Appendix A: A Further Application

As second application we consider data derived from the National Survey of Maternal and Child Health in Guatemala in 1987. The data is available from the R-package `mlmRev` (Bates et al., 2014) and was also analysed by Rodriguez and Goldman (2001). The data contains observations of children that were born in the 5-year period before the survey. In our analysis we include 1211 children living in 45 communities. One observes a minimal number of 20, a maximal number of 50 and an average number of 26.9 pregnancies per community. The response y_{ij} is a binary outcome with $y_{ij} = 0$ for traditional prenatal care and $y_{ij} = 1$ for modern prenatal care, for example by doctors or nurses. The response is modelled by a logistic regression model $\text{logit}(P(y_{ij} = 1)) = \eta_{ij}$. The heterogeneity of communities is modelled by the alternative approaches considered here. The distribution of the response as well as of the two binary and five categorical explanatory variables that characterize the children's mothers and their families are given in Table 3.

An overview of the estimated coefficients when using a generalized mixed model (GMM), tree-structured clustering (TSC) and a finite mixture model (FIN) is given in Table 4. The 95% confidence intervals were obtained by 2000 bootstrap samples. It can be seen from the results that the age of the mother at the time of the survey as well as the employment status of the husband do not have a significant effect on the form of prenatal care. The educational level of the mother as well as of the husband, however, have a strong impact. For births, where the mother at least finished primary or the husband finished primary, modern prenatal care was provided more likely compared to births of parents without any graduation. Indigenous mothers (speaking and not speaking Spanish) are also more likely to use traditional prenatal care than non-indigenous mothers. The existence of a modern toilet in the household does not favour the use of modern prenatal care, whereas it is preferred by families using the television regularly.

A comparison of the estimates obtained by the three methods does not show strong distinctions and no clear tendency. Differences occur for variable ethnicity (first rows in Table 4), for which the two estimates of the mixed model are larger than for TSC and FIN and for mothers that finished secondary (fourth row) for which the estimate of the finite mixture model is larger than for TSC and GMM.

TABLE 3: *Description and distribution of the response and the covariates used for the analysis of the Guatemala survey.*

Variable	Description	Categories	Frequency
y	Modern prenatal care	no	733
		yes	478
ethn	Mother's ethnicity	non-indigenous (Ladino)	612
		indigenous, not speaking Spanish	286
		indigenous, speaking Spanish	313
momEd	Mother's level of education	not finished primary	571
		finished primary	607
		finished secondary	33
husEd	Husband's level of education	not finished primary	430
		finished primary	598
		finished secondary	67
		unknown	116
husEmpl	Husband's employment status	unskilled	45
		professional	120
		agricultural, self-employed	420
		agricultural, employee	407
		skilled service	219
telev	Frequency of TV usage	never	1034
		not daily	52
		daily	125
momAge	Mother 25 years or older	no	583
		yes	628
toilet	Modern toilet in house	no	112
		yes	1099

The estimated community-specific intercepts obtained by tree-structured clustering and the finite mixture model are given in the lower panel of Table 4. Using the tree-structured model results in three clusters of communities that differ in terms of their probability to use modern prenatal care. The finite mixture identifies only two clusters. We prefer to use model selection by BIC as it showed more stable estimates in the simulations with binary response. The detected partitions and the high variance obtained by the mixed model indicate that heterogeneity of communities is definitely present. Nevertheless, only a few clusters of communities have to be distinguished. There is a strong similarity between the third cluster of the tree-structured model ($\beta_{i0}^{(3)} = 1.448$) and the second cluster of the finite mixture model ($\beta_{i0}^{(2)} = 1.465$). With the exception of community 13 they are composed of the same units and have nearly the same estimate. The big cluster with 33 communities, obtained by the finite mixture model, is further

TABLE 4: *Estimation results of the Guatemala survey using the generalized mixed model, tree-structured clustering and the finite mixture model.*

Predictor	**GMM**		**TSC**		**FIN**	
	Coefficient	95%-CI	Coefficient	95%-CI	Coefficient	95%-CI
ethn						
not spanish	-1.370	[-2.101,-0.774]	-1.090	[-2.469,-0.387]	-0.995	[-2.280,-0.556]
spanish	-0.720	[-1.235,-0.244]	-0.434	[-1.425, 0.005]	-0.335	[-1.338, 0.011]
momEd						
primary	0.645	[0.331, 1.048]	0.673	[0.298, 1.122]	0.646	[0.317, 1.078]
secondary	1.385	[0.303, 2.955]	1.405	[0.268, 3.046]	1.735	[0.364, 2.944]
husEd						
primary	0.785	[0.445, 1.236]	0.817	[0.437, 1.303]	0.843	[0.444, 1.301]
secondary	0.194	[-0.809, 1.186]	0.049	[-0.922, 1.286]	0.291	[-0.846, 1.311]
unknown	0.398	[-0.113, 0.951]	0.520	[-0.101, 1.006]	0.428	[-0.106, 0.962]
husEmpl						
professional	-0.210	[-1.150, 0.670]	-0.095	[-1.301, 0.820]	-0.408	[-1.336, 0.667]
agricult, self	-0.119	[-0.975, 0.721]	-0.065	[-1.044, 0.798]	-0.266	[-1.065, 0.716]
agricult, empl	-0.158	[-1.024, 0.656]	-0.100	[-1.092, 0.750]	-0.238	[-1.103, 0.723]
skilled	-0.199	[-1.079, 0.606]	-0.125	[-1.123, 0.661]	-0.300	[-1.134, 0.607]
telev						
not daily	0.355	[-0.497, 1.292]	0.226	[-0.601, 1.286]	0.241	[-0.548, 1.283]
daily	0.867	[0.312, 1.560]	0.928	[0.290, 1.570]	0.735	[0.307, 1.524]
momAge	0.099	[-0.208, 0.403]	0.061	[-0.241, 0.411]	0.061	[-0.219, 0.401]
toilet	-0.869	[-1.833,-0.055]	-1.008	[-1.875, 0.092]	-0.839	[-1.808,-0.154]
β_0	-0.011	[-1.223, 1.166]	—	—	—	—
σ^2_{rand}	1.250	[1.233, 2.416]	—	—	—	—

Community-specific intercepts	**TSC**			**FIN**		
	Cluster	Size	Coefficient	Cluster	Size	Coefficient
β_{i0}	6,7,8,9,10,11,12,18,22, 24,31,32,34,37,42	15	-1.286	2,4,6,7,8,9,10,11,12,13,15, 16,17,18,19,22,20,21,23, 24,25,26,27,29,31,32,34, 37,42,33,36,39,42	33	-0.696
	2,4,15,16,17,19,20,21,23, 25,26,27,29,33,36,39,42	17	-0.214			
	1,3,5,13,14,28,30,35, 38,40,41,44,45	13	1.448	1,3,5,14,28,30,35, 38,40,41,44,45	12	1.465

split into two clusters by the tree-structured model. As the estimated coefficients (-1.286 and -0.214) and therefore the probabilities to use modern prenatal care are quite different, the solution of the tree-structured model seems more sensible.

Appendix B: Further Simulations

Display of Simulation Results

In the following we give the results of the settings of the simulations described in Section 6 that were not displayed in the manuscript. By analogy with Figures 5 to 7 in the manuscript, the figures contain the MSEs of the unit-specific intercepts, the MSEs of the linear term and the selected number of clusters. In addition, the tables contain the TPR and FPR as the average over 100 replications, respectively.

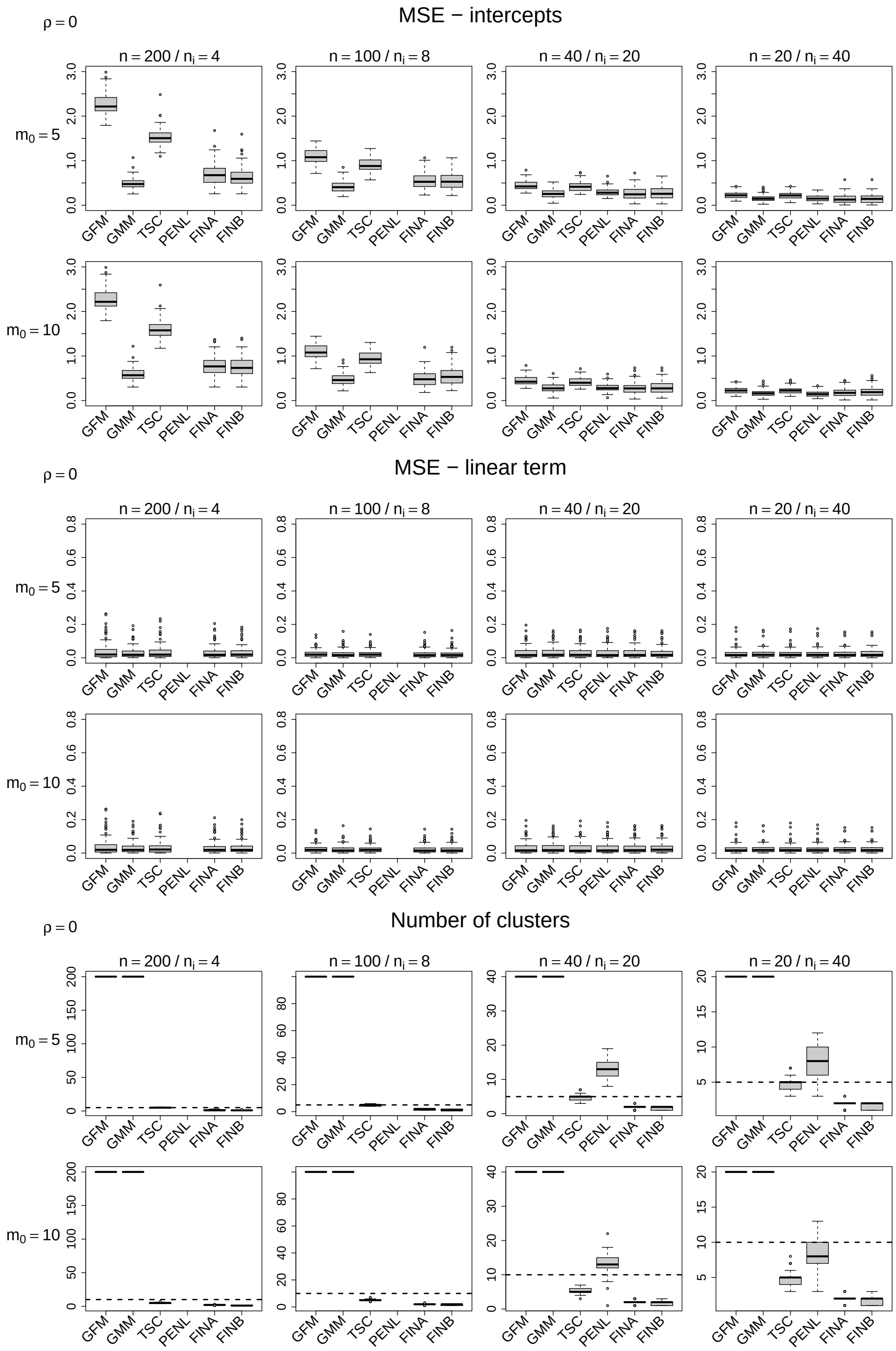

FIGURE 10: *Results for the settings with normal response, chi-squared distributed intercepts and* $\rho = 0$.

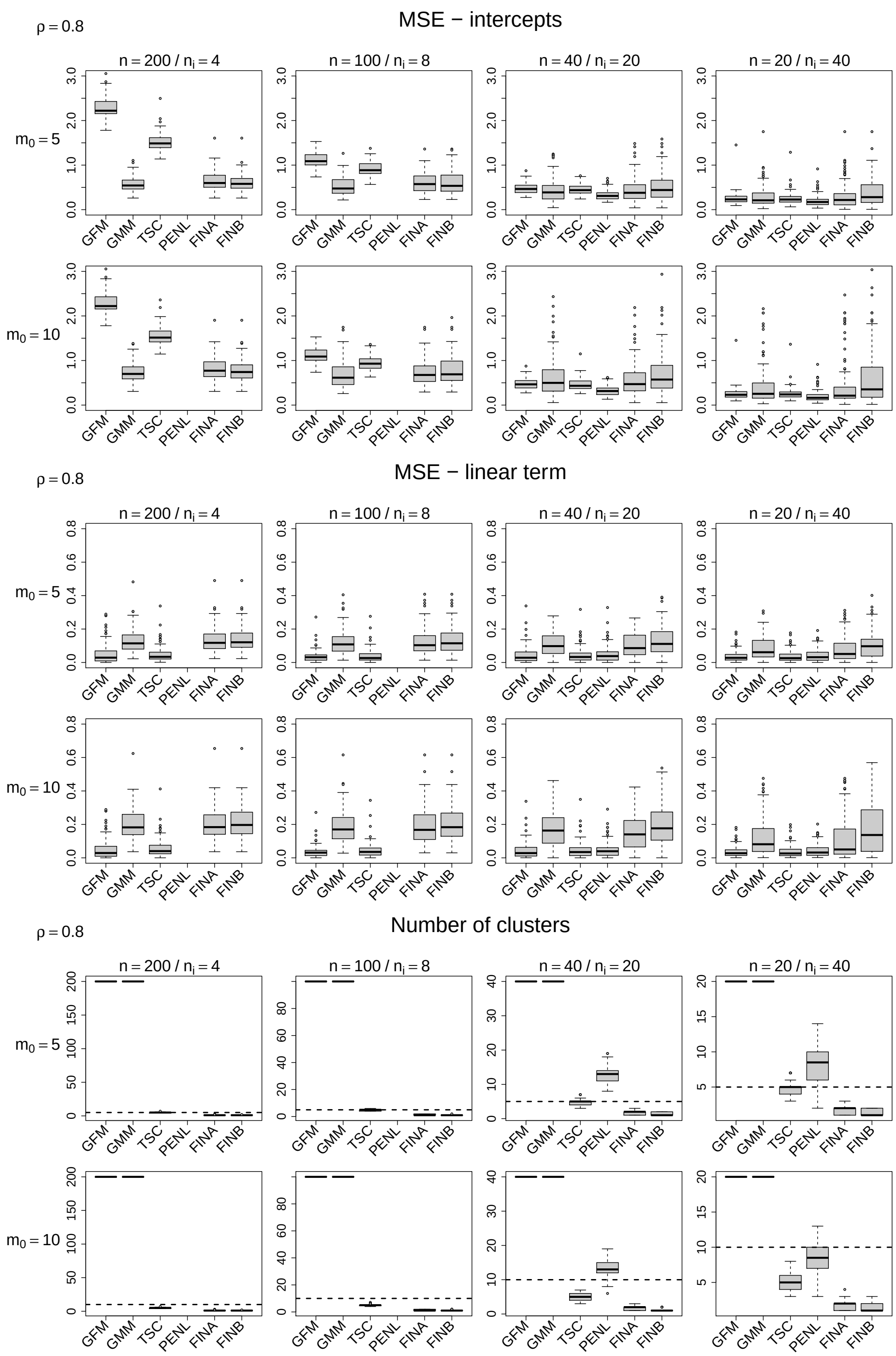

FIGURE 11: *Results for the settings with normal response, chi-squared distributed intercepts and* $\rho = 0.8$.

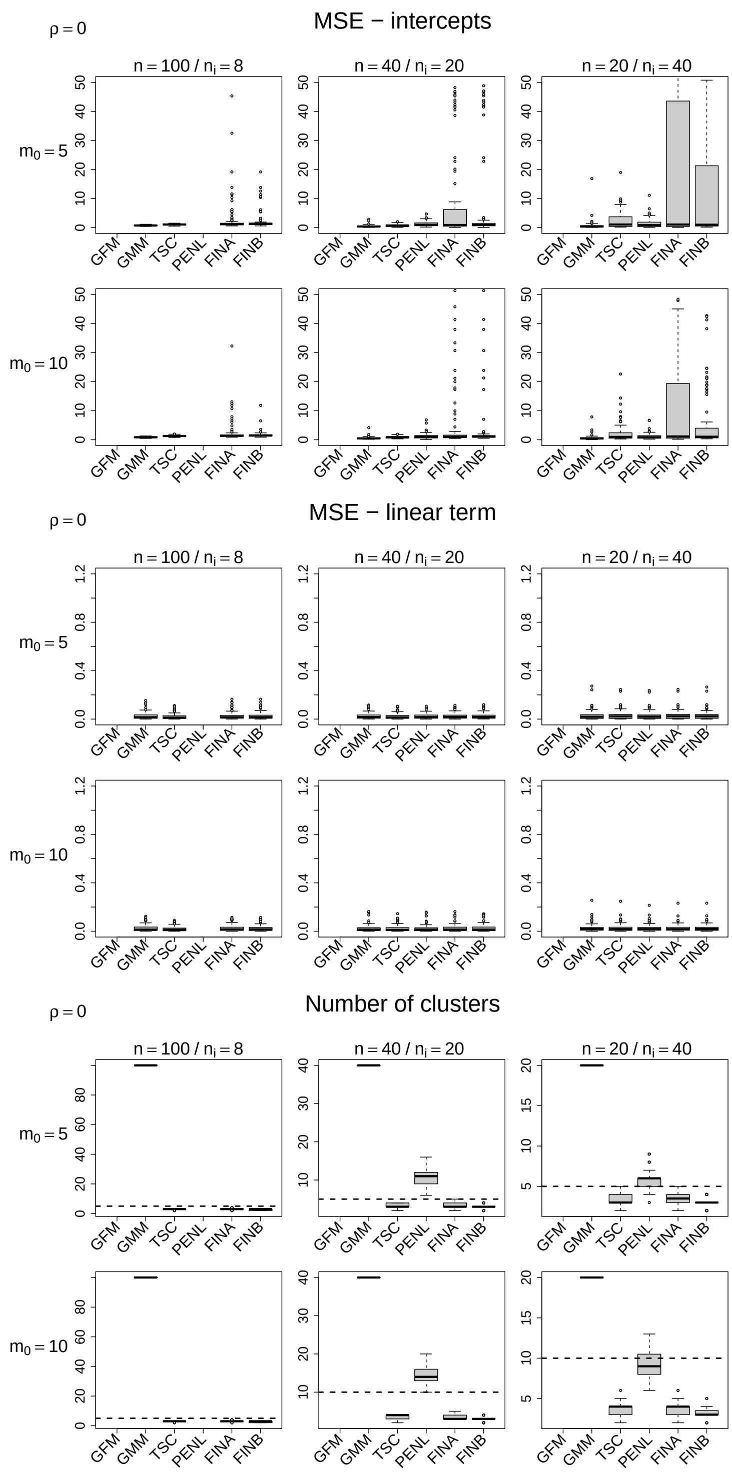

FIGURE 12: *Results for the settings with binary response, normally distributed intercepts and $\rho = 0$.*

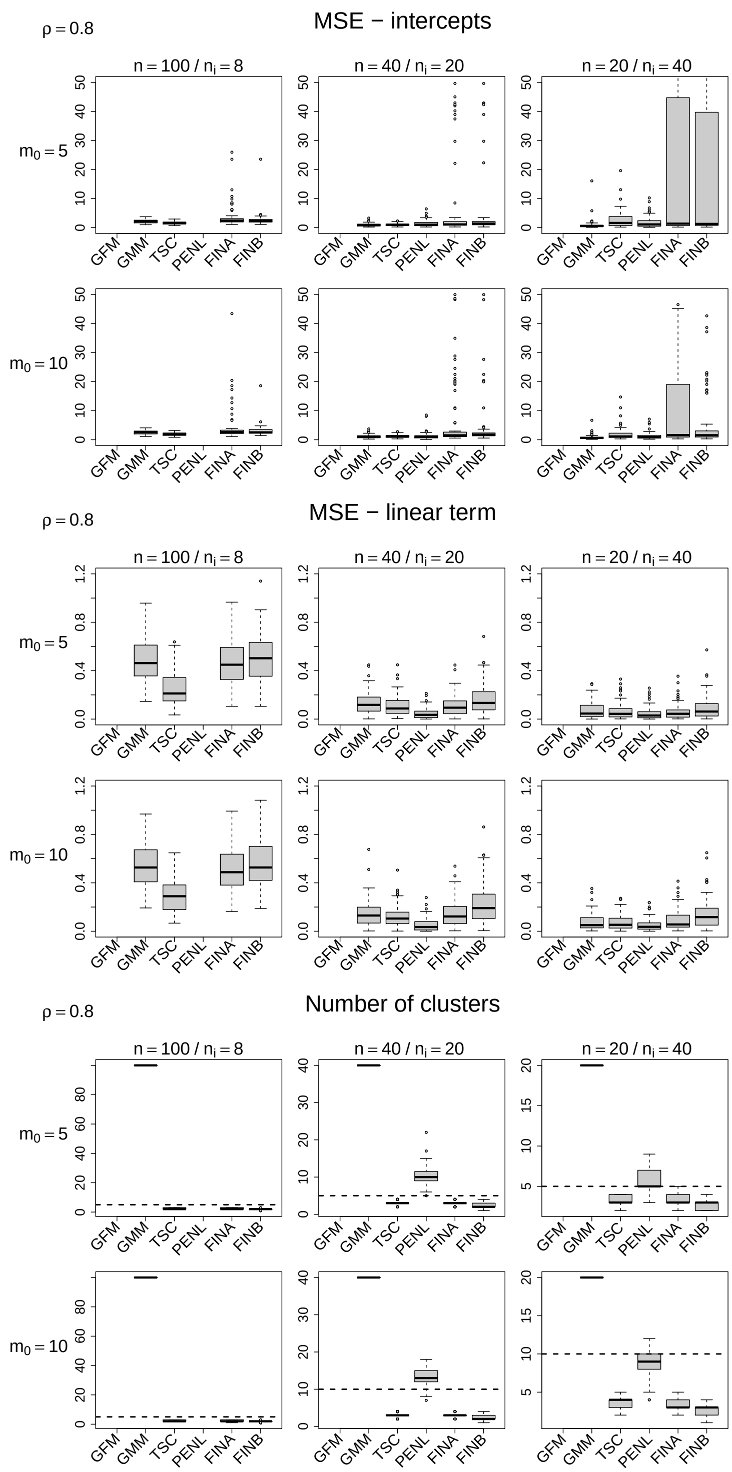

FIGURE 13: *Results for the settings with binary response, normally distributed intercepts and $\rho = 0.8$.*

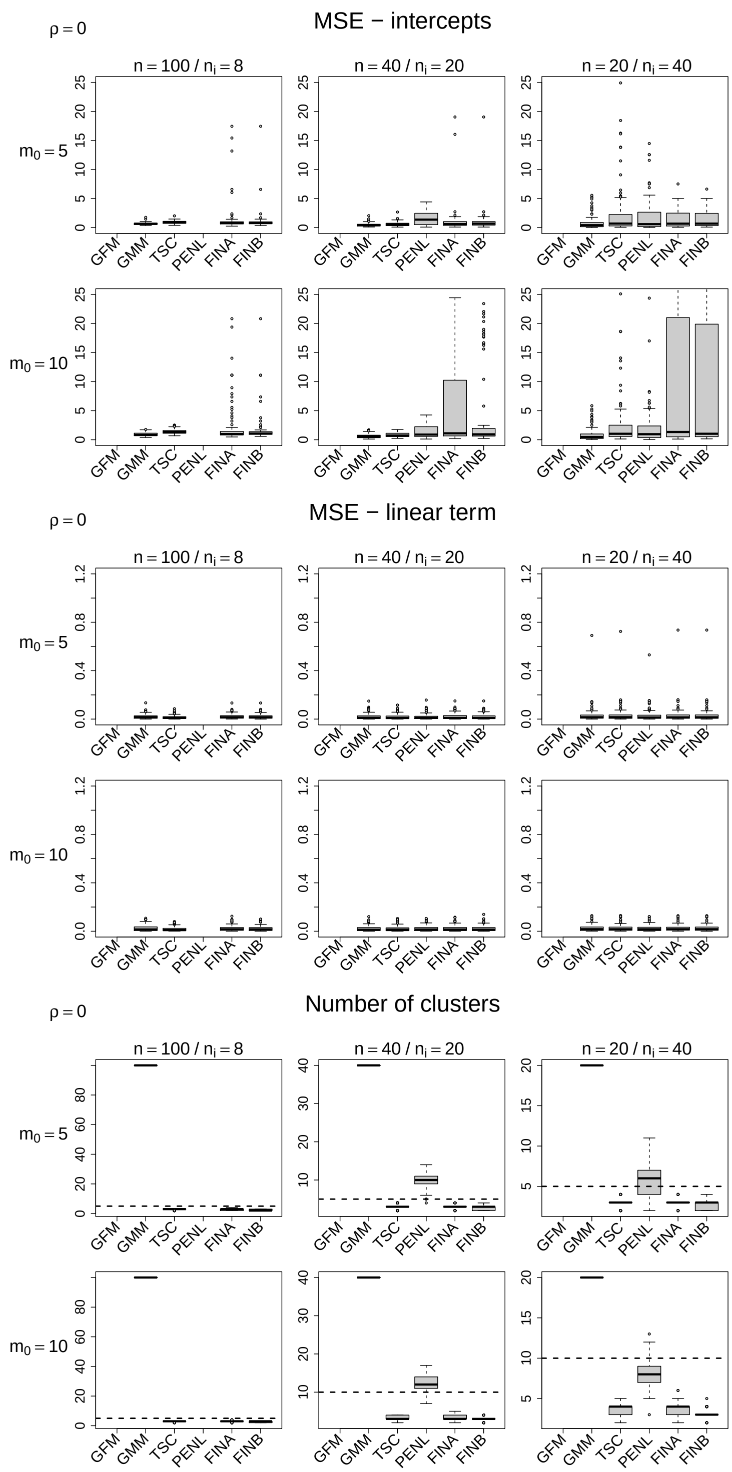

FIGURE 14: *Results for the settings with binary response, chi-squared distributed intercepts and* $\rho = 0$.

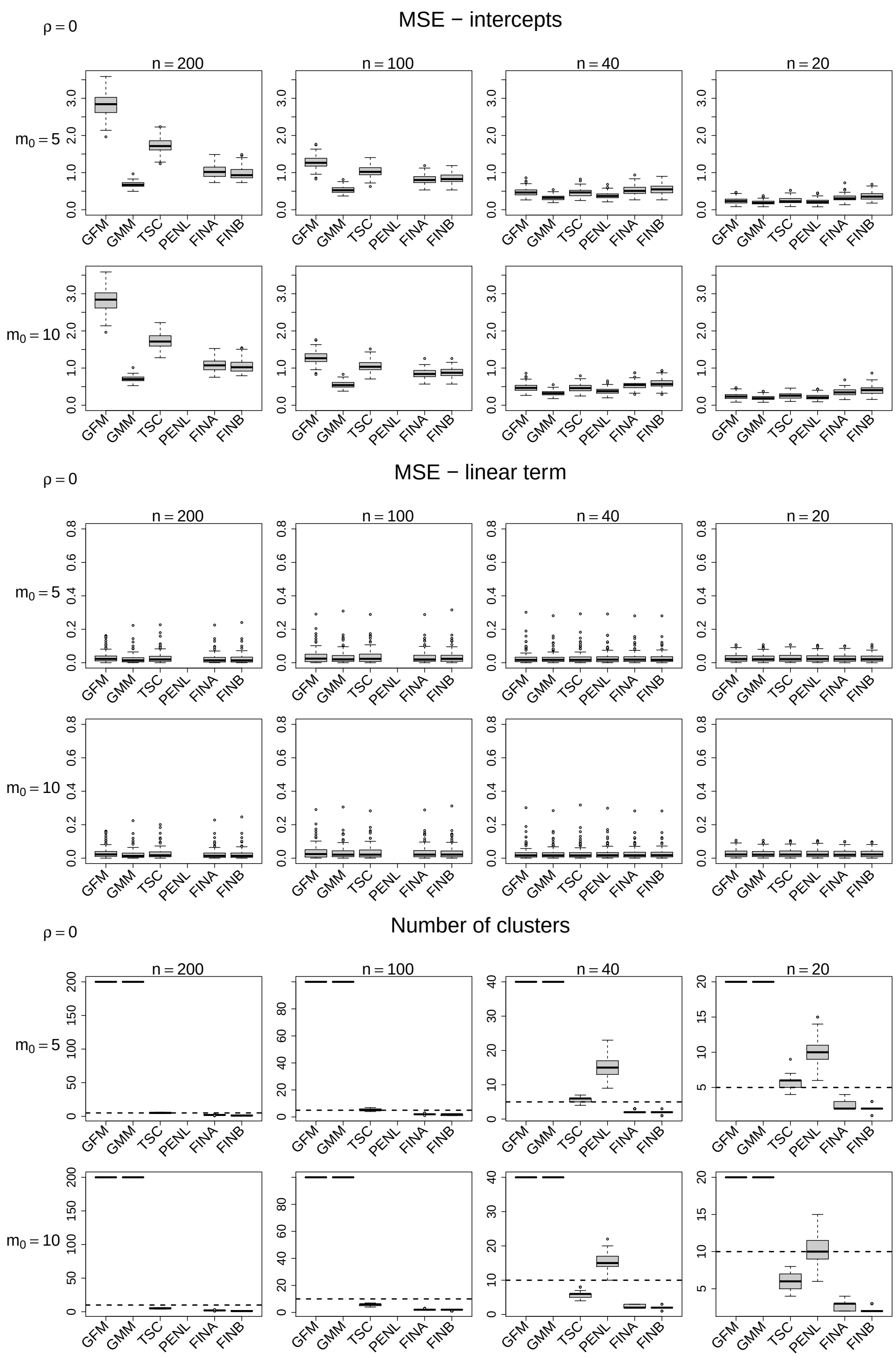

Figure 15: *Results for the settings with normal response, normally distributed intercepts and $\rho = 0$ with unbalanced design.*

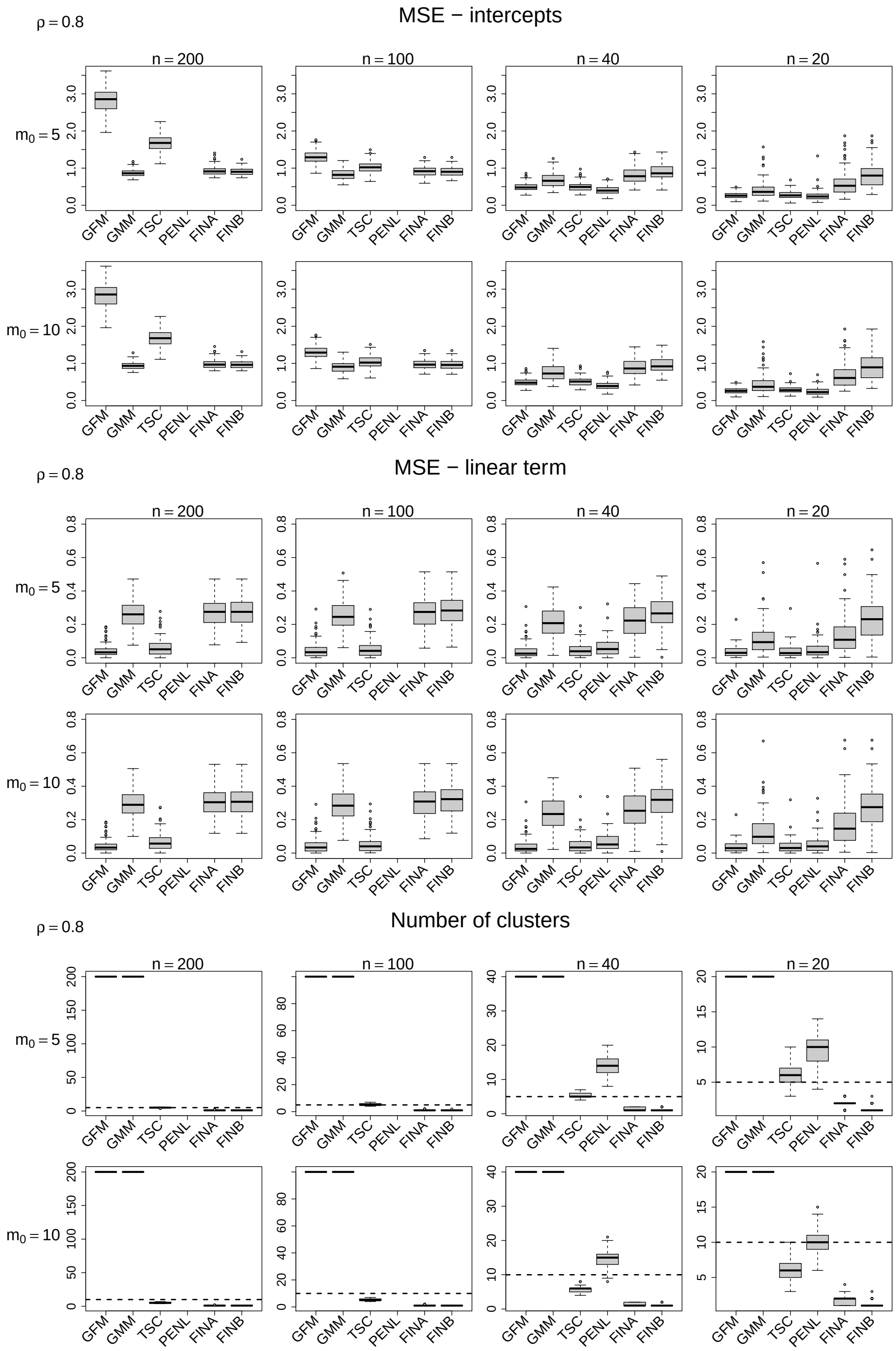

FIGURE 16: *Results for the settings with normal response, normally distributed intercepts and $\rho = 0.8$ with unbalanced design.*

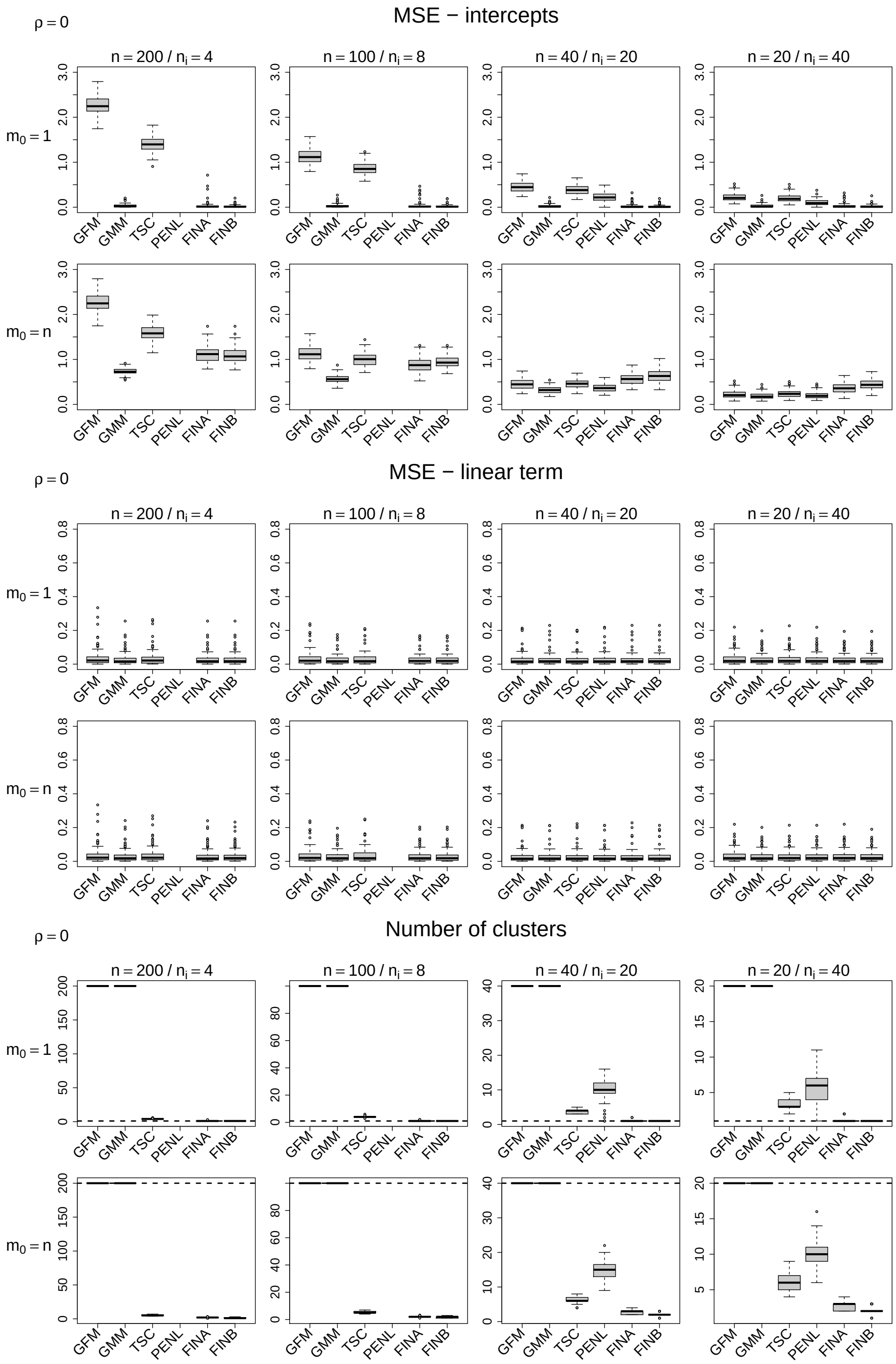

FIGURE 17: *Results for the extreme settings with normal response, normally distributed intercepts and $\rho = 0$.*

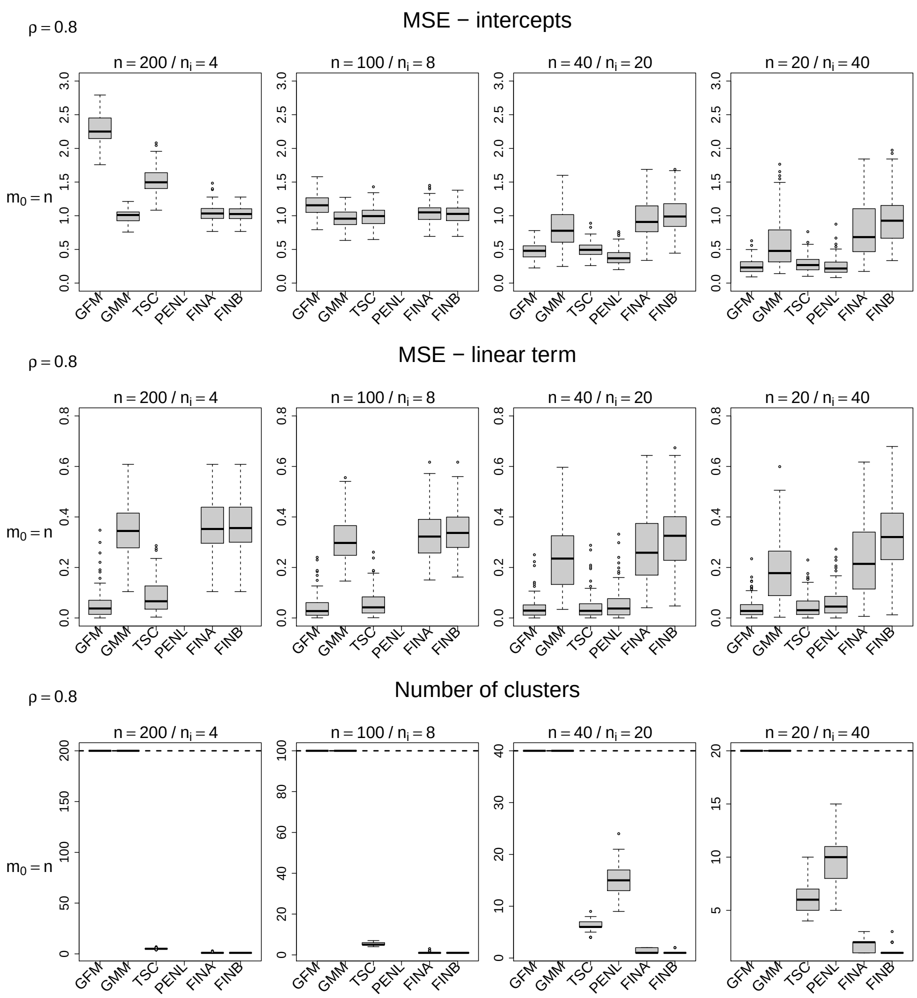

FIGURE 18: *Results for the extreme settings with normal response, normally distributed intercepts and $\rho = 0.8$.*

Table 5: *Average FPR and FNR for the settings with normal response and normally distributed intercepts.*

		$\rho = \mathbf{0}$				$\rho = \mathbf{0.8}$			
		$m_0 = 5$		$m_0 = 10$		$m_0 = 5$		$m_0 = 10$	
		FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR
$n = 200$	GFM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
$n_i = 4$	GMM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	TSC	0.70	0.23	0.70	0.24	0.70	0.24	0.71	0.24
	PENL								
	FINA	0.28	0.64	0.28	0.64	0.03	0.96	0.03	0.97
	FINB	0.11	0.85	0.12	0.84	0.00	1.00	0.00	1.00
$n = 100$	GFM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
$n_i = 8$	GMM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	TSC	0.68	0.21	0.69	0.22	0.69	0.22	0.69	0.23
	PENL								
	FINA	0.29	0.54	0.29	0.57	0.07	0.92	0.06	0.94
	FINB	0.19	0.69	0.21	0.68	0.00	1.00	0.00	0.99
$n = 40$	GFM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
$n_i = 20$	GMM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	TSC	0.65	0.16	0.64	0.18	0.66	0.17	0.67	0.18
	PENL	0.83	0.08	0.83	0.09	0.81	0.10	0.82	0.11
	FINA	0.25	0.47	0.26	0.49	0.14	0.79	0.13	0.83
	FINB	0.19	0.56	0.21	0.56	0.02	0.96	0.00	1.00
$n = 20$	GFM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
$n_i = 40$	GMM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	TSC	0.57	0.13	0.56	0.17	0.60	0.14	0.60	0.17
	PENL	0.75	0.09	0.74	0.11	0.74	0.11	0.74	0.12
	FINA	0.18	0.40	0.20	0.43	0.15	0.67	0.14	0.74
	FINB	0.12	0.52	0.12	0.56	0.03	0.91	0.02	0.95

Table 6: *Average FPR and FNR for the settings with normal response and chi-squared distributed intercepts.*

		$\rho = \mathbf{0}$				$\rho = \mathbf{0.8}$			
		$m_0 = 5$		$m_0 = 10$		$m_0 = 5$		$m_0 = 10$	
		FPR	FNR	FPR	FNR	FPR	FNR	FPR	FNR
$n = 200$	GFM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
$n_i = 4$	GMM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	TSC	0.71	0.25	0.70	0.25	0.71	0.25	0.71	0.24
	PENL								
	FINA	0.12	0.86	0.15	0.81	0.04	0.95	0.06	0.94
	FINB	0.03	0.96	0.05	0.93	0.00	1.00	0.00	0.99
$n = 100$	GFM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
$n_i = 8$	GMM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	TSC	0.69	0.24	0.69	0.24	0.69	0.24	0.68	0.25
	PENL								
	FINA	0.16	0.76	0.11	0.81	0.08	0.90	0.06	0.93
	FINB	0.05	0.90	0.05	0.89	0.00	1.00	0.00	0.99
$n = 40$	GFM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
$n_i = 20$	GMM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	TSC	0.66	0.22	0.64	0.24	0.65	0.23	0.65	0.24
	PENL	0.79	0.15	0.79	0.16	0.79	0.15	0.79	0.16
	FINA	0.12	0.67	0.10	0.74	0.11	0.80	0.07	0.87
	FINB	0.06	0.77	0.04	0.83	0.03	0.92	0.02	0.96
$n = 20$	GFM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
$n_i = 40$	GMM	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	TSC	0.61	0.21	0.60	0.23	0.62	0.22	0.60	0.24
	PENL	0.68	0.22	0.69	0.22	0.69	0.21	0.68	0.22
	FINA	0.07	0.64	0.09	0.69	0.09	0.73	0.08	0.79
	FINB	0.03	0.72	0.04	0.78	0.04	0.86	0.03	0.92

Extension with Constant Covariates

As noted in Section 4, the tree based approach can be seen as a regularization method that is able to also fit models with constant covariates $\boldsymbol{x}_i$. By using the tree-structured model one can separate the clustered intercepts $\beta_{i0}^{(k)}$ from the effects $\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\gamma}$ if they are not collinear.

However, for the inclusion of constant covariates the basic algorithm described in Section 4 has to be modified such that the fixed effects model is not needed during the fitting procedure. Therefore, the order of units is defined by the estimation of random intercepts instead of fixed intercepts. In addition, the null hypothesis to obtain a splitting decision is tested by using the score test instead of the likelihood-ratio test.

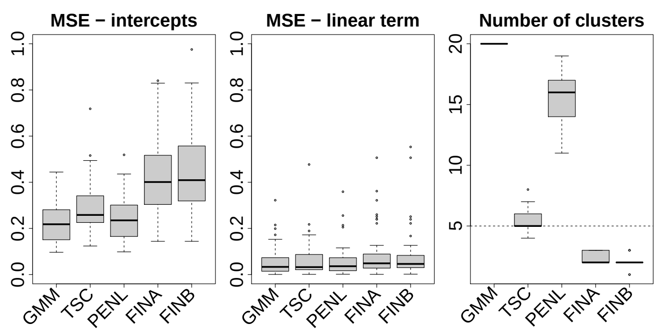

FIGURE 19: *Results of the simulation with a covariate that is constant across observation units. The performance of the proposed TSC is quite satisfactory.*

To illustrate the potential of the method to analyse this kind of data we show the results of a small simulation. We consider data with $n = 20$, $n_i = 40$, $m_0 = 5$ and normally distributed responses y_{ij} with error variance $\sigma_\epsilon^2 = 3^2$. The linear predictors are composed of normally distributed, clustered intercepts, one continuous covariate x_1 with $x_{ij1} \sim N(0, 1)$, one binary covariate x_2 with $x_{ij2} \sim B(1, 0.5)$ and one continuous covariate x_3 with $x_{i3} \sim N(0, 1)$. That means x_3 is the same for each observation belonging to one unit. This variable can, for example, represent characteristics of persons, like height or weight, that do not change over measurements. The true coefficients are $\beta_1 = \beta_2 = \beta_3 = 1$. There is no correlation between the unit-specific intercepts and the covariates.

It is seen from the results in Figure 19 that the performance of the tree-structured model is quite satisfactory. The linear term is estimated with sufficient accuracy by all the approaches. Regarding the clustering, the tree-structured model is very close to the true number of clusters, whereas the penalty approach forms many small clusters and the finite mixture model again forms very few large clusters. Certainly, further research is needed to investigate the performance if constant covariates are present, in particular the case of categorical covariates raises problems since then the clusters can be strongly linked to the categorical variables and identifiability might be endangered.