SUPPLEMENT TO "DETECTING MULTIPLE RANDOM CHANGEPOINTS IN BAYESIAN PIECEWISE GROWTH MIXTURE MODELS"

Abstract

This supplementary document contains additional simulation results and illustrations for the `BayesianPGMM` package for R. Section 1 describes a simulation to assess changepoint detection and recovery for a single class model with up to five changepoints. Section 2 describes an illustrative example with up to four latent classes, and over-specification of the number of latent classes. Section 2 also serves as a brief tutorial for the `BayesianPGMM` R package, with commands to reproduce the results and figures shown. Section 3 describes a simulation study using different Dirichlet priors for the latent class memberships, and Section 4 describes a simulation study using alternative priors for the variances of the random effects.

1. Multiple change-point simulation

Here we describe a simulation study in which longitudinal data are generated with anywhere from 0 to 5 changepoints, and we assess both accuracy in detecting the number of changepoints and estimation accuracy for the mean location of the changepoints. All data are simulated according to the following model, which reflects Equation (3) in the main manuscript:

$$y_{ij} = \beta_{i,0} + \beta_{i,1}x_{ij} + \sum_{k=1}^{5} \beta_{i,k+1}(x_{ij} - \lambda_{i,k})^{+}\mathbb{1}_{\{k \leq \mathcal{K}\}} + \epsilon_{ij},$$

with

- 30 individuals ($i = 1, \ldots, 30$)
- 20 time points ($x_{ij} = 0, \ldots, 19$)
- Intercepts $\beta_{i,0} \sim N(0, 0.05)$
- Potential changepoints

  $\lambda_{i,1} \sim N(3, \sigma_\lambda^2), \lambda_{i,2} \sim N(6, \sigma_\lambda^2), \lambda_{i,3} \sim N(9, \sigma_\lambda^2), \lambda_{i,4} \sim N(12, \sigma_\lambda^2), \lambda_{i,5} \sim N(15, \sigma_\lambda^2)$
- Potential slope changes $\beta_{i,k+1} \sim N\left((-1)^k, 0.05\right)$ for $k = 0, \ldots, 5$
- Error $\epsilon_{ij} \sim N(0, 0.5)$.

The manipulated conditions are the number of changepoints present, $\mathcal{K} = \{0, 1, 2, 3, 4, 5\}$, and the standard deviation of the individual changepoints, $\sigma_\lambda = \{0.2, 0.5\}$.

We generate 30 replicated datasets for each combination of the initial conditions, yielding $30 \times 6 \times 2 = 360$ simulated datasets. For each simulation, we estimate the model as described in the main manuscript with a maximum of 5 possible changepoints ($K = 5$). The prior for the number of changepoints is specified as $\mathcal{K} \sim \text{Binomial}(5, 0.5)$, where 5 is the number of potential changepoints and 0.5 is the probability of including each changepoint.

Table 1.

True number of changepoints ($\mathcal{K}$) and mean posterior probability $\hat{\mathcal{K}}$.

$\sigma_\lambda = \mathbf{0.2}$	$\mathcal{K} = 0$	$\mathcal{K} = 1$	$\mathcal{K} = 2$	$\mathcal{K} = 3$	$\mathcal{K} = 4$	$\mathcal{K} = 5$
$\hat{\mathcal{K}} = 0$	**1.000**	0.00	0.000	0.000	0.000	0.000
$\hat{\mathcal{K}} = 1$	0.000	**0.977**	0.000	0.000	0.000	0.000
$\hat{\mathcal{K}} = 2$	0.000	0.022	**0.996**	0.000	0.000	0.000
$\hat{\mathcal{K}} = 3$	0.000	0.000	0.004	**0.987**	0.000	0.000
$\hat{\mathcal{K}} = 4$	0.000	0.000	0.000	0.0134	**0.999**	0.011
$\hat{\mathcal{K}} = 5$	0.000	0.000	0.00	0.000	0.001	**0.989**

$\sigma_\lambda = \mathbf{0.5}$	$\mathcal{K} = 0$	$\mathcal{K} = 1$	$\mathcal{K} = 2$	$\mathcal{K} = 3$	$\mathcal{K} = 4$	$\mathcal{K} = 5$
$\hat{\mathcal{K}} = 0$	**1.000**	0.00	0.000	0.000	0.000	0.000
$\hat{\mathcal{K}} = 1$	0.000	**0.933**	0.000	0.000	0.000	0.000
$\hat{\mathcal{K}} = 2$	0.000	0.067	**0.998**	0.000	0.054	0.022
$\hat{\mathcal{K}} = 3$	0.000	0.000	0.002	**0.997**	0.080	0.343
$\hat{\mathcal{K}} = 4$	0.000	0.000	0.000	0.003	**0.867**	0.113
$\hat{\mathcal{K}} = 5$	0.000	0.000	0.00	0.000	0.000	**0.522**

The resulting posterior distributions for $\mathcal{K}$ are shown in Table 1, averaged over the 30 replications for each cell. When $\sigma_\lambda = 0.2$ the correct number of changepoints, is generally recovered correctly with high posterior probability; the true number of changepoints always has average posterior probability > 0.95. When $\sigma_\lambda = 0.5$ the true number of changepoints has the highest average posterior probability for all cases, but is frequently underestimated when there are 4 or 5 changepoints.

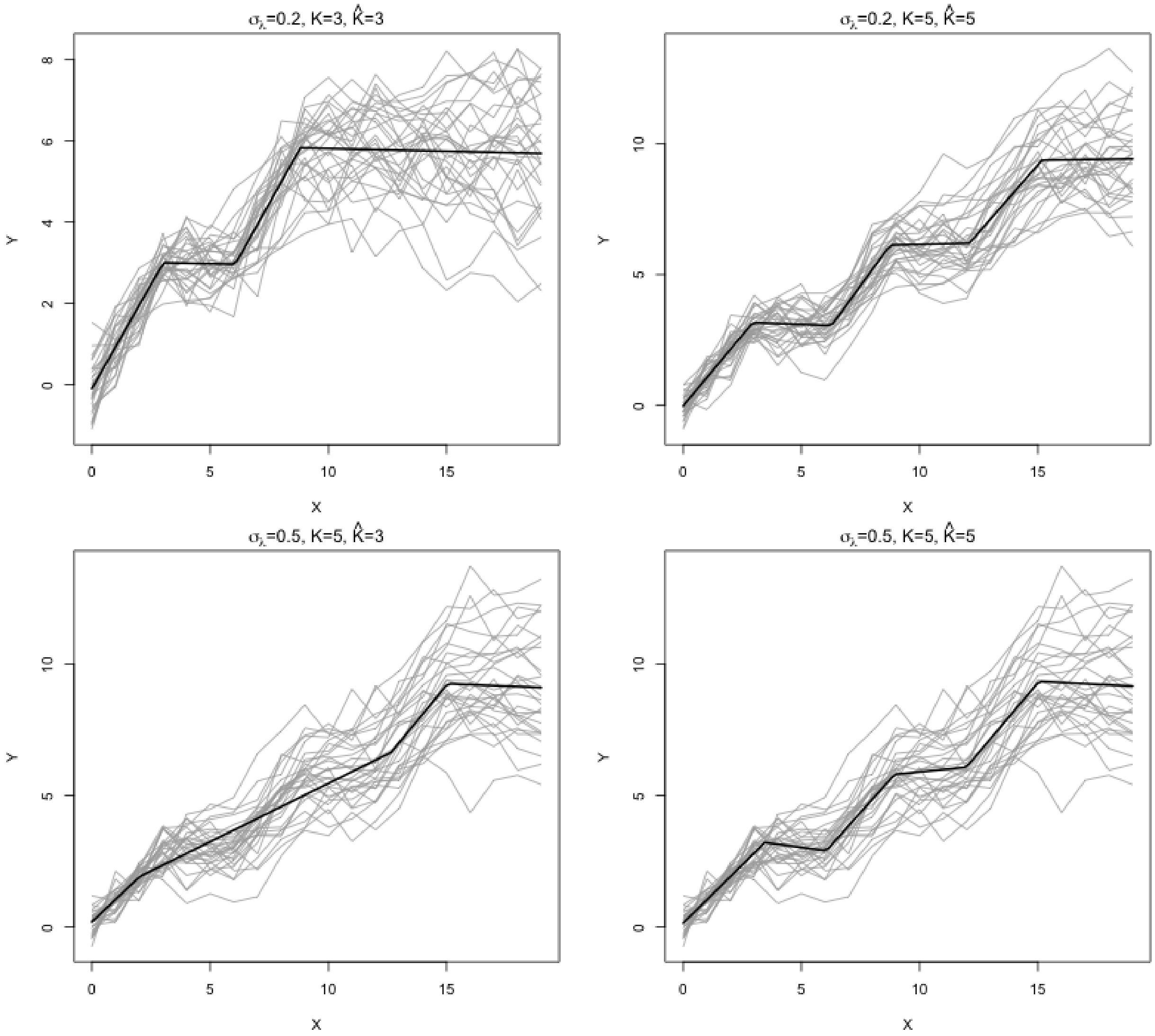

Figure 1.

Spaghetti plot of individual trajectories (gray) and posterior estimates (black) for select simulations.

The lack of precision in detecting several changepoints when $\sigma_\lambda = 0.5$, evident in Table 1, is because the higher variance of the changepoints makes them more difficult to distinguish. This is apparent in Figure 1, which illustrates the data and resulting mean model fit for different situations. The bottom two panels show two different models for the same dataset with 5 changepoints and $\sigma_\lambda = 0.5$. The bottom left panel shows the posterior mean when the number of changepoints is under specified as $\hat{\mathcal{K}} = 3$; this under specified model has substantial posterior probability, illustrating the difficulty in detecting the true number of changepoints. The posterior mean for the same dataset under the correct number of changepoints ($\hat{\mathcal{K}} = 5$) is shown in the bottom right panel.

Table 2 shows the estimated mean changepoint locations, when the number of changepoints are correctly specified, for each of the 10 simulation scenarios that involve at least one changepoint ($\mathcal{K} = 1, 2, 3, 4, 5$; $\sigma_\lambda = 0.2, 0.5$). For each scenario the recovery of the changepoints are generally accurate. The accuracy in estimating each changepoint does not suffer greatly as the number of changepoints increase.

To assess the effect of prior specification on the posterior number of changepoints, we repeat the entire simulation above with alternative priors $\mathcal{K} \sim \text{Binomial}(5, 0.25)$ or $\mathcal{K} \sim \text{Binomial}(5, 0.75)$. The results are summarized in Table 3, which shows the average posterior probability across simulation scenarios under-estimating the true number of changepoints $\mathcal{K}$ ($\hat{\mathcal{K}} < \mathcal{K}$), correctly estimating $\mathcal{K}$ ($\hat{\mathcal{K}} = \mathcal{K}$), and over-estimating $\mathcal{K}$ ($\hat{\mathcal{K}} > \mathcal{K}$), for the different prior specifications. The correct number of changepoints has the highest posterior probability in all scenarios, but as expected smaller values of the prior binomial probability p tend to bias the results toward under-estimation, and larger values bias the result toward over-estimation. Thus, for a more conservative prior that will avoid over-detecting changepoints, one can use a binomial prior with a small probability hyper-parameter.

Table 2.

Mean changepoint locations (and standard deviation of replications) for the $\mathcal{K} = 1, 2, 3, 4,$ and 5 changepoint scenarios, for posterior draws where the number of changepoints are correctly detected.

$\sigma_\lambda = \mathbf{0.2}$	$\mathcal{K} = 1$	$\mathcal{K} = 2$	$\mathcal{K} = 3$	$\mathcal{K} = 4$	$\mathcal{K} = 5$
$\lambda_1 = 3$	2.96 (0.14)	2.99 (0.16)	3.01 (0.14)	3.03 (0.18)	2.98 (0.18)
$\lambda_2 = 6$		5.94 (0.12)	5.97 (0.17)	6.01 (0.16)	5.99 (0.17)
$\lambda_3 = 9$			8.98 (0.15)	8.99 (0.17)	8.99 (0.20)
$\lambda_4 = 12$				11.99 (0.18)	12.00 (0.20)
$\lambda_5 = 15$					15.00 (0.16)

$\sigma_\lambda = \mathbf{0.5}$	$\mathcal{K} = 1$	$\mathcal{K} = 2$	$\mathcal{K} = 3$	$\mathcal{K} = 4$	$\mathcal{K} = 5$
$\lambda_1 = 3$	3.01 (0.11)	2.99 (0.20)	2.97 (0.20)	3.02 (0.18)	2.92 (0.22)
$\lambda_2 = 6$		6.03 (0.16)	5.92 (0.21)	6.02 (0.23)	5.99 (0.21)
$\lambda_3 = 9$			9.07 (0.21)	8.95 (0.21)	8.98 (0.22)
$\lambda_4 = 12$				12.03 (0.18)	12.03 (0.21)
$\lambda_5 = 15$					14.96 (0.27)

Table 3.

Average posterior probability across simulation scenarios under-estimating the true number of changepoints $\mathcal{K}$ ($\hat{\mathcal{K}} < \mathcal{K}$), correctly estimating $\mathcal{K}$ ($\hat{\mathcal{K}} = \mathcal{K}$), and over-estimating $\mathcal{K}$ ($\hat{\mathcal{K}} > \mathcal{K}$), for different prior specifications.

$\sigma_\lambda = \mathbf{0.2}$	Under-estimation	Correct estimation	Over-estimation
$\mathcal{K} \sim \text{Binom}(5, \mathbf{0.25})$	0.367	0.633	0.001
$\mathcal{K} \sim \text{Binom}(5, \mathbf{0.5})$	0.002	0.991	0.007
$\mathcal{K} \sim \text{Binom}(5, \mathbf{0.75})$	0.006	0.883	0.111

$\sigma_\lambda = \mathbf{0.5}$	Under-estimation	Correct estimation	Over-estimation
$\mathcal{K} \sim \text{Binom}(5, \mathbf{0.25})$	0.441	0.559	0.000
$\mathcal{K} \sim \text{Binom}(5, \mathbf{0.5})$	0.10	0.886	0.012
$\mathcal{K} \sim \text{Binom}(5, \mathbf{0.75})$	0.027	0.877	0.095

2. Multi-class illustration

Here we describe a simple example to illustrate the clustering properties and simultaneous changepoint detection of the `BayesianPGMM` package. This section also serves as a brief tutorial, with commands to reproduce the results below after the package is installed and loaded to the R workspace.

We generate the data shown in Figure 2. These data can be loaded in R via the command `data(SimData4classes)` after the package is installed, and can be visualized as shown using the command `plotPGMM(X,Y)`. These data consist of four latent classes, each with ten individuals with measurements for the same 10 time points. Each latent class has a different number of changepoints 0, 1, 2, or 3.

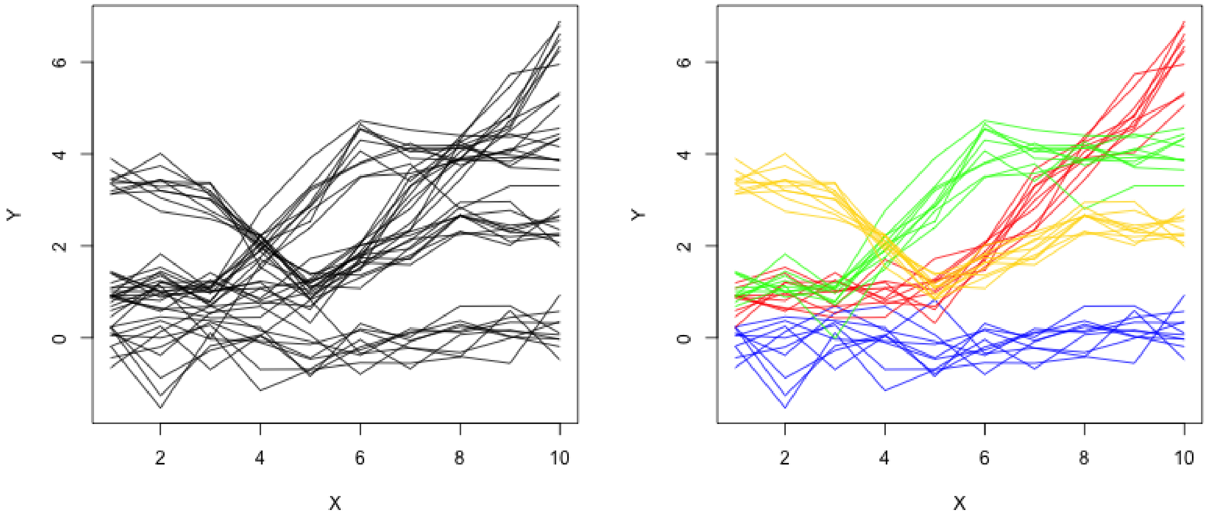

Figure 2.

Spaghetti plot of generated data without showing classes (left) and colored by latent classes (right). The blue class has 0 changepoints, the red has 1 changepoint, the green has 2 changepoints, and the gold has 3 changepoints.

We estimate the posterior model with four latent classes, and up to 3 changepoints in each class, using the command

```
Fit <- BayesPGMM(X,Y,max_cp=3,n_clust=4) .
```

The resulting class clustering and mean fits can be visualized using the command `plotPGMM(X,Y,Fit)`, as shown in the top left panel of Figure 3. The resulting clustering matches the true latent classes, and the correct number of changepoints are detected for each class.

To illustrate robustness to over-specification of the number of classes we also fit the model with the same specification (four classes, 3 potential changepoints) to a reduced dataset with one latent class removed. Specifically, we remove the 10 individuals belonging to the fourth class, leaving three latent classes with 0, 1 and 2 changepoints. Thus, we use a four class model to estimate data with three classes:

```
Fit <- BayesPGMM(X[1:30,],Y[1:30,],max_cp=3,n_clust=4) .
```

The results can again be visualized using `plotPGMM(X[1:30,],Y[1:30,],Fit)`, as in the top right panel of Figure 3. The latent classes and number of changepoints in each class are again recovered correctly. In particular, only three of the possible four latent classes are represented, leaving the extraneous fourth class empty. We similarly fit the model with four classes and 3 potential changepoints to data with only two of the classes (with 0 and 1 changepoints), and to data with only one class (with 0 changepoints). The results, shown in the bottom two panels of Figure 3, again recover the true clustering and number of changepoints, leaving extraneous classes empty.

For each of the four simulated datasets above, the recovery of the clustering and true number of changepoints were validated with 10 independent replications.

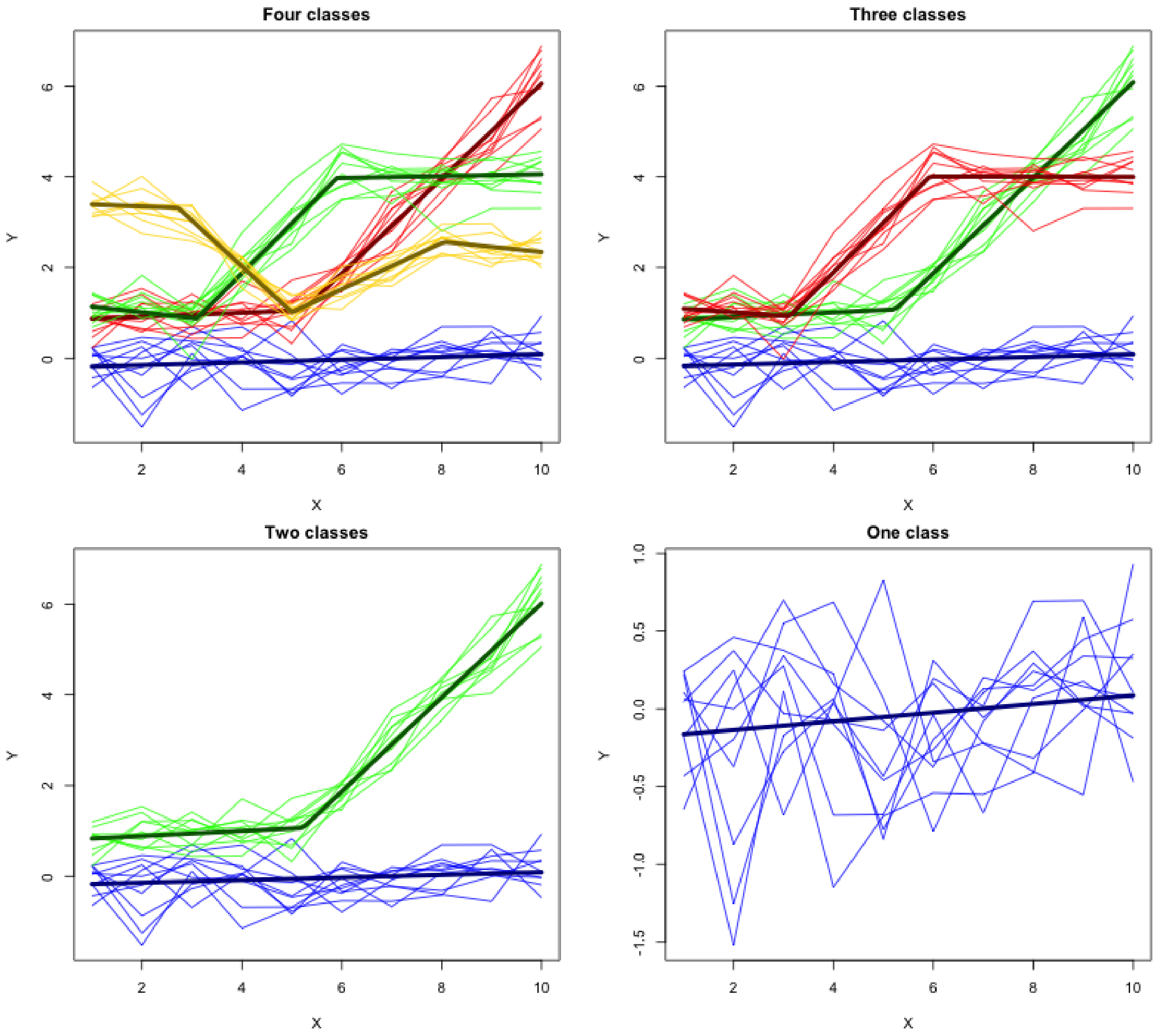

Figure 3.

Spaghetti plot of the simulated data with different number of latent classes present, with colors showing the estimated class clustering. The trajectory defined by the mean parameters for each class are shown in **bold**.

3. Clustering prior simulation

Here we describe a simulation to illustrate the effect of the concentration parameters for the Dirichlet clustering prior on the posterior. It is common to set each value of the C-dimensional concentration parameter, where C is the number of clusters, to a constant α: Dirichlet(α, ..., α). Smaller values of α suggest less parity in the class sizes (e.g., one class is much larger than the other), while larger values of α suggest more parity in the class sizes. To illustrate, we consider a two-class model, for which the Dirichlet(α, α) distribution is equivalent to a Beta(α, α) distribution for the proportion of one class. By default we use $\alpha = 1$, which is equivalent to a Uniform$(0, 1)$ distribution; more generally, a Dirichlet$(1, \ldots, 1)$ distribution is uniform over the unit simplex.

Herein in addition to the two-class model with $\alpha = 1$, we consider Dirichlet priors with $\alpha = 0.25$, $\alpha = 0.5$, $\alpha = 1$, $\alpha = 2$, $\alpha = 4$, and $\alpha = 8$. The resulting prior distributions for a single class probability ν_1 ($\nu_2 = 1 - \nu_1$) are shown in Figure 4. Note that $\alpha = 0.5$ corresponds to a Jeffrey's prior (Jeffreys, 1946).

We simulate an additional 100 realizations of the simulation scheme in Section 5 of the main manuscript, under the application-motivated scenario with $N = 60$, $M_i = 50$, $\nu_1 = 0.80$, and $\mathcal{K}_2 = 2$. We compute the posterior for each of $\alpha = \{0.25, 0.5, 2, 4, 8\}$ for 20 realizations, with otherwise the same settings as those used in the main simulation with $\alpha = 1$. The average of the posterior means for the latent proportion of the smaller class $\nu_2 = 0.2$ is shown for each value of α in Table 1. The estimated latent proportion tends to increase above 0.2 for higher values of α; this is expected, as higher values of α tends to bias estimates toward equal class proportions. However, this appears to have little affect on the overall accuracy of the posterior: the misallocation rate of the latent class memberships is not substantially affected (Table 4), and the posterior accuracy of other model parameters are also not substantially affected (Table 5).

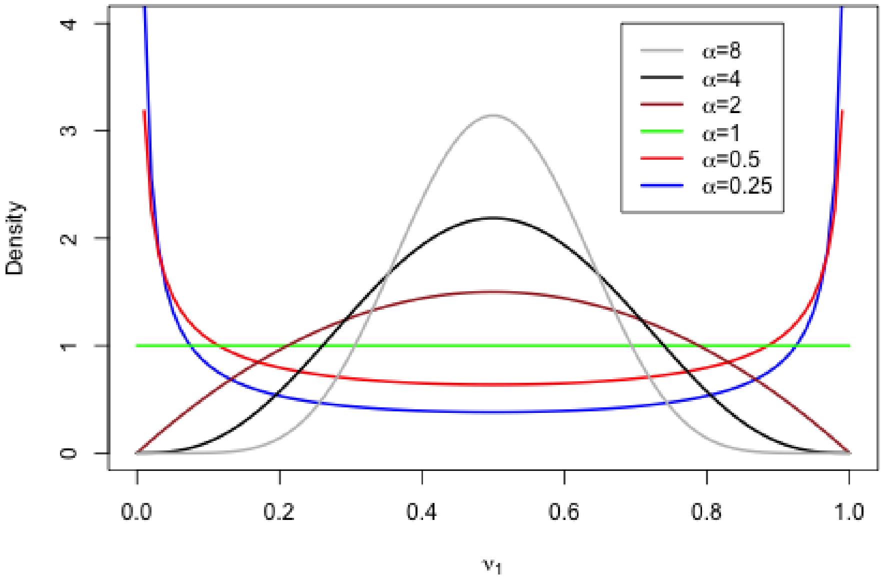

Figure 4.

Prior density of ν_1 for different Dirichlet(α, α) distributions.

Table 4.

Mean class 2 proportion ν_2 ($\nu_2 = 0.2$), and mean class misallocation rate, for Dirichlet concentration parameter α.

	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$	$\alpha = 4$	$\alpha = 8$
$\hat{\nu}_2$	0.18	0.19	0.21	0.26	0.26	0.31
Misallocation	0.15	0.13	0.12	0.10	0.12	0.12

4. Variance prior simulation

Here we describe a simulation in which we consider alternative priors for the variance (or standard deviation) of the random effects. By default we have used a uniform prior for the

Table 5.

Summary of mean parameter estimates in Class 1 and 2 for concentration parameter $\alpha < 1$ ($\alpha = 0.25$ or $\alpha = 0.5$), $\alpha = 1$, or $\alpha > 1$ ($\alpha = 2$ or $\alpha = 4$ or $\alpha = 8$).

	Class 1	$\alpha < 1$	$\alpha = 1$	$\alpha > 1$	**Class 2**	$\alpha < 1$	$\alpha = 1$	$\alpha > 1$
σ_ϵ	**3.16**	3.17	3.17	3.18	**3.16**	3.17	3.17	3.18
β_1	**-0.002**	-0.003	-0.003	-0.003	**-0.005**	-0.0009	-0.0008	-0.0008
β_2	**0.194**	0.185	0.186	0.192	**0.060**	0.068	0.066	0.062
β_3	**-0.171**	-0.153	-0.154	-0.164	**0.081**	0.0217	0.0325	0.0434
σ_{β_1}	**0.010**	0.010	0.010	0.010	**0.008**	0.033	0.025	0.016
σ_{β_2}	**0.064**	0.058	0.057	0.053	**0.027**	0.062	0.059	0.054
σ_{β_3}	**0.079**	0.089	0.087	0.081	**0.068**	0.105	0.099	0.090
λ_1	**362**	362	363	363	**321**	328	327	328
σ_{λ_1}	**93.6**	98.7	98.2	95.1	**132**	136	143	151
λ_2	**643**	650	650	645	**726**	719	708	708
σ_{λ_2}	**149**	147	146	144	**128**	140	147	161

standard deviation, with a lower bound of 0 and an upper bound that depends on the context of the parameter (see Section 3.1 of the main manuscript). An alternative prior for the standard deviation is the half-Cauchy prior (Polson et al., 2012), which is a Cauchy distribution truncated above 0:

$$p(x \mid \gamma) = \frac{2}{\pi\gamma\,(1 + (x/\gamma)^2)} \quad \text{for } x > 0,$$

where γ is a scale parameter. We implement the half-Cauchy prior for all random-effects parameters $\left(\{\sigma^2_{c,\beta_k}\}, \{\sigma^2_{c,\lambda_k}\}\right)$, and under two different strategies to select γ,

1. **Scaled**, in which γ depends on the parameter. Here, γ is selected such that the 90th percentile of the resulting half-Cauchy distribution is given by the upper bound used for the default uniform distribution. For example, under the default uniform prior $\sigma_{c,\lambda_1} \overset{iid}{\sim} \text{Uniform}(0, b)$ where $b = \frac{\max(X)-\min(X)}{4}$, while under the scaled Cauchy prior $P(\sigma_{c,\lambda_1} < b) = 0.9$.
2. **Unscaled**, in which $\gamma = 25$ for all parameters; this is suggested as the default half-Cauchy prior for a scale parameter in the `laplacesDemon` R package (Statisticat, 2015).

As another alternative, we consider an $IG(0.001, 0.001)$ distribution for the variances of the random effects.

We repeat 100 simulations from Section 5 of the main manuscript, under the application-motivated scenario with $N = 60$, $M_i = 50$, $\nu_1 = 0.80$, and $\mathcal{K}_2 = 2$. For each replication we consider, in addition to the default uniform prior, a the scaled half-cauchy prior, unscaled half-cauchy prior, and inverse-gamma prior for the random effects.

The diffuse inverse-gamma prior (here $\text{IG}(0.001, 0.001)$) is generally not recommended for modeling the variance of hierarchical random effects (Gelman et al., 2006), partly because its density is unstable as the variance approaches 0. This is especially worrisome when the number of random effects are small, or in the context of mixture models, where the number of observations within a class may be small and vary during posterior sampling. Indeed, our implementation of $\text{IG}(0.001, 0.001)$ priors failed during posterior sampling for each replication, because of numerical errors caused by extreme values.

The resulting average parameter estimates for the uniform, scaled half-Cauchy, and unscaled half-Cauchy priors are shown in Table 6. The results for the scaled half-Cauchy priors are mostly comparable to the results for the default uniform priors, although mean estimates for the standard deviations for the changepoint locations in Class 2 (the smaller class) are inflated. For the unscaled half-Cauchy priors the posterior standard deviations for the random coefficients in Class 2 are highly inflated, and other parameter estimates for Class 2 are generally less accurate. These results demonstrate that appropriate scaling of the prior for hierarchical random effects is important, especially for the accurate identification of latent classes that have a small number of individuals.

In the `BayesianPGMM` package, we have implemented the scaled half-Cauchy prior as an option, in addition to the default uniform prior. The half-Cauchy has the advantage of not having a hard constraint (e.g., as in the uniform upper bound) and facilitating some shrinkage; however, the half-Cauchy can give non-trivial probability to unreasonably large standard deviations in the right tail of the distribution, and the uniform prior has the advantage of being simple to interpret.

References

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London a: mathematical, physical and engineering sciences*, volume 186, pages 453–461. The Royal Society.

Table 6.

Summary of mean parameter estimates in Class 1 and 2 with different prior choices for the model random effects, including the default uniform priors, scaled half-Cauchy and unscaled half-Cauchy (HC(0, 25)).

	Class 1	Uniform	Scaled HC	$\mathrm{HC}(0,25)$	**Class 2**	Uniform	Scaled HC	$\mathrm{HC}(0,25)$
σ_ϵ	**3.16**	3.17	3.18	3.19	**3.16**	3.17	3.18	3.19
β_1	**-0.002**	-0.003	-0.003	-0.003	**-0.005**	-0.0008	-0.002	-0.001
β_2	**0.194**	0.186	0.190	0.178	**0.060**	0.066	0.070	0.045
β_3	**-0.171**	-0.154	-0.159	-0.142	**0.081**	0.0325	0.034	0.028
σ_{β_1}	**0.010**	0.010	0.010	0.010	**0.008**	0.025	0.017	5.12
σ_{β_2}	**0.064**	0.057	0.054	0.065	**0.027**	0.059	0.027	8.61
σ_{β_3}	**0.079**	0.087	0.082	0.103	**0.068**	0.099	0.081	8.51
λ_1	**362**	363	362	362	**321**	327	335	337
σ_{λ_1}	**93.6**	98.2	95.1	97.1	**132**	143	263	176
λ_2	**643**	650	646	654	**726**	708	711	710
σ_{λ_2}	**149**	146	142	142	**128**	147	235	145

Polson, N. G., Scott, J. G., et al. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

Statisticat, L. (2015). LaplacesDemon: complete environment for Bayesian inference. *R package version*, 15(01).