for finite mixture models where the observations are iid, Nobile (1994, chap. 3) uses the Doob's Theorem to prove their posterior consistency. It is expected that such results can be extended to the mixture of regression models in (2.1), if there are a finite number of covariates, the model is identifiable, the prior on K assigns a nonzero probability to the true number of subgroups, and the form of the mixture components (that is, the linear regression model based on a subset of the covariates within each group) are correctly specified.

To establish consistency when the number of covariates increases with n is very challenging (and it may have more theoretical than practical implications). Here we provide a brief and heuristic discussion. First, to achieve consistency, it is necessary to require the hyperparameters in the spike and slab prior for the regression coefficients to depend on the number of covariates. To see this, suppose that $(1 - \pi_0)$, the prior inclusion probability of a given imaging feature, is fixed. Then the prior assigns disappearing probabilities to models of size $o(p)$, failing to achieve sparsity. On the other hand, suppose that τ^*, the variance of the slab part, is fixed. Then the prior will be increasingly informative and overwhelm information from the sample. Indeed, had the subgroups been correctly identified, for each subgroup, there would be various consistency results for Bayesian variable selection models in the literature. One such result is given in Narisetty and He (2014) on strong selection consistency, in the sense that the model that corresponds to the true subset of covariates is assigned the highest posterior probability as sample size increases. Essentially, the conditions needed for the prior in the context of our problem are that the prior expected number of selected variables is bounded and that τ^* grows to infinity as the number of covariates grows. We have incorporated such considerations when specifying priors in our numerical studies.

2. Additional discussions on updating the subgroup memberships

If the subgroup-specific parameters can be integrated out, and the marginal likelihood has a closed form, which is generally the case when conjugate priors are used, then one can use the algorithm referred to as “Algorithm 3” in Miller and Harrison (2018) when updating the latent-subgroup memberships. In our model, non-conjugate priors for the subgroup-specific parameters are used, and the marginal likelihood cannot be easily computed. Hence we choose to adopt the auxiliary variable approach, which is referred to as “Algorithm 8” in Miller and Harrison (2018).

The proposed sampler can be potentially coupled with the split-merge type algorithm, which is referred to as Jain-Neal algorithm in Miller and Harrison (2018). The main purpose of using the split-merge type algorithm is to improve the mixing on the number of subgroups. We instead use a relatively big number of auxiliary variables, $m = 10$, when updating the subgroup memberships, which we have found helpful for the mixing on the number of subgroups and sufficient for the efficiency of the proposed sampler (Neal, 2000).

3. Additional Numerical Results

Table S.1: Simulation results for Scenarios 3 and 4: mean(sd) based on 100 replicates. $\hat{K}$: estimated number of subgroups; ARI: adjusted rand index; TPR: true positive rate; FPR: false positive rate.

	Subgroup	Scenario 3				Scenario 4			
		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		0.860(0.05)	0.703(0.10)	0.787(0.13)		0.807(0.05)	0.653(0.12)	0.604(0.15)	
Main I Effects									
TPR	1	0.988(0.06)	0.504(0.36)	0.100(0.15)	0.951(0.09)	1.000(0.00)	0.890(0.22)	0.420(0.30)	0.952(0.10)
	2	1.000(0.00)	0.844(0.27)	0.950(0.18)		1.000(0.00)	0.828(0.25)	0.290(0.28)	
FPR	1	0.001(0.00)	0.023(0.03)	0.001(0.00)	0.762(0.08)	0.000(0.00)	0.010(0.01)	0.001(0.00)	0.728(0.05)
	2	0.001(0.00)	0.004(0.01)	0.001(0.00)		0.001(0.00)	0.005(0.01)	0.000(0.00)	
I-E Interactions									
TPR	1	0.976(0.09)	0.657(0.22)	0.440(0.17)	0.958(0.07)	1.000(0.00)	0.892(0.19)	0.609(0.22)	0.952(0.10)
	2	1.000(0.00)	0.967(0.08)	0.962(0.13)		1.000(0.00)	0.882(0.18)	0.583(0.22)	
FPR	1	0.011(0.01)	0.035(0.02)	0.015(0.00)	0.770(0.08)	0.004(0.00)	0.030(0.01)	0.013(0.00)	0.731(0.05)
	2	0.007(0.01)	0.091(0.02)	0.018(0.01)		0.003(0.00)	0.042(0.01)	0.012(0.00)	
Discrete E variables, a block-diagonal covariance matrix for I features									
$\hat{K}$		2.000(0.00)				2.000(0.00)			
ARI		0.863(0.05)	0.751(0.08)	0.791(0.15)		0.809(0.06)	0.672(0.10)	0.631(0.15)	
Main I Effects									
TPR	1	0.980(0.10)	0.690(0.33)	0.174(0.20)	0.950(0.07)	1.000(0.00)	0.914(0.17)	0.566(0.32)	0.948(0.09)
	2	1.000(0.00)	0.914(0.20)	0.934(0.19)		1.000(0.00)	0.840(0.20)	0.434(0.32)	
FPR	1	0.001(0.00)	0.037(0.04)	0.001(0.00)	0.769(0.07)	0.000(0.00)	0.010(0.01)	0.001(0.00)	0.726(0.05)
	2	0.000(0.00)	0.009(0.01)	0.002(0.00)		0.000(0.00)	0.004(0.01)	0.001(0.00)	
I-E Interactions									
TPR	1	0.976(0.12)	0.744(0.19)	0.444(0.18)	0.956(0.06)	1.000(0.00)	0.911(0.15)	0.683(0.26)	0.948(0.09)
	2	1.000(0.00)	0.988(0.04)	0.954(0.14)		1.000(0.00)	0.912(0.15)	0.661(0.25)	
FPR	1	0.027(0.01)	0.033(0.02)	0.014(0.00)	0.776(0.07)	0.013(0.00)	0.021(0.01)	0.010(0.00)	0.729(0.05)
	2	0.018(0.01)	0.081(0.02)	0.020(0.01)		0.013(0.00)	0.034(0.01)	0.011(0.00)	
Continuous E variables, a banded covariance matrix for I features									
$\hat{K}$		2.000(0.00)				2.001(0.10)			
ARI		0.850(0.05)	0.713(0.08)	0.798(0.10)		0.814(0.06)	0.673(0.13)	0.624(0.16)	
Main I Effects									
TPR	1	0.988(0.06)	0.446(0.34)	0.080(0.13)	0.940(0.08)	1.000(0.00)	0.860(0.22)	0.374(0.30)	0.946(0.11)
	2	1.000(0.00)	0.830(0.27)	0.972(0.12)		1.000(0.00)	0.802(0.22)	0.298(0.28)	
FPR	1	0.001(0.00)	0.023(0.03)	0.000(0.00)	0.738(0.08)	0.001(0.00)	0.011(0.01)	0.000(0.00)	0.734(0.05)
	2	0.001(0.00)	0.007(0.02)	0.001(0.00)		0.001(0.00)	0.005(0.01)	0.001(0.00)	
I-E Interactions									
TPR	1	0.971(0.09)	0.618(0.22)	0.408(0.15)	0.948(0.07)	0.998(0.01)	0.886(0.18)	0.606(0.21)	0.946(0.11)
	2	1.000(0.00)	0.972(0.06)	0.976(0.10)		0.997(0.02)	0.867(0.19)	0.556(0.22)	
FPR	1	0.008(0.01)	0.035(0.02)	0.016(0.00)	0.746(0.07)	0.003(0.00)	0.030(0.01)	0.013(0.00)	0.738(0.05)
	2	0.005(0.01)	0.088(0.02)	0.018(0.01)		0.003(0.00)	0.042(0.01)	0.012(0.00)	
Discrete E variables, a banded covariance matrix for I features									
$\hat{K}$		2.000(0.00)				2.010(0.10)			
ARI		0.847(0.05)	0.752(0.08)	0.772(0.18)		0.819(0.06)	0.701(0.08)	0.660(0.16)	
Main I Effects									
TPR	1	0.982(0.09)	0.616(0.32)	0.144(0.19)	0.945(0.08)	1.000(0.00)	0.904(0.17)	0.520(0.32)	0.938(0.10)
	2	1.000(0.00)	0.928(0.17)	0.906(0.24)		1.000(0.00)	0.852(0.18)	0.430(0.31)	
FPR	1	0.001(0.00)	0.036(0.03)	0.001(0.00)	0.757(0.06)	0.000(0.00)	0.011(0.01)	0.001(0.00)	0.748(0.05)
	2	0.001(0.00)	0.007(0.02)	0.002(0.01)		0.001(0.00)	0.005(0.01)	0.001(0.00)	
I-E Interactions									
TPR	1	0.974(0.10)	0.708(0.19)	0.444(0.17)	0.952(0.07)	1.000(0.00)	0.932(0.11)	0.661(0.22)	0.938(0.10)
	2	1.000(0.00)	0.978(0.11)	0.923(0.20)		1.000(0.00)	0.915(0.10)	0.659(0.26)	
FPR	1	0.027(0.01)	0.033(0.01)	0.015(0.00)	0.770(0.08)	0.013(0.00)	0.022(0.01)	0.011(0.00)	0.751(0.05)
	2	0.020(0.01)	0.076(0.02)	0.020(0.01)		0.013(0.00)	0.035(0.01)	0.011(0.00)	

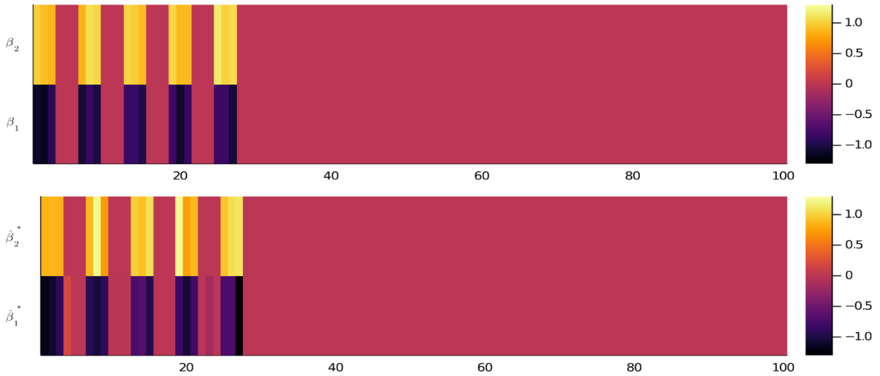

Fig. S.1: Simulation setup 1 and Scenario 1, heatmaps of β_1^* and β_2^*: true (upper) and estimated with one simulation replicate (lower).

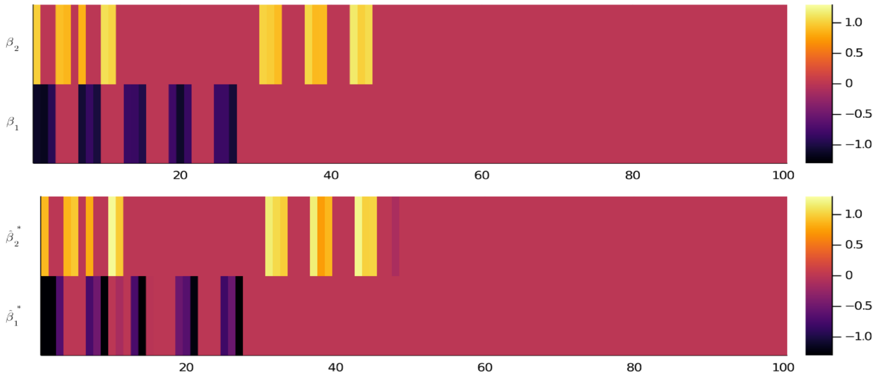

Fig. S.2: Simulation setup 1 and Scenario 3, heatmaps of β_1^* and β_2^*: true (upper) and estimated with one simulation replicate (lower).

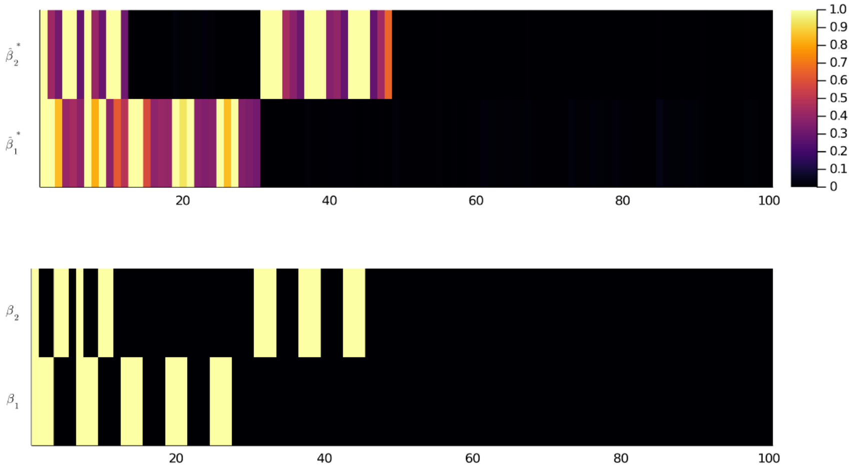

Fig. S.3: Simulation setup 1 and Scenario 3, inclusion probabilities of β_1^* and β_2^*: the estimated with one simulation replicate (upper) and the locations of the true non-zero coefficients in yellow and zero coefficients in black (lower).

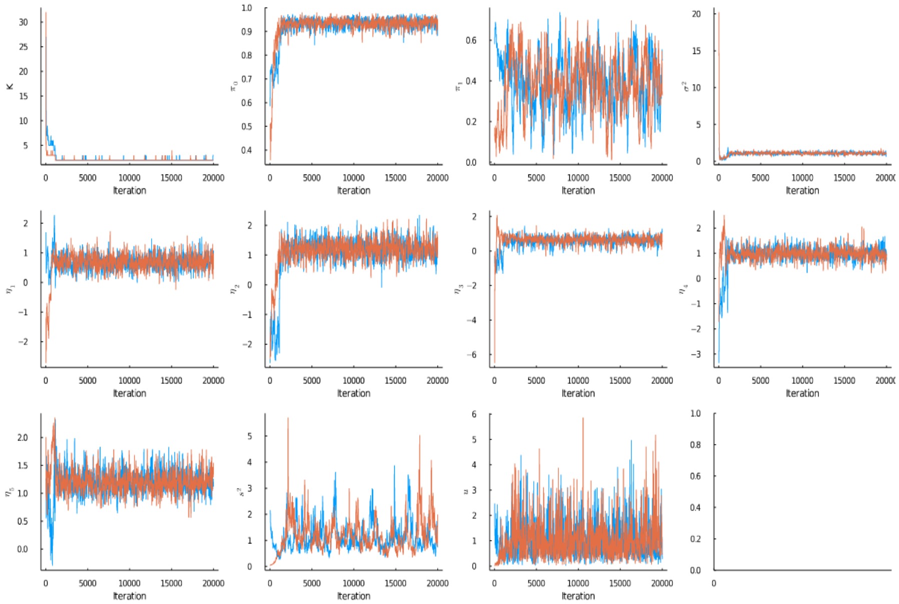

Fig. S.4: Simulation Scenario 1, trace plots for the label-invariant variables, thinned at every 20th iterations of two MCMC chains.

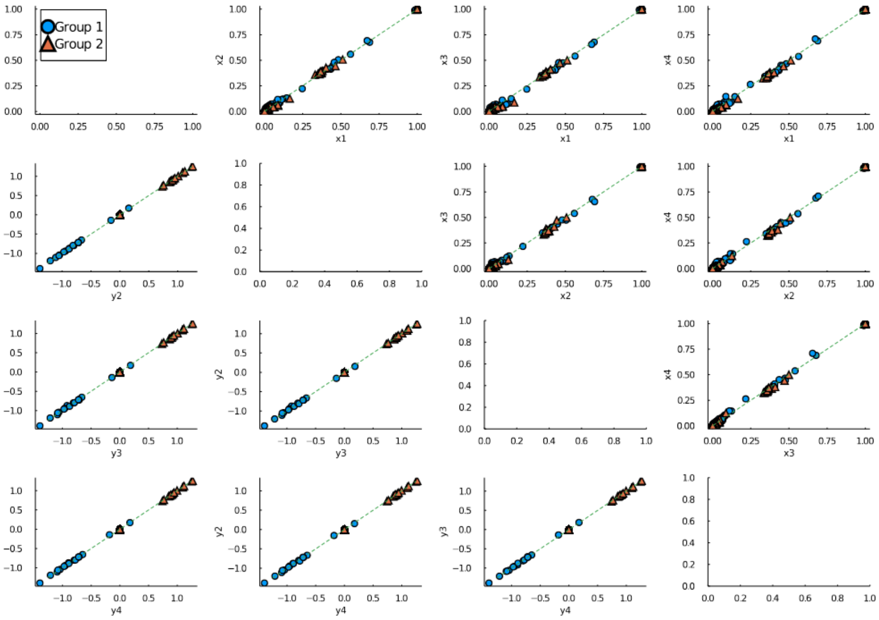

Fig. S.5: Simulation Scenario 1, comparison of four independent MCMC runs. Plots in the upper and lower triangles compare the posterior inclusion probabilities and estimates of β_1^* and β_2^* for run i and j, respectively.

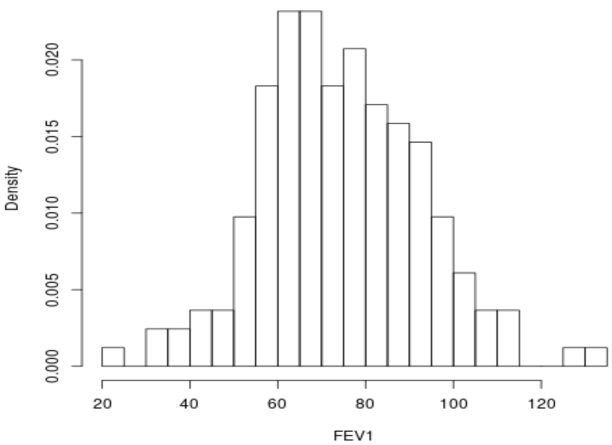

Fig. S.6: Data analysis, histogram of FEV1.

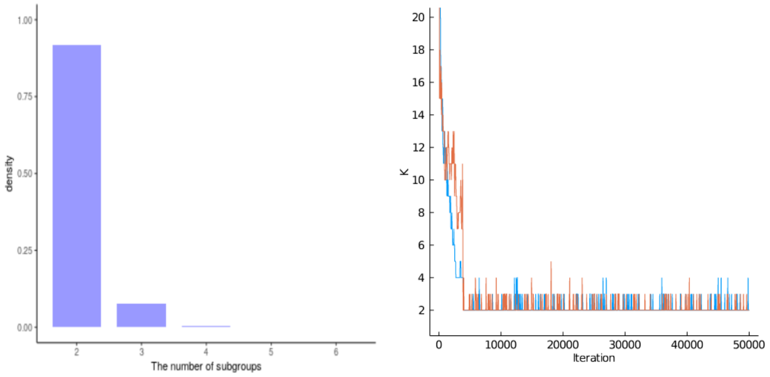

Fig. S.7: Data analysis, posterior distribution of the number of subgroups (left) and its trace plot (right; two of the four MCMC runs are plotted to improve visualization).

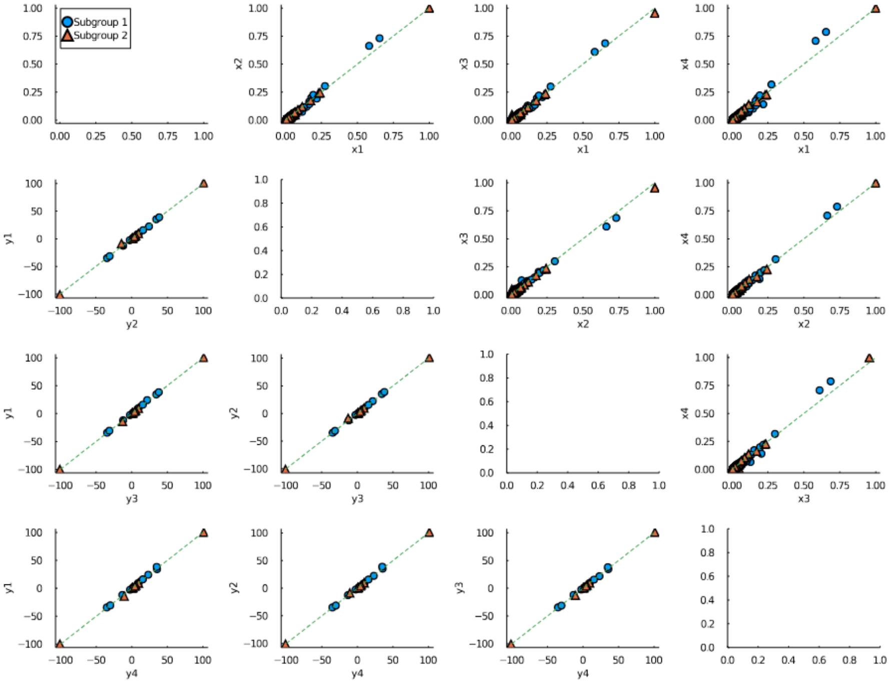

Fig. S.8: Data analysis, comparison of the four independent MCMC runs. Upper triangle: comparison of the posterior inclusion probabilities for the two subgroups. Lower triangle: comparison of the estimates for the top ten β^*_{1jl} and β^*_{2jl} with the highest inclusion probabilities for the two subgroups.

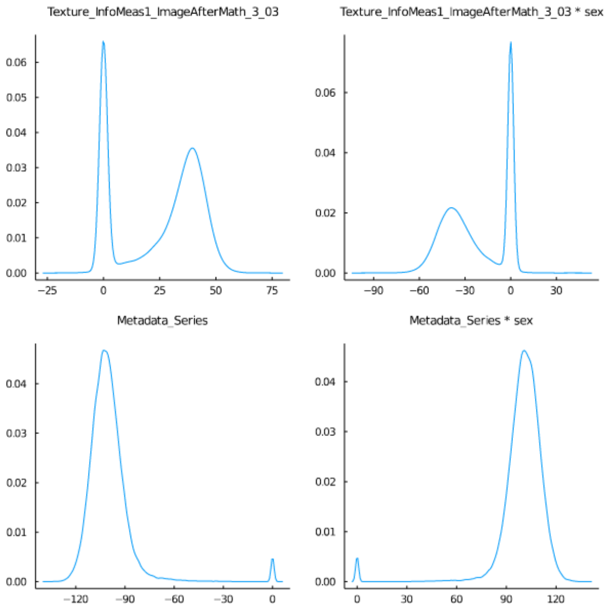

Fig. S.9: Data analysis, density estimation of the posterior distribution of β^*_{djl} for the selected top imaging features for subgroup 1 (upper) and 2 (lower).

References

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, 23–27.

Kim, D. and Lindsay, B.G. (2015). Mixture densities, maximum likelihood and the em algorithm. *Annals of the Institute of Statistical Mathematics* **67**, 745–772.

Miller, J. W. (2018). A detailed treatment of doob's theorem. *arXiv preprint arXiv:1801.03122.*

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.*

Narisetty, N. N. and He, X. *(2014). Bayesian variable selection with shrinking and diffusing priors.* Annals of Statistics ***42**(2), 789–817.*

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

Nobile, A. *(1994). Bayesian analysis of finite mixture distributions.* Ph.D. dissertation, Dept. Statistics, Carnegie Mellon Univ., Pittsburgh.

Redner, R. A. and Walker, H. F. *(1984). Mixture densities, maximum likelihood and the em algorithm.* SIAM Review ***26**(2), 195–239.*

Schwartz, L. *(2008). On bayes procedures.* Probability Theory and Related Fields *4(1), 10–26.*