Modeling Context-Dependent Latent Effect Heterogeneity (Online Supplementary Material)

Diogo Ferrari

December 6, 2018

Contents

1 Riemann Manifold Hamiltonian Monte Carlo

The proposed model from which we derive the Riemman Manifold Hamiltonian Monte Carlo (RMHMC) within Gibbs is defined as follows

$$
\begin{aligned}
& V_l \mid \alpha_o \sim \text{Beta}(1, \alpha_o) \\
& \pi_k = \begin{cases} V_1 & , k = 1 \\ V_k \prod_{l=1}^{k-1} (1 - V_l) & , k > 1 \end{cases} \\
& Z_i \mid \pi \sim \text{Cat}(\pi) && , \pi \in \Delta^{\infty} \\
& \tau_d \sim p(\tau_d) && , d = 1, \dots, D_x + 1 \\
& \theta_{kj} \mid Z_{ik}, \tau, C_{ij}, W \sim p(\theta_{jk} \mid W, \tau) && , j = 1, \dots, J \\
& y_i \mid Z_{ik}, \theta_{kj}, X_i, C_{ij} \sim p(y_i \mid Z_{ik}, X_i, \theta_{kj}) && \ni \quad \mathbb{E}[y_i \mid Z_{ik}, \theta_{kj}, X_i, C_{ij}] = g^{-1}(X_i^T \theta_{kj}) \\
& && \quad\;\; p(y_i \mid Z_{ik}, X_i, \theta_{kj}) \text{ from exponential family}
\end{aligned}
\tag{1}
$$

As discussed in the main paper, when the outcome variable y_i in the model (1) is binomial or multinomial distributed the Gibbs sampler developed in the paper cannot be used anymore for the parameters θ (or β). We use a RMHMC update within Gibbs to sample the β coefficients in these cases. The random variable of interest is $\beta_{kj} \in \mathbb{R}^{D+x+1}$ and we use $v \in \mathbb{R}^{D_x+1}$ as the ancillary variable (momentum) such that $v \sim N_{D_x+1}(0, G(\beta_{kj}))$. The Hamiltonian is defined by

$$H(\beta_{kj}, v) = U(\beta_{kj}, v) + K(\beta_{kj}, v) = -(\beta_{kj} \mid \cdot) + \frac{D_x+1}{2} \ln(2\pi) + \frac{1}{2}\left[\ln\left(\det[G(\beta_{kj})]\right) + v^T G(\beta_{kj})^{-1} v\right] \tag{2}$$

whose solution is

$$\begin{aligned} \nabla_v H(\beta_{kj}, v) &= G(\beta_{kj})^{-1} v \\ \nabla_{\beta_{kj}} H(\beta_{kj}, v) &= -\left[\nabla_{\beta_{kj}} U(\beta_{kj}, v) - \frac{1}{2} \mathrm{tr}\left\{G(\beta_{kj})^{-1} \nabla_{\beta_{kj}} G(\beta_{kj})\right\}\right. \\ &\qquad \left. + \frac{1}{2}\left(v^T G(\beta_{kj})^{-1} G(\beta_{kj})^{-1} v\right) \nabla_{\beta_{kj}} G(\beta_{kj})\right] \end{aligned} \tag{3}$$

The Hamiltonian equations are solved using the generalized Stormer-Verlet leapfrog integrator (Calin and Chang, 2006; Girolami and Calderhead, 2011). So for L leapfrog steps with size ε, and $l = 1, \ldots L$, we have

$$\begin{aligned} v^{l+\frac{\varepsilon}{2}} &= v^l - \frac{\varepsilon}{2} \nabla_{\beta_{kj}} H\left(\beta_{kj}^l, v^{l+\frac{\varepsilon}{2}}\right) \\ \beta_{kj}^{l+\varepsilon} &= \beta_{kj}^l + \frac{\varepsilon}{2}\left[\nabla_v H\left(\beta_{kj}^l, v^{l+\frac{\varepsilon}{2}}\right) + \nabla_v H\left(\beta_{kj}^{l+\varepsilon}, v^{l+\frac{\varepsilon}{2}}\right)\right] \\ v^{l+\varepsilon} &= v^{l+\frac{\varepsilon}{2}} - \frac{\varepsilon}{2} \nabla_{\beta_{kj}} H\left(\beta_{kj}^{l+\varepsilon}, v^{l+\frac{\varepsilon}{2}}\right) \end{aligned} \tag{4}$$

When y_i is binomial, that is, the distribution of y_i is defined by

$$y_i \sim \mathrm{Bin}(p_{kj}) \quad , \quad p_{kj} = \frac{1}{1 + e^{-X_i^T \beta_{kj}}}$$

then the elements of the RMHMC for the model when $k \in Z_j^*$ are defined by the following equations:

$$\begin{aligned} U(\beta_{kj}) = -\ln p(\beta_{kj} \mid \cdot) \propto -\Bigg[&-\frac{D_x+1}{2} \ln 2\pi - \frac{1}{2} \ln(\det(\Sigma_\beta)) - \frac{1}{2}(\beta_{kj} - (W_j^T \tau)^T)^T \Sigma_\beta^{-1} (\beta_{kj} - (W_j^T \tau)^T) \\ &- \sum_{i \in I_k} y_i \ln\left(1 + e^{-X_i^T \beta_{kj}}\right) - \sum_{i \in I_k} (1 - y_i) \ln\left(1 + e^{X_i^T \beta_{kj}}\right)\Bigg] \\ \nabla_{\beta_{kj}} U(\beta_{kj}) = -\Bigg[&-(\beta_{kj} - (W_j^T \tau)^T)^T \Sigma_\beta^{-1} + \sum_{i \in I_k} X_i y_i p(y_i = 0 \mid \cdot) - \sum_{i \in i_k} X_i (1 - y_i) p(y_i = 1 \mid \cdot)\Bigg] \end{aligned}$$

In practice we use $G(\beta_{kj}) = I_{(D_x+1)\times(D_x+1)}$, which is the most widely used approach in applications (Neal et al., 2011; Liu, 2008). It also simplify the equation (2), (3), and (4). Using $v \sim N_{D_x+1}(0, I)$, the integrator reduces to the standard Stormer-Verlet leapfrog integrator (Duane et al., 1987; Neal et al.,

2011) and we have the following equations for the hierarchical Dirichlet Process Generalized Linear Model (hdpGLM):

$$
\begin{aligned}
\nabla_{\beta_{kj}} H(\beta_{kj}, v) &= \nabla_{\beta_{kj}} U(\beta_{kj}) - v \\
v^{l+\frac{\varepsilon}{2}} &= v^{l} - \frac{\varepsilon}{2} \nabla_{\beta_{kj}} U(\beta_{kj}) \\
\beta_{kj}^{l+\varepsilon} &= \beta_{kj}^{l} + \varepsilon v^{l+\frac{\varepsilon}{2}} \\
v^{l+\varepsilon} &= v^{l+\frac{\varepsilon}{2}} - \frac{\varepsilon}{2} \nabla_{\beta_{kj}} U(\beta_{kj})
\end{aligned}
$$

With these definitions, the RMHMC is presented in the algorithm 1

Algorithm 1 Riemman Manifold Hamiltonian Monte Carlo

Require: $Z^{(t)}, \beta^{(t)}, \pi^{(t)}, \tau^{(t)}$

for $j = 1, ..., J$ and for $k \in Z_j^*$ **do**
  sample $v^{\text{current}} \sim N_{D_x+1}(0, G(\beta_{kj}))$
  Let $v \leftarrow v^{\text{current}}, \beta_{kj} \leftarrow \beta_{kj}^{(t)}$
  Set $v \leftarrow v - \frac{\varepsilon}{2} \nabla_{\beta_{kj}} U(\beta_{kj})$
  for $l = L - 1$ **do**
    $\beta_{kj} \leftarrow \beta_{kj} + \varepsilon \left(G^{-1}(\beta_{kj}) v \right)$
    $v \leftarrow v - \frac{\varepsilon}{2} \nabla_{\beta_{kj}} U(\beta_{kj})$
  end for
  Set $v \leftarrow v - \frac{\varepsilon}{2} \nabla_{\beta_{kj}} U(\beta_{kj})$
  Set $v = -v$
  sample $u \sim U(0, 1)$
  if $u < \min \left\{ 1, \exp \left\{ -H(\beta_{kj}, v) + H(\beta_{kj}^{(t)}, v^{\text{current}}) \right\} \right\}$ **then**
    $\beta_{kj}^{(t+1)} \leftarrow \beta_{kj}$
  else
    $\beta_{kj}^{(t+1)} \leftarrow \beta_{kj}^{(t)}$
  end if
end for

2 Prior Perturbation Analysis

This section evaluates the effect of prior perturbations on three quantities: "bias" (the distance between the estimated posterior expectation and the true), the size of the 95% HPD intervals, and the estimated number of clusters.

As mentioned in the main paper, the MC exercises were conducted using the following prior parameters: $(\mu_{\tau_d}, \sigma_{\tau_d} I, \sigma_{\beta_{kj}} I, s^2, \nu, \alpha_0) = (0, 10I, 10I, 10, 10, 1)$, where I represents the identity matrix. The choice of those values was based on experimentation. To evaluate the effect of prior perturbation, I estimated the model 1215 times, each time using a different combination of the prior parameter whose

Table 1: Sensitivity of estimated bias ($\beta - \hat{\mathbb{E}}[\beta \mid \cdot]$) to prior specification: worst case versus priors used in the main MC study

Sample size	Priors: (α, s^2, ν, σ_β)	Absolute Maximum Estimated Bias	Average Estimated Bias
Worst Case (average computed across priors and linear coefficients)			
500	(5,1,4,5)	9.2910	0.0665
1000	(0.5,1,10,5)	8.5466	-0.1051
2000	(0.5,4,1,10)	8.3330	-0.2286
5000	(0.5,4,10,5)	7.7365	-0.0364
10000	(5,4,4,10)	8.0995	0.0255
Values used in the main MC study (average computed across linear coefficients)			
500	(1,10,10,10)	0.1531	0.0115
1000	(1,10,10,10)	0.1647	-0.0520
2000	(1,10,10,10)	0.0800	0.0053
5000	(1,10,10,10)	0.0657	-0.0020
10000	(1,10,10,10)	0.2383	0.0043

values were selected from the set described in the following table:

Prior parameters	Set of values
α	$\{.5, 1, 5\}$
s^2	$\{1, 4, 10\}$
ν	$\{1, 4, 10\}$
σ_τ	$\{1, 5, 10\}$
σ_β	$\{1, 5, 10\}$
n (sample size)	$\{500, 1000, 2000, 5000, 10000\}$

I used the same setting with three clusters and three covariates for all estimations under different prior combinations to properly compare the effect of prior perturbation. The Gibbs sampler ran for 17000 iterations, and the last 7000 were recorded.

The results are summarized in Tables 1 to 3 and Figure 1. Table 1 shows summaries of how the "bias" is affected by the prior specification. For each sample size n, the upper half of the table shows the prior setting containing the linear coefficient β that produced the maximum absolute value of the bias. The last column of the upper half shows the average bias computed across prior combinations and linear coefficients. The bottom half of the table shows the same two measures - the maximum absolute value of the bias and the average bias - for each sample size, but considering only the benchmark with prior parameter as used in the MC exercises. Table 2 displays similar information but with consequences of the prior selection on the 95% HPD intervals. Table 3 compares the estimated number of clusters in the worst case (more distance from the true) against the priors used in the MC study. Finally, Figure 1 shows the number of clusters, the bias, and the sizes of the 95% HPD intervals for all prior combinations and sample sizes. In the upper half of the figure, the dots represent the number of clusters activated in each estimation. In the bottom-left, the dots represent the bias of each linear coefficient β, and in the bottom right, the dots represent the size of the 95% HPD interval, also for each linear coefficient. Red lines in the figure show the average values.

We can see from the tables and the figure that the estimation is not very sensitive on average to the choice of those prior parameters in the range considered here, but for certain combinations, in the worst

Table 2: Sensitivity of the size of the 95% HPD interval to prior specification: worst case versus priors used in the main MC study

Sample size	Priors: (α, s^2, ν, σ_β)	Maximum size of the 95% HPD intervals	Mean size of the 95% HPD intervals
Worst Case (average computed across priors and linear coefficients)			
500	(0.5,1,1,10)	15.7945	3.7669
1000	(0.5,1,1,10)	12.9130	3.2080
2000	(5,1,1,1)	14.6177	3.4737
5000	(1,1,1,5)	13.0680	2.5827
10000	(5,1,1,5)	14.7897	2.0746
Values used in the main MC study (average computed across linear coefficients)			
500	(1,10,10,10)	0.4253	0.3785
1000	(1,10,10,10)	0.2747	0.2509
2000	(1,10,10,10)	0.1724	0.1651
5000	(1,10,10,10)	0.1159	0.1087
10000	(1,10,10,10)	1.5535	0.8132

Table 3: Sensitivity of estimated number of clusters to prior specification: worst case versus priors used in the main MC study

Sample size	Priors: (α, s^2, ν, σ_β)	K estimated	Mean across different priors	Percentage equal to true across different priors
Worst Case				
500	(5,1,1,10)	21	4.740741	66.67 %
1000	(5,1,1,5)	20	5.654321	66.67 %
2000	(5,1,1,10)	43	5.530864	72.84 %
5000	(1,1,1,5)	60	10.246914	55.56 %
10000	(5,1,1,5)	88	12.913580	59.26 %
Values used in the main MC study				
500	(1,10,10,10)	3	–	–
1000	(1,10,10,10)	3	–	–
2000	(1,10,10,10)	3	–	–
5000	(1,10,10,10)	3	–	–
10000	(1,10,10,10)	8	–	–

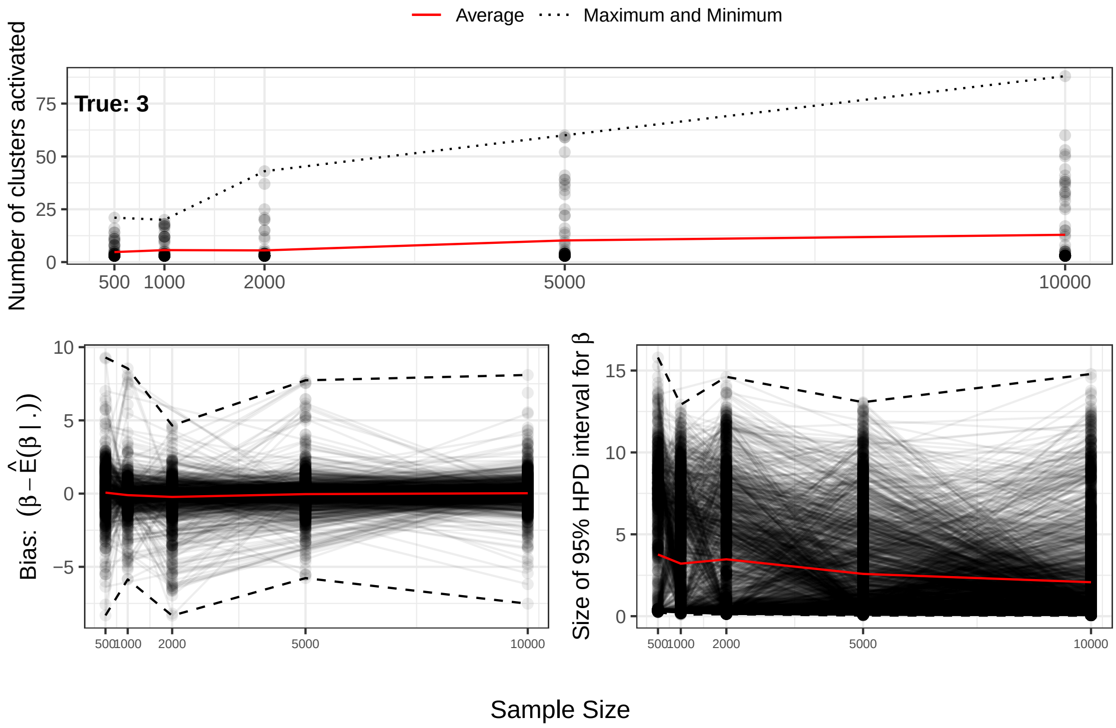

Figure 1: Sensitivity of number of clusters, bias $(\beta - \mathbb{E}[\beta \mid \cdot])$, and 95% HPD interval for different sample sizes and prior parameters

case, the model can demand very large data sets to escape the influence of the prior specification. This is true specially for extreme values of the concentration parameter α and values that produce highly dispersed inverse-scaled-χ^2 distribution, which can be generated by low values (below five) of the scale parameter s^2.

In particular, the estimated number of clusters can be sensitive to α. That is expected and it is a feature of models using Dirichlet Process Prior (DPP). In Dirichlet processes, large values of α tend to produce large number of clusters. Antoniak (1974) has shown that for any sample size n, $\mathbb{E}[K \mid n, \alpha] = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}$, where K denotes a random variable that captures the number of clusters in n observations from a Dirichlet process model with concentration parameter α. So it is clear that $\mathbb{E}[K \mid n, \alpha] \to n$ if $\alpha \to \infty$, and $\mathbb{E}[K \mid n, \alpha] \to 0$ it $\alpha \to 0$. Table 4 shows $\mathbb{E}[K \mid n, \alpha]$ for the values of α and n used in this section. The table shows the expectation of K if the DPP was the true data generating process (DGP) and K was a random variable denoting the number of clusters. Note that those are not posterior expectations, but the expectation of the number of clusters when DPP is actually the DGP. When K is finite, fixed, but unknown, the results presented here and in the main paper shows that DPP can be used to approximate the true unknown finite mixture model, that K can be estimated as a result, that on average the estimation produces good approximation, but that in the worst case it can be affected by choice of α. A much larger chain and large data set may be needed for extreme values of α.

In all the simulated data sets and discussions in the paper, K is fixed but unknown, and we approximate it using the hierarchical DPP model. For all estimations I selected a fixed value for the

Table 4: Prior Expectation of the number of clusters for different sample sizes (n) and values of the concentration parameter α

α	Sample size (n)	$E[K \mid n, \alpha]$
0.5	500	4.0891
0.5	1000	4.4356
0.5	2000	4.7822
0.5	5000	5.2404
0.5	10000	5.5869
1.0	500	6.7928
1.0	1000	7.4855
1.0	2000	8.1784
1.0	5000	9.0945
1.0	10000	9.7876
5.0	500	23.5873
5.0	1000	27.0306
5.0	2000	30.4852
5.0	5000	35.0599
5.0	10000	38.5234

concentration parameter: $\alpha = 1$. Using a fixed value for α reduces the computational cost of the Monte Carlo study substantially, and it also supported the derivation of the full Gibbs as presented in the paper. The results in the paper and this section indicate that $\alpha = 1$ is a sensible choice. The benchmark values in the MC exercises have shown good results in a variety of data sets and parameter values. It is important to keep in mind, however, that such choice may not be appropriate for all data sets. If the chain is taken too long to display good convergence diagnostics or multiple estimations seems to be generating different results, practitioners can adjust α and the other fixed parameters to reflect their prior beliefs about α. It may be advisable also centering and scaling the data to estimate effects of covariates in terms of standard deviation of covariates (see Rossi (2014) for similar recommendations).

The model can be extended to incorporate prior uncertainty on α. Some authors have proposed models with Gamma priors on α (West, Muller and Escobar, 1994; Mukhopadhyay and Gelfand, 1997; Gelfand, Kottas and MacEachern, 2005). The difficulty then becomes to select priors for that distribution, which can impact on the posterior distribution of α. In those approaches, learning about that parameter can be difficult, and it may require large data sets. For that reason, Dorazio (2009) proposes a prior selection procedure based on minimizing a Kullback-Leibler divergence computed with the prior distribution of the number of clusters K induced by the prior on α and a ancillary distribution on $K = 1, \ldots, n$ that can capture our uncertainty about K. The model can be extended in those directions, although it will increase the computational cost for the estimation.

3 Details of the Simulation Study

This section presents some details of the simulation study, the extension of the exercise from 100 to 1000 replications, and the tables with the MC error. Details of the prior parametrization are in the section 2. The MCMC ran until the MCMC error was below 0.05 for all parameters (Flegal, 2008;

Table 5: Summary of Monte Carlo error across replications of each parameter set

Parameter Set	N. Clusters	N. Covariates	MC error			
			Mean	Std. Dev.	Minimum	Maximum
1	7	4	0.1968	0.3862	0.0384	1.6139
2	5	3	0.1286	0.2148	0.0379	0.8725
3	4	3	0.1578	0.2158	0.0389	0.7418
4	10	3	0.1747	0.1958	0.0449	0.8134
5	7	5	0.1089	0.2361	0.0376	1.4820
6	1	0	0.1673	0.1673	0.1673	0.1673
7	4	2	0.3718	0.4727	0.0372	1.5248
8	2	5	0.3882	0.4810	0.0325	1.3367
9	10	2	0.3842	0.2792	0.0534	0.9490
10	3	2	0.2480	0.2413	0.0490	0.8220

Gong and Flegal, 2015)

Table 5 displays the summary statistics of the MC error for the simulation study presented in the main paper. Note that the average value of the MC error is quite small, as well as the maximum value. That is true for all parameter sets. Table 6 shows the estimated posterior expectation, the MC error, and the true value for each β in each cluster and each parameter set. We can see that the MC errors are in general small and the estimated values very close to the true values.

Table 6: True, Monte Carlo error, and posterior expectation of each β per parameter set and cluster

Parameter	Estimate	Cluster									
		1	2	3	4	5	6	7	8	9	10
Parameter Set 1 (N. of Clusters: 7; N. of Covariates: 4)											
β_1	True	2.8461	-0.2693	4.5358	4.9792	-4.3710	7.9719	-2.8122	–	–	–
β_1	MCMC Mean	2.8340	-0.2810	4.2452	4.9691	-4.3731	7.9653	-2.8170	–	–	–
β_1	MCMC error	0.0513	0.0505	1.0963	0.0464	0.0428	0.0452	0.0456	–	–	–
β_2	True	0.4319	1.1724	6.6526	-4.4481	1.8016	8.1086	-9.1892	–	–	–
β_2	MCMC Mean	0.4258	1.1740	6.2130	-4.4441	1.7929	8.1082	-9.1888	–	–	–
β_2	MCMC error	0.0471	0.0384	1.6139	0.0472	0.0502	0.0458	0.0458	–	–	–
β_3	True	-6.4458	9.7919	-3.3883	3.8350	-3.3731	8.4948	-6.7871	–	–	–
β_3	MCMC Mean	-6.4347	9.7833	-3.1728	3.8333	-3.3793	8.4908	-6.7782	–	–	–
β_3	MCMC error	0.0444	0.0473	0.8053	0.0439	0.0481	0.0528	0.0466	–	–	–
β_4	True	-9.4002	4.0393	-4.1365	2.8764	-5.6639	-6.9705	-0.2020	–	–	–
β_4	MCMC Mean	-9.3907	4.0328	-3.8762	2.8833	-5.6537	-6.9739	-0.2108	–	–	–
β_4	MCMC error	0.0525	0.0458	0.9850	0.0436	0.0499	0.0505	0.0479	–	–	–
β_5	True	5.4716	6.6912	-3.8207	-9.5771	9.7805	3.3063	-6.9065	–	–	–
β_5	MCMC Mean	5.4706	6.6822	-3.5633	-9.5764	9.7745	3.3096	-6.9055	–	–	–
β_5	MCMC error	0.0563	0.0461	0.9658	0.0444	0.0495	0.0457	0.0497	–	–	–
Parameter Set 2 (N. of Clusters: 5; N. of Covariates: 3)											
β_1	True	9.1445	-6.0408	-8.7230	2.2769	-1.7914	–	–	–	–	–
β_1	MCMC Mean	9.1411	-6.0349	-8.7143	2.2665	-1.7694	–	–	–	–	–
β_1	MCMC error	0.0413	0.0417	0.0411	0.0462	0.1481	–	–	–	–	–
β_2	True	-4.1189	6.0930	-7.7552	-7.8354	-3.8545	–	–	–	–	–
β_2	MCMC Mean	-4.1114	6.0878	-7.7426	-7.8315	-3.8024	–	–	–	–	–
β_2	MCMC error	0.0379	0.0393	0.0389	0.0413	0.3828	–	–	–	–	–
β_3	True	-9.5885	-9.8816	-4.8716	5.4633	8.1354	–	–	–	–	–

Table 6: True, Monte Carlo error, and posterior expectation of each β per parameter set and cluster *(continued)*

		Cluster									
Parameter	Estimate	1	2	3	4	5	6	7	8	9	10
β_3	MCMC Mean	-9.5946	-9.8767	-4.8742	5.4583	8.0071	–	–	–	–	–
β_3	MCMC error	0.0434	0.0379	0.0388	0.0437	0.8725	–	–	–	–	–
β_4	True	4.9159	7.8931	8.7228	9.0426	-4.3706	–	–	–	–	–
β_4	MCMC Mean	4.9136	7.8858	8.7227	9.0325	-4.2889	–	–	–	–	–
β_4	MCMC error	0.0420	0.0462	0.0473	0.0419	0.4999	–	–	–	–	–
Parameter Set 3 (N. of Clusters: 4; N. of Covariates: 3)											
β_1	True	2.1420	-3.2250	-5.9593	5.7489	–	–	–	–	–	–
β_1	MCMC Mean	1.8353	-3.2095	-5.9560	5.7361	–	–	–	–	–	–
β_1	MCMC error	0.7418	0.0743	0.0479	0.0471	–	–	–	–	–	–
β_2	True	-2.2197	-1.6789	-4.9497	-8.0287	–	–	–	–	–	–
β_2	MCMC Mean	-2.0263	-1.6679	-4.9482	-8.0293	–	–	–	–	–	–
β_2	MCMC error	0.4735	0.0439	0.0389	0.0431	–	–	–	–	–	–
β_3	True	-2.3680	-5.3547	4.6992	-5.3455	–	–	–	–	–	–
β_3	MCMC Mean	-2.1589	-5.3479	4.7003	-5.3371	–	–	–	–	–	–
β_3	MCMC error	0.4878	0.0847	0.0401	0.0411	–	–	–	–	–	–
β_4	True	-0.5086	-0.3742	7.8034	7.7347	–	–	–	–	–	–
β_4	MCMC Mean	-0.4262	-0.3669	7.8005	7.7390	–	–	–	–	–	–
β_4	MCMC error	0.2304	0.0459	0.0460	0.0392	–	–	–	–	–	–
Parameter Set 4 (N. of Clusters: 10; N. of Covariates: 3)											
β_1	True	6.4346	8.6388	-7.1063	-0.5870	7.9331	-4.6517	8.7052	-2.8011	3.6076	4.1293
β_1	MCMC Mean	6.4254	8.6307	-7.1046	-0.4646	–	-4.6518	8.7140	-2.7246	3.5984	3.9971
β_1	MCMC error	0.0632	0.0562	0.0537	0.2964	–	0.0583	0.0595	0.3052	0.0938	0.3760
β_2	True	-2.6849	8.7468	9.8020	1.9841	-4.0242	-9.8518	-2.7650	-0.6405	7.9583	-4.3867
β_2	MCMC Mean	-2.6908	8.7506	9.7952	1.7767	–	-9.8542	-2.7652	-0.6221	7.9330	-4.1721
β_2	MCMC error	0.0610	0.0527	0.0515	0.4119	–	0.0449	0.0588	0.1044	0.2131	0.4874
β_3	True	-5.5527	1.6805	6.3605	2.7285	-1.0697	4.8846	-6.6493	4.5761	0.5479	1.6996
β_3	MCMC Mean	-5.5470	1.6855	6.3594	2.4642	–	4.8889	-6.6377	4.4818	0.5594	1.6212
β_3	MCMC error	0.0564	0.0526	0.0530	0.5743	–	0.0528	0.0586	0.3773	0.0648	0.2616
β_4	True	8.6219	-7.7376	-0.5569	2.6615	2.6063	5.7916	1.3942	-8.9141	-1.4563	0.6422
β_4	MCMC Mean	8.6174	-7.7381	-0.5505	2.3886	–	5.7896	1.3884	-8.7106	-1.4624	0.6404
β_4	MCMC error	0.0556	0.0574	0.0468	0.5946	–	0.0491	0.0605	0.8134	0.0696	0.1417
Parameter Set 5 (N. of Clusters: 7; N. of Covariates: 5)											
β_1	True	-3.7746	-5.4668	-7.6802	1.1977	5.9336	-0.2147	-6.1483	–	–	–
β_1	MCMC Mean	-3.7702	-5.4786	-7.6816	1.1949	5.9273	-0.2204	-6.1439	–	–	–
β_1	MCMC error	0.0533	0.0662	0.0466	0.0588	0.0426	0.0731	0.0468	–	–	–
β_2	True	-1.2831	-7.8355	-6.5404	-0.3373	-7.0447	-6.2112	8.5538	–	–	–
β_2	MCMC Mean	-1.2776	-7.8253	-6.5314	-0.3387	-7.0344	-5.7822	8.5433	–	–	–
β_2	MCMC error	0.0512	0.0463	0.0457	0.0467	0.0521	1.4820	0.0440	–	–	–
β_3	True	-1.3126	-8.3251	4.5719	-6.6823	0.0608	-1.0889	-6.4825	–	–	–
β_3	MCMC Mean	-1.3157	-8.3086	4.5622	-6.6700	0.0616	-1.0417	-6.4765	–	–	–
β_3	MCMC error	0.0525	0.0549	0.0471	0.0459	0.0379	0.1973	0.0474	–	–	–
β_4	True	-3.0599	-9.7882	-9.0171	-8.2836	6.8602	-2.4649	4.5642	–	–	–
β_4	MCMC Mean	-3.0467	-9.7698	-9.0050	-8.2824	6.8634	-2.3245	4.5694	–	–	–
β_4	MCMC error	0.0607	0.0544	0.0376	0.0507	0.0470	0.5041	0.0473	–	–	–
β_5	True	6.6265	3.6617	-3.4436	3.7347	7.0128	0.4728	6.0935	–	–	–
β_5	MCMC Mean	6.6150	3.6623	-3.4362	3.7322	7.0067	0.4606	6.0952	–	–	–
β_5	MCMC error	0.0510	0.0477	0.0513	0.0545	0.0497	0.0741	0.0486	–	–	–
β_6	True	5.6638	1.5949	6.4922	-3.5569	-1.2739	1.9962	-9.9017	–	–	–
β_6	MCMC Mean	5.6558	1.5807	6.4879	-3.5475	-1.2793	1.8697	-9.8919	–	–	–
β_6	MCMC error	0.0591	0.0623	0.0511	0.0489	0.0493	0.4384	0.0473	–	–	–

Table 6: True, Monte Carlo error, and posterior expectation of each β per parameter set and cluster *(continued)*

		Cluster									
Parameter	Estimate	1	2	3	4	5	6	7	8	9	10
Parameter Set 6 (N. of Clusters: 1; N. of Covariates: 0)											
β_1	True	-0.1968	–	–	–	–	–	–	–	–	–
β_1	MCMC Mean	-0.2171	–	–	–	–	–	–	–	–	–
β_1	MCMC error	0.1673	–	–	–	–	–	–	–	–	–
Parameter Set 7 (N. of Clusters: 4; N. of Covariates: 2)											
β_1	True	3.3510	5.5764	7.1851	-5.8302	–	–	–	–	–	–
β_1	MCMC Mean	3.3476	5.5469	7.1918	-5.3099	–	–	–	–	–	–
β_1	MCMC error	0.0994	0.2005	0.0372	1.5248	–	–	–	–	–	–
β_2	True	-4.3097	2.9101	7.9174	4.1559	–	–	–	–	–	–
β_2	MCMC Mean	-4.2880	2.8864	7.9131	3.8197	–	–	–	–	–	–
β_2	MCMC error	0.1681	0.1204	0.0384	0.9978	–	–	–	–	–	–
β_3	True	-9.4016	-6.3153	4.9380	-3.4102	–	–	–	–	–	–
β_3	MCMC Mean	-9.3762	-6.2819	4.9363	-3.1650	–	–	–	–	–	–
β_3	MCMC error	0.2617	0.2057	0.0404	0.7669	–	–	–	–	–	–
Parameter Set 8 (N. of Clusters: 2; N. of Covariates: 5)											
β_1	True	5.6029	6.0953	–	–	–	–	–	–	–	–
β_1	MCMC Mean	5.6057	5.8913	–	–	–	–	–	–	–	–
β_1	MCMC error	0.0365	1.0385	–	–	–	–	–	–	–	–
β_2	True	3.7438	-5.3878	–	–	–	–	–	–	–	–
β_2	MCMC Mean	3.7405	-5.1924	–	–	–	–	–	–	–	–
β_2	MCMC error	0.0344	0.9935	–	–	–	–	–	–	–	–
β_3	True	-3.4657	1.3461	–	–	–	–	–	–	–	–
β_3	MCMC Mean	-3.4630	1.2993	–	–	–	–	–	–	–	–
β_3	MCMC error	0.0366	0.2394	–	–	–	–	–	–	–	–
β_4	True	-9.3687	-1.4674	–	–	–	–	–	–	–	–
β_4	MCMC Mean	-9.3684	-1.4309	–	–	–	–	–	–	–	–
β_4	MCMC error	0.0390	0.2103	–	–	–	–	–	–	–	–
β_5	True	3.0410	3.6361	–	–	–	–	–	–	–	–
β_5	MCMC Mean	3.0409	3.5077	–	–	–	–	–	–	–	–
β_5	MCMC error	0.0325	0.6258	–	–	–	–	–	–	–	–
β_6	True	-3.5824	-7.7044	–	–	–	–	–	–	–	–
β_6	MCMC Mean	-3.5858	-7.4373	–	–	–	–	–	–	–	–
β_6	MCMC error	0.0347	1.3367	–	–	–	–	–	–	–	–
Parameter Set 9 (N. of Clusters: 10; N. of Covariates: 2)											
β_1	True	-8.0023	7.2055	5.0725	-8.2809	-8.7811	4.2359	-9.9581	2.4104	-3.1409	-0.7061
β_1	MCMC Mean	-7.9718	6.8629	–	-8.0508	-8.7664	4.1149	-9.9473	2.3620	-2.9757	-0.5916
β_1	MCMC error	0.2386	0.7669	–	0.7421	0.2083	0.4606	0.0578	0.2952	0.3749	0.3023
β_2	True	8.0038	0.9477	3.5049	-5.4635	7.8988	9.4809	-3.3568	5.0702	-3.6214	2.0435
β_2	MCMC Mean	7.9600	0.9294	–	-5.2767	7.8755	9.2670	-3.3558	4.8840	-3.4042	1.7063
β_2	MCMC error	0.2188	0.0927	–	0.5682	0.1650	0.7722	0.0534	0.5710	0.4900	0.5177
β_3	True	5.3229	1.7302	-8.6640	-1.1130	-3.8013	1.9706	-8.7511	-7.4686	6.7933	3.5699
β_3	MCMC Mean	5.2997	1.6535	–	-1.0755	-3.7790	1.9068	-8.7519	-7.1621	6.4443	2.8895
β_3	MCMC error	0.1487	0.1612	–	0.1376	0.1766	0.2024	0.0584	0.8681	0.7768	0.9490
Parameter Set 10 (N. of Clusters: 3; N. of Covariates: 2)											
β_1	True	0.0924	0.1113	-7.4930	–	–	–	–	–	–	–
β_1	MCMC Mean	0.0387	0.1130	-7.4726	–	–	–	–	–	–	–
β_1	MCMC error	0.1826	0.0490	0.1489	–	–	–	–	–	–	–
β_2	True	3.6043	-6.8301	6.8375	–	–	–	–	–	–	–
β_2	MCMC Mean	3.2693	-6.7974	6.8269	–	–	–	–	–	–	–

Table 7: Summary of the performance of the hdpGLM when estimating number of clusters (K) and linear coefficients (β) across 1000 replications generated by 2 different parameter sets.

	Number of Clusters (K)							
		Estimates across replications				Coverage and HPD of linear coefficients (β)		
Number of Covariates	True	Mean	Minimum	Maximum	Correct (%)	Minimum.	Average	95% HPD (largest average)
4	3	3.03	3	4	96.7	88.21	94.22	(6.8238, 7.3827)
1	5	5.35	5	9	70.2	92.62	97.97	(-1.4744, 4.0107)

Table 6: True, Monte Carlo error, and posterior expectation of each β per parameter set and cluster *(continued)*

		Cluster									
Parameter	Estimate	1	2	3	4	5	6	7	8	9	10
β_2	MCMC error	0.8220	0.2302	0.1298	–	–	–	–	–	–	–
β_3	True	2.6229	6.4300	-1.2320	–	–	–	–	–	–	–
β_3	MCMC Mean	2.4522	6.4049	-1.2277	–	–	–	–	–	–	–
β_3	MCMC error	0.4166	0.1980	0.0546	–	–	–	–	–	–	–

Table 7 shows values of the MC study using 1000 simulations for two parameter sets. As we can see, the number of estimated clusters, coverage, and HPD intervals are reinforced the results of the MC exercise with 100 replications for ten different parameter sets. The same is true for the MC error displayed in tables 8 and 9.

Table 9: True, MC error, and posterior expectation of eacn β per parameter set and cluster

		Cluster				
Parameter	Estimate	1	2	3	4	5
Parameter Set 1 (N. of Clusters: 5; N. of Covariates: 1)						
β_1	True	-9.7860	-6.5590	-5.0161	3.2058	1.8956
β_1	MCMC Mean	-9.6749	–	-4.7423	2.3554	–
β_1	MCMC error	0.4887	–	0.6616	0.9238	–
β_2	True	-3.0997	8.9872	4.6558	1.6063	7.3254
β_2	MCMC Mean	-3.0461	–	4.5645	1.7286	–
β_2	MCMC error	0.2578	–	0.5309	0.3695	–
Parameter Set 2 (N. of Clusters: 3; N. of Covariates: 4)						
β_1	True	-1.6338	2.8461	-0.2693	–	–
β_1	MCMC Mean	-1.5796	2.8450	-0.2690	–	–
β_1	MCMC error	0.2595	0.0416	0.0522	–	–
β_2	True	7.3505	0.4319	1.1724	–	–

Table 8: Summary of MC error across replications of each parameter set

			MC error			
Parameter Set	N. Clusters	N. Covariates	Mean	Std. Dev.	Minimum	Maximum
1	5	1	0.5387	0.2339	0.2578	0.9238
2	3	4	0.2136	0.3248	0.0300	1.2298

Table 9: True, MC error, and posterior expectation of eacn β per parameter set and cluster *(continued)*

		Cluster				
Parameter	Estimate	1	2	3	4	5
β_2	MCMC Mean	7.1099	0.4178	1.1766	–	–
β_2	MCMC error	1.2298	0.0442	0.0334	–	–
β_3	True	-2.9529	-6.4458	9.7919	–	–
β_3	MCMC Mean	-2.8634	-6.4548	9.7907	–	–
β_3	MCMC error	0.4717	0.0461	0.0565	–	–
β_4	True	-2.2035	-9.4002	4.0393	–	–
β_4	MCMC Mean	-2.1458	-9.4045	4.0415	–	–
β_4	MCMC error	0.3626	0.0300	0.0357	–	–
β_5	True	-2.3907	5.4716	6.6912	–	–
β_5	MCMC Mean	-2.3058	5.4648	6.6988	–	–
β_5	MCMC error	0.4575	0.0368	0.0470	–	–

4 Example with binary outcome

This section displays a simulation using binary outcome. As the main paper discusses, for binary outcome variable the algorithm uses RMHMC update within Gibbs for the parameter β. So the MCMC algorithm can take a long time before satisfactory convergence diagnostics are achieved. Therefore, we present results for three simulated cases only, one with no heterogeneity, one with two clusters, and one with three clusters. The procedures adopted are similar to those described in the main paper. We used 80000 burn-in iterations, and 30000 samples were recorded after that. The figures 2 to 4 show the estimation for the three cases.

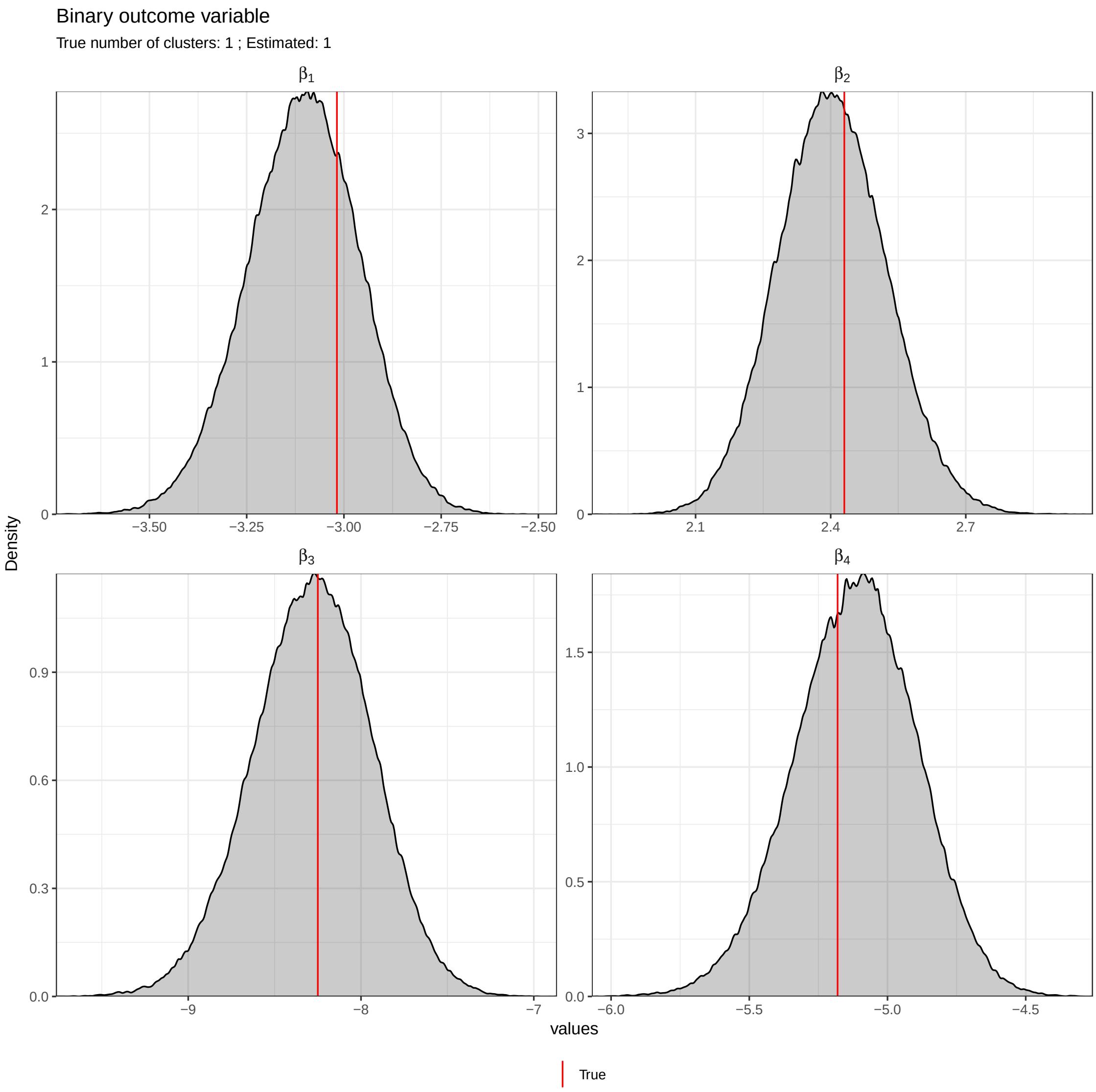

Figure 2: Estimation of the hdpGLM model with binary outcome variable and one cluster $(K = 1)$

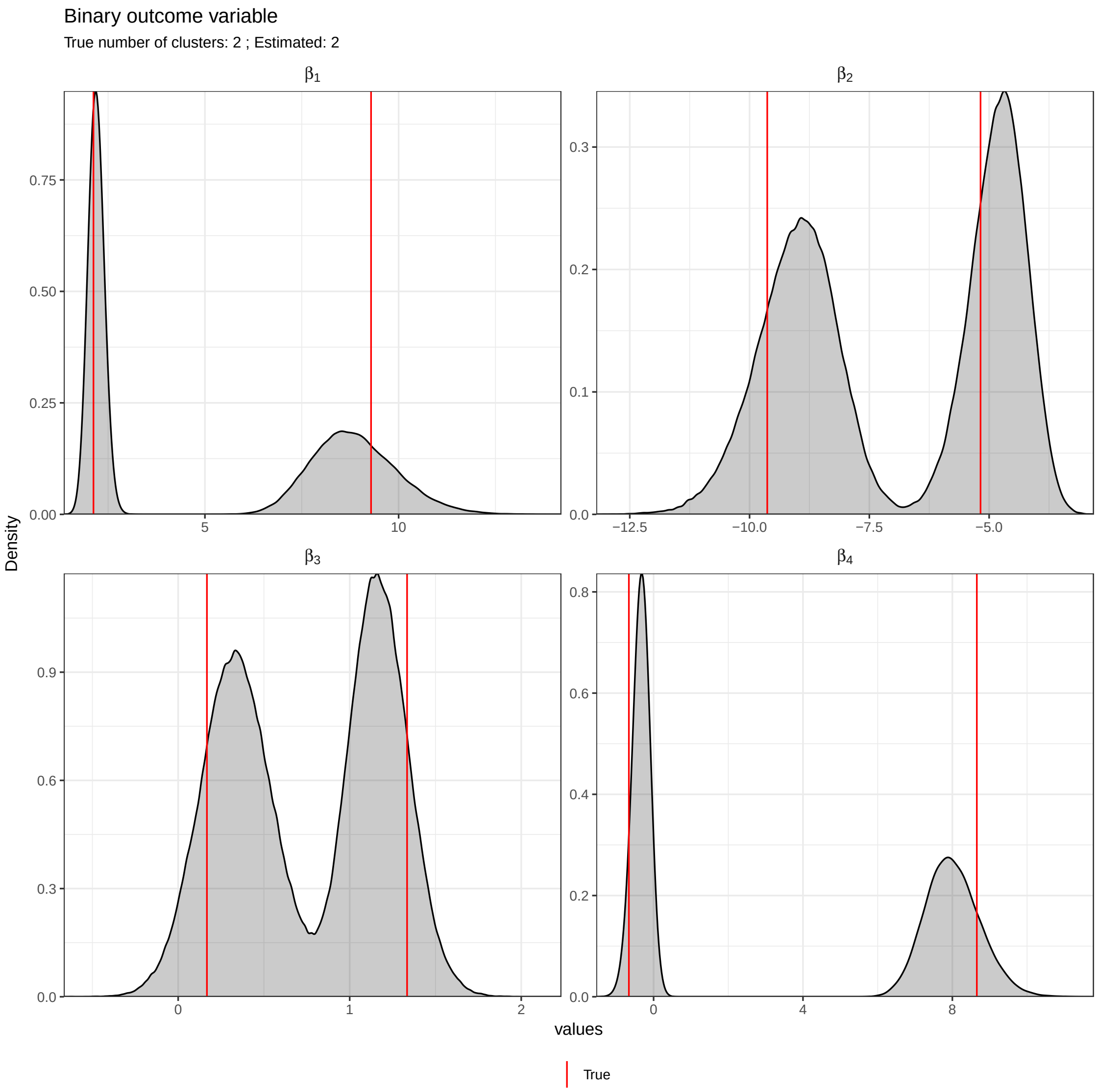

Figure 3: Estimation of the hdpGLM model with binary outcome variable and two clusters ($K = 2$)

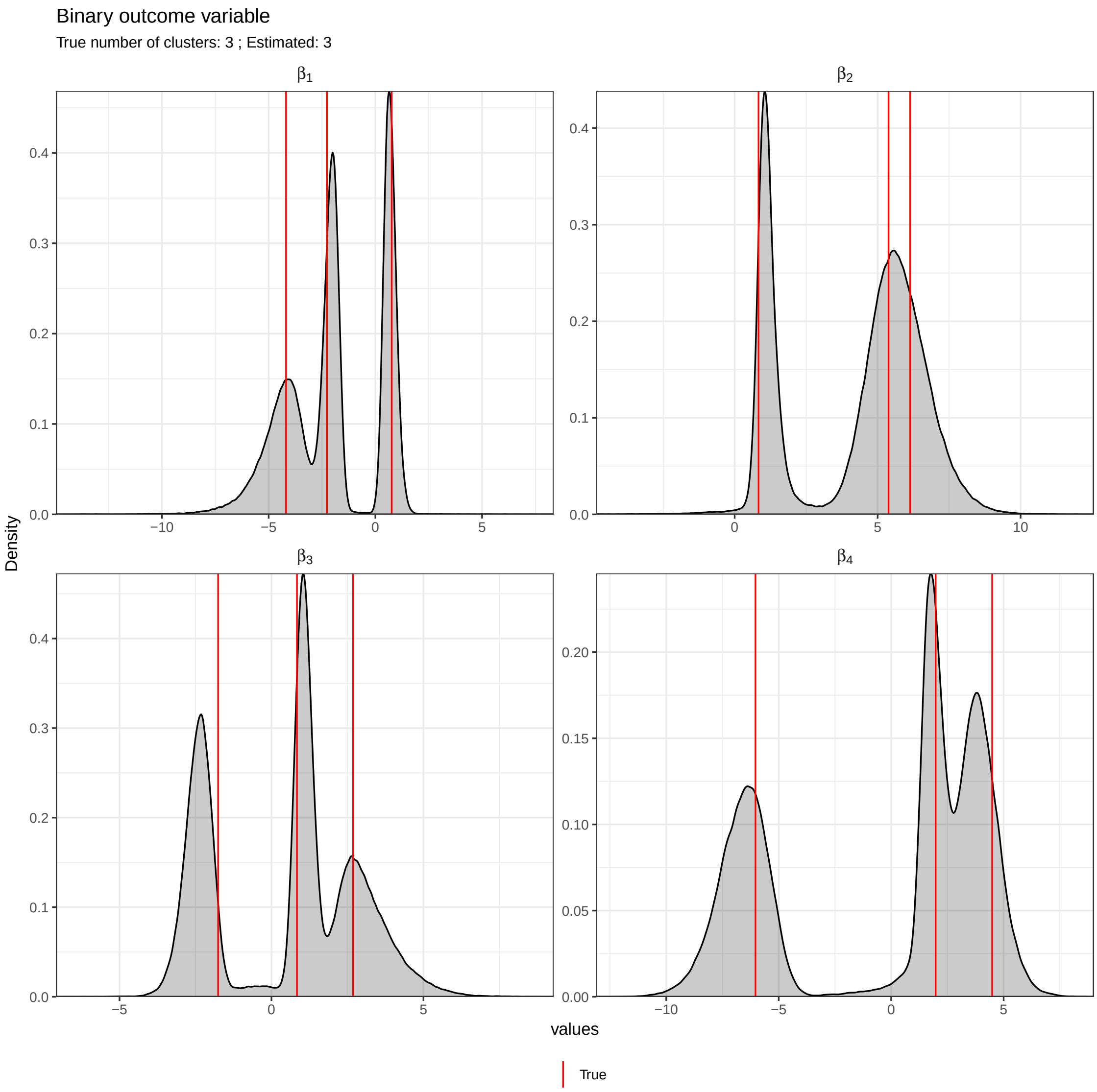

Figure 4: Estimation of the hdpGLM model with binary outcome variable and three clusters ($K = 3$)

5 Marginal Densities of β

As described in the main paper, we randomly generated 10 parameter sets and, for each one, we generated 100 data sets. The hdpGLM was estimated for each one of the 1000 data sets. The Figures 5 to 13 display the marginal posterior density plots for the linear coefficients β for each data set generated by the parameter sets. The HPDI and the MCMC average in the figures were computed across data sets. As we can see, the posterior densities are located around the true, and the MCMC posterior average is quite close to the true value that generated the data. The estimated number of clusters and the coverage probability of the linear parameters are presented in the main paper.

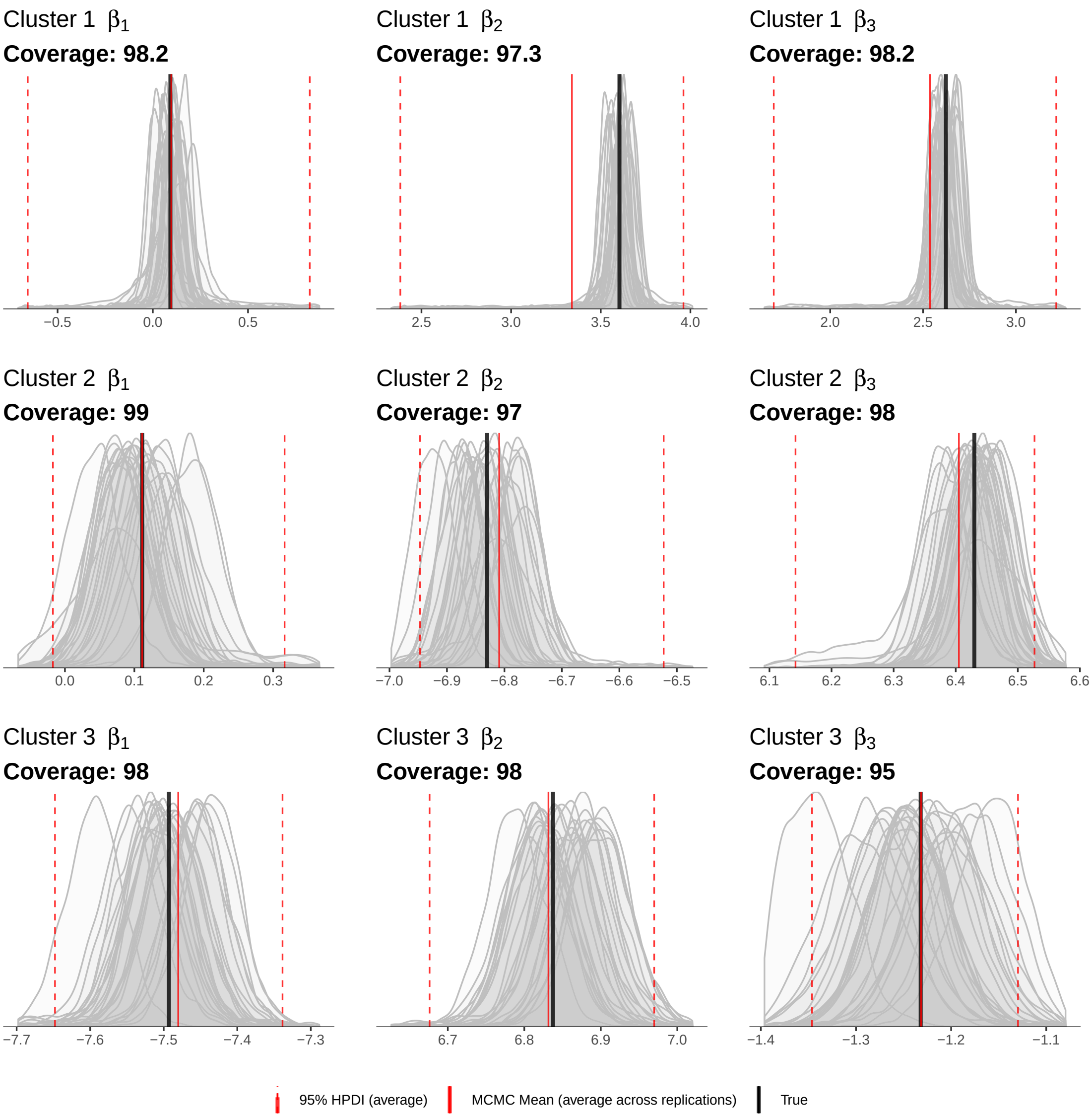

Figure 5: Marginal Posterior Distribution of β for the estimation of the 50 data sets generated using the parameter set with 3 clusters and 2 covariates.

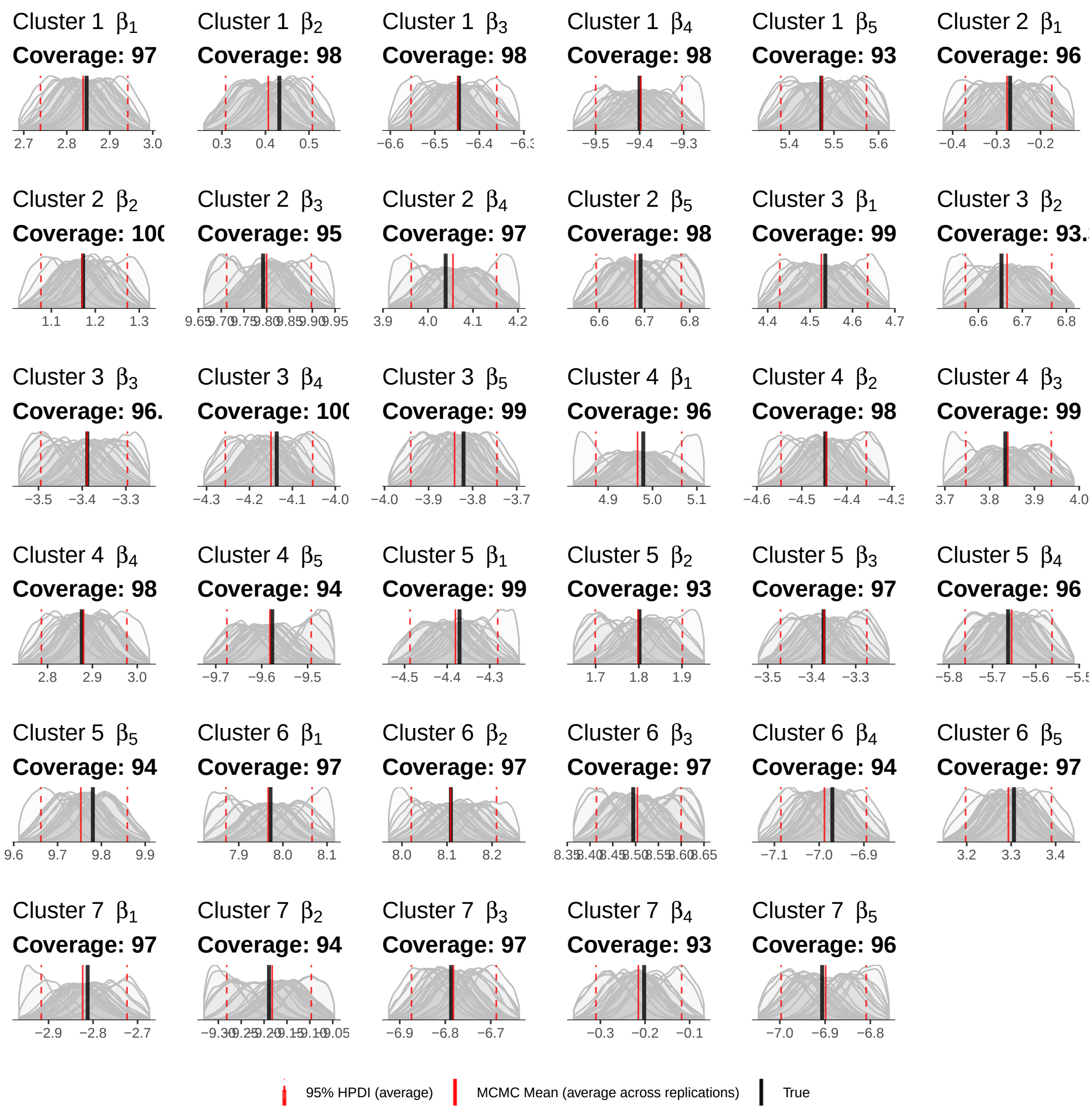

Figure 6: Marginal Posterior Distribution of β for the estimation of the 50 data sets generated using the parameter set with 7 clusters and 4 covariates.

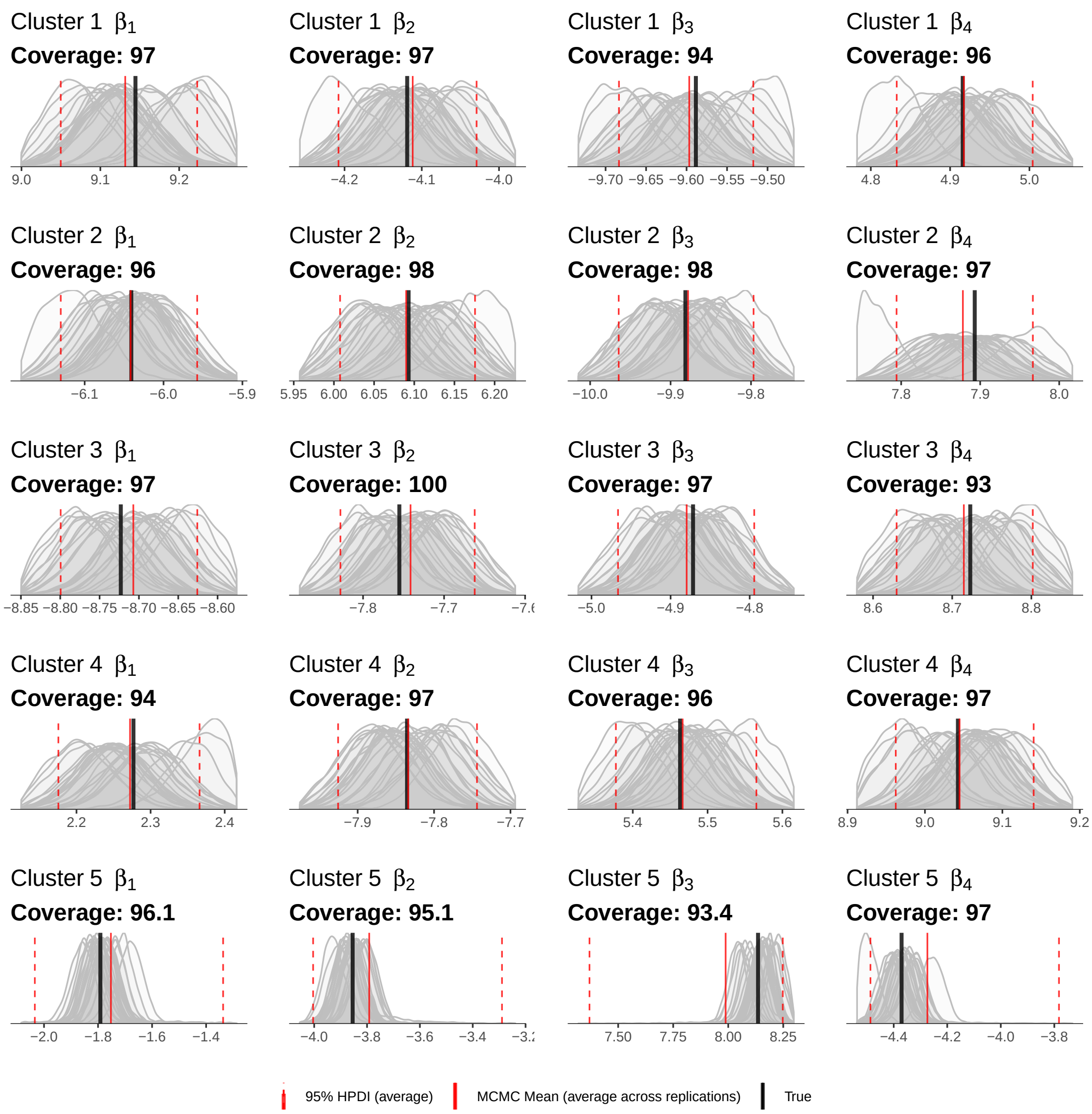

Figure 7: Marginal Posterior Distribution of β for the estimation of the 50 data sets generated using the parameter set with 5 clusters and 3 covariates.

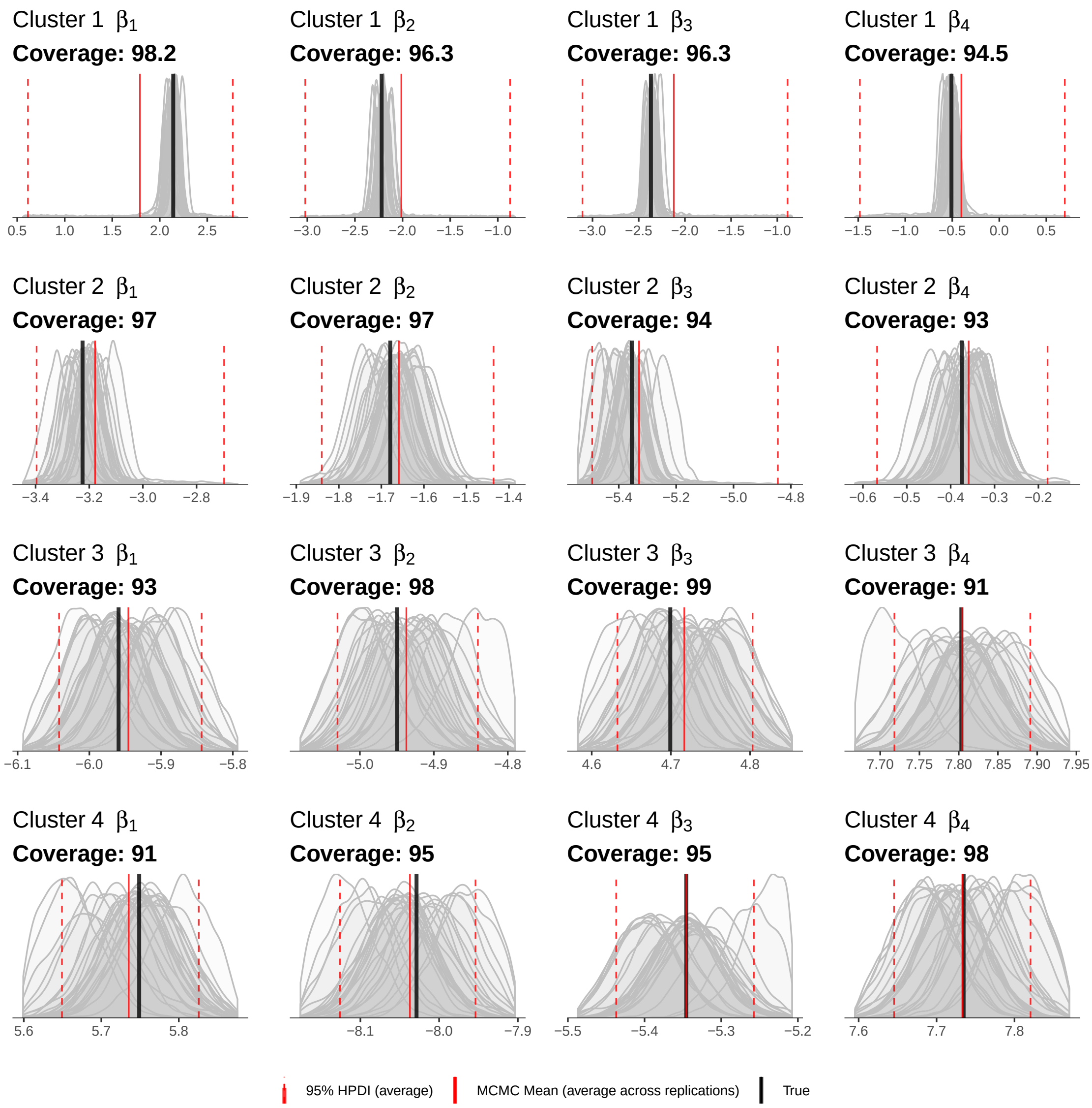

Figure 8: Marginal Posterior Distribution of β for the estimation of the 50 data sets generated using the parameter set with 4 clusters and 3 covariates.

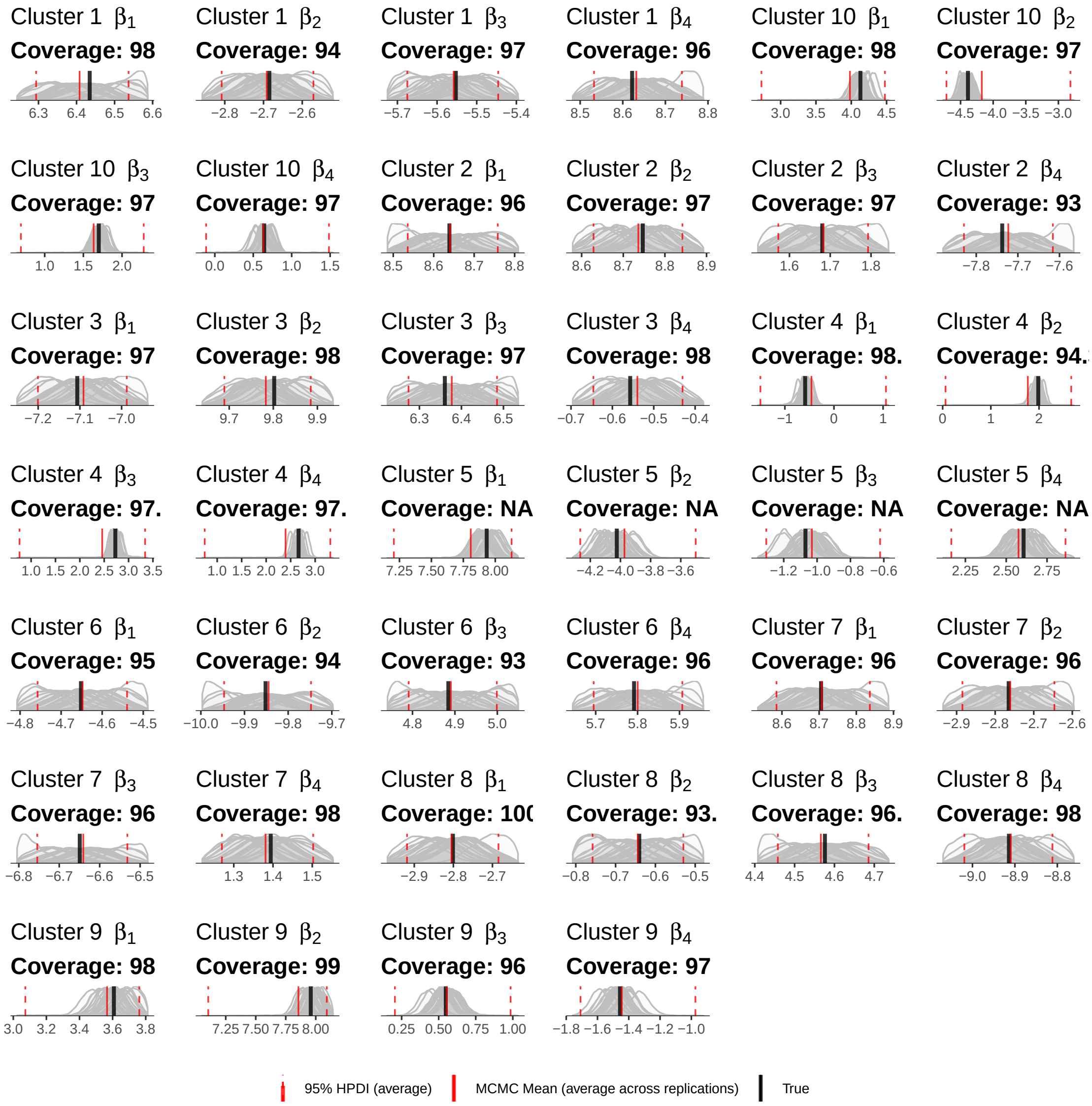

Figure 9: Marginal Posterior Distribution of β for the estimation of the 50 data sets generated using the parameter set with 10 clusters and 3 covariates.

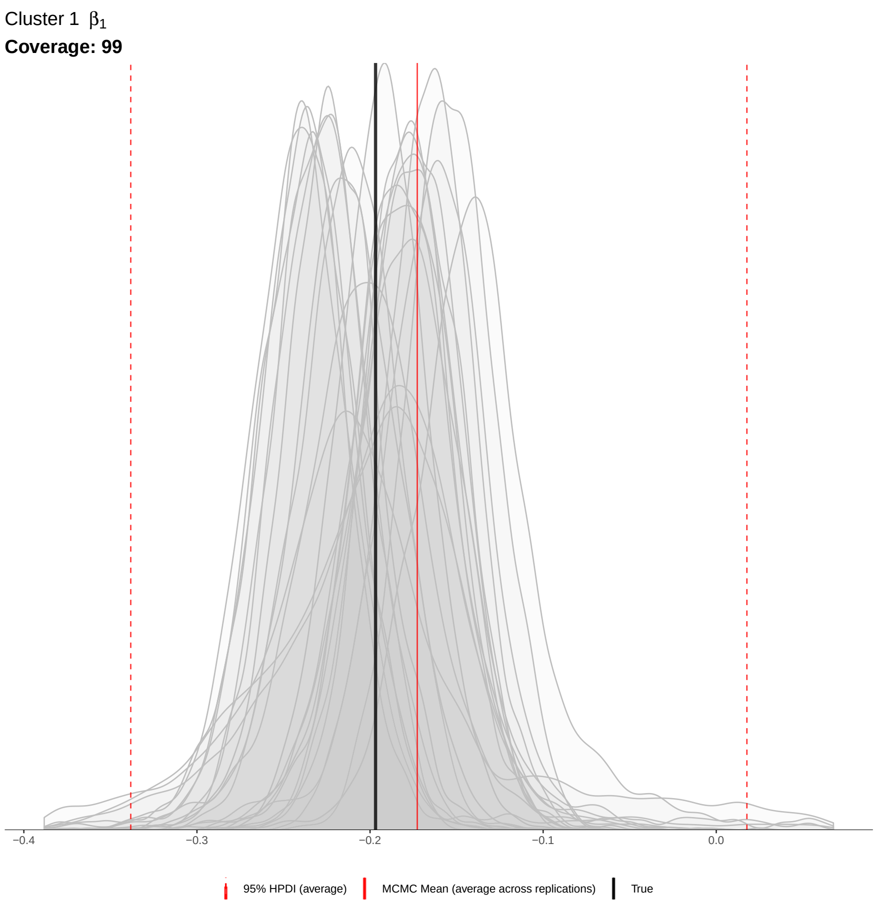

Figure 10: Marginal Posterior Distribution of β for the estimation of the 50 data sets generated using the parameter set with 1 cluster and 0 covariates.

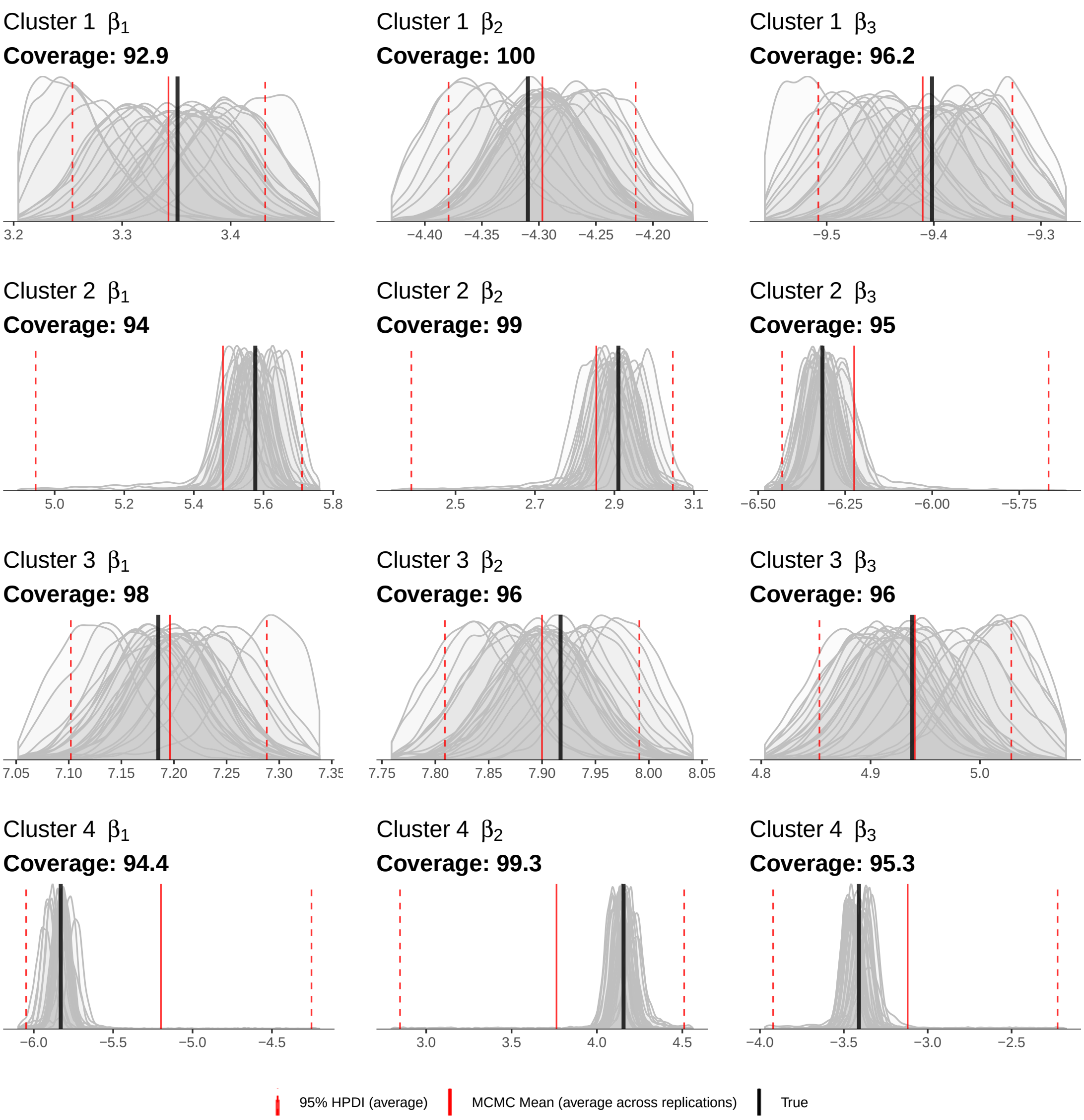

Figure 11: Marginal Posterior Distribution of β for the estimation of the 50 data sets generated using the parameter set with 4 clusters and 2 covariates.

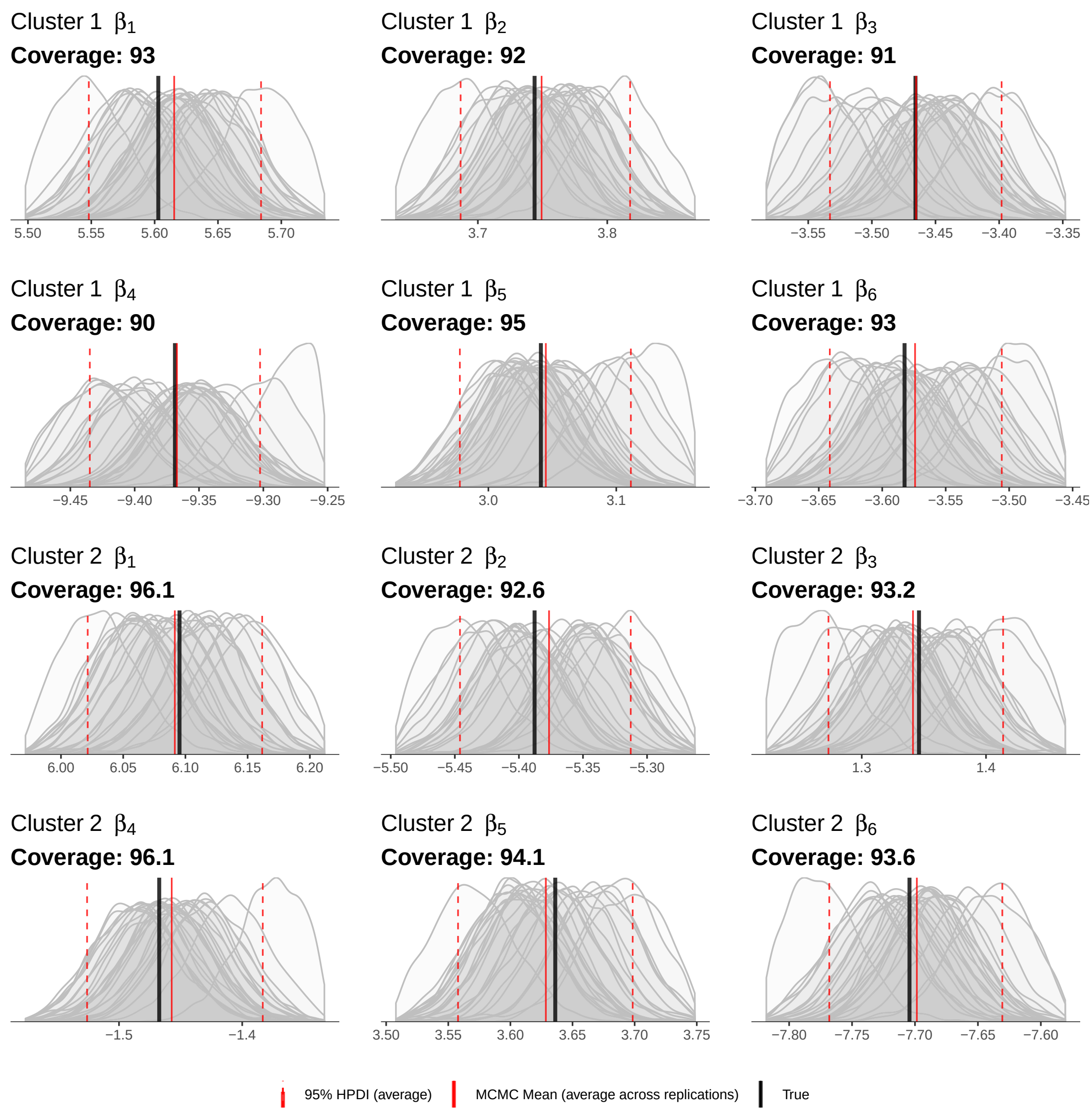

Figure 12: Marginal Posterior Distribution of β for the estimation of the 50 data sets generated using the parameter set with 2 clusters and 5 covariates.

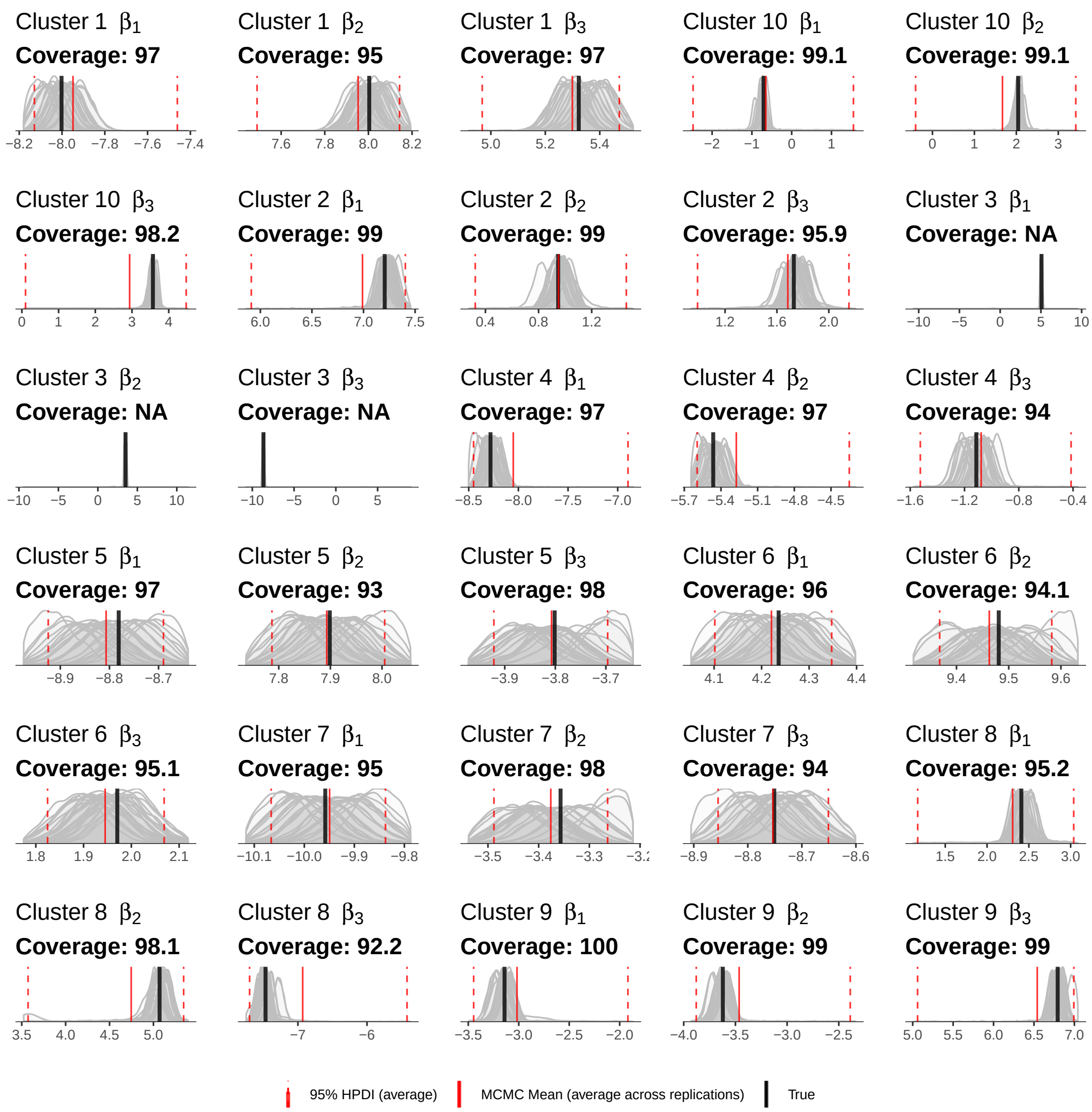

Figure 13: Marginal Posterior Distribution of β for the estimation of the 50 data sets generated using the parameter set with 10 clusters and 2 covariates.

References

Antoniak, Charles E. 1974. "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The annals of statistics* pp. 1152–1174.

Calin, Ovidiu and Der-Chen Chang. 2006. *Geometric mechanics on Riemannian manifolds: applications to partial differential equations*. Springer Science & Business Media.

Dorazio, Robert M. 2009. "On selecting a prior for the precision parameter of Dirichlet process mixture models." *Journal of Statistical Planning and Inference* 139(9):3384–3390.

Duane, Simon, Anthony D Kennedy, Brian J Pendleton and Duncan Roweth. 1987. "Hybrid monte carlo." *Physics letters B* 195(2):216–222.

Flegal, James Marshall. 2008. Monte Carlo Standard Errors for Markov Chain Monte Carlo PhD thesis University of Minnesota.

Gelfand, Alan E, Athanasios Kottas and Steven N MacEachern. 2005. "Bayesian nonparametric spatial modeling with Dirichlet process mixing." *Journal of the American Statistical Association* 100(471):1021–1035.

Girolami, Mark and Ben Calderhead. 2011. "Riemann manifold langevin and hamiltonian monte carlo methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(2):123–214.

Gong, Lei and James M Flegal. 2015. "A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo." *Journal of Computational and Graphical Statistics* (just-accepted):00–00.

Liu, Jun S. 2008. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

Mukhopadhyay, Saurabh and Alan E Gelfand. 1997. "Dirichlet process mixed generalized linear models." *journal of the American Statistical Association* 92(438):633–639.

Neal, Radford M et al. 2011. MCMC using Hamiltonian dynamics. Vol. 2 CRC Press New York, NY.

Rossi, Peter. 2014. *Bayesian Non- and Semi-parametric Methods and Applications*. Princeton University Press.

West, Mike, Peter Muller and Michael D Escobar. 1994. Hierarchical priors and mixture models, with application in regression and density estimation. Wiley pp. 363–386.