Sparse Bayesian modelling of underreported count data

Supplement

Michaela Dvorzak and Helga Wagner

May 3, 2015

A Identifiability (additional figures)

In Section 2.1 of the main paper, we illustrate non-identification in the Pogit model for simulated data and show that additional information on the reporting process, provided either through an informative prior distribution or a sample of validation data, allows parameter identification in a simple version of the Pogit model with only an intercept in both parts of the model. Here we show additional figures for Section 2.1 with varying the amount of information on the reporting process. We simulated the data with $\beta_0^{\text{true}} = 1$ and $\alpha_0^{\text{true}} = 1.8$ (which is similar to $\hat{\alpha}_0$ in Table 5 of Section 5 in the main paper for the cervical cancer data set). Observed data were generated from a Poisson distribution with $\mathrm{E}_i = 100$, $i = 1, \ldots, 1000$, $\lambda = \exp(\beta_0^{\text{true}})$ and $p = \text{logit}^{-1}(\alpha_0^{\text{true}})$. In the following figures, we plot the scaled log-posterior distribution $\log p(\beta_0, \alpha_0|\mathbf{y})^* = \max(\log p(\beta_0, \alpha_0|\mathbf{y}))/\log p(\beta_0, \alpha_0|\mathbf{y})$. Figure A1 shows contour and surface plots of the scaled log-posterior of the parameters (β_0, α_0) with varying the amount of prior information

$$p(\alpha_0) \sim \mathcal{N}(\alpha_0^{\text{true}}, M_{\alpha,0}),$$

in terms of the prior variance $M_{\alpha,0}$. As the prior variance $M_{\alpha,0}$ and thus, the uncertainty on α_0 decreases, the shape of the resulting posterior distribution becomes more and more elliptical.
Figure A2 shows again contour and surface plots of the scaled log-posterior of (β_0, α_0), but now for varying sample sizes of validation data $m = 10, 50, 100$ (top) and $m = 250, 500, 1000$ (bottom).

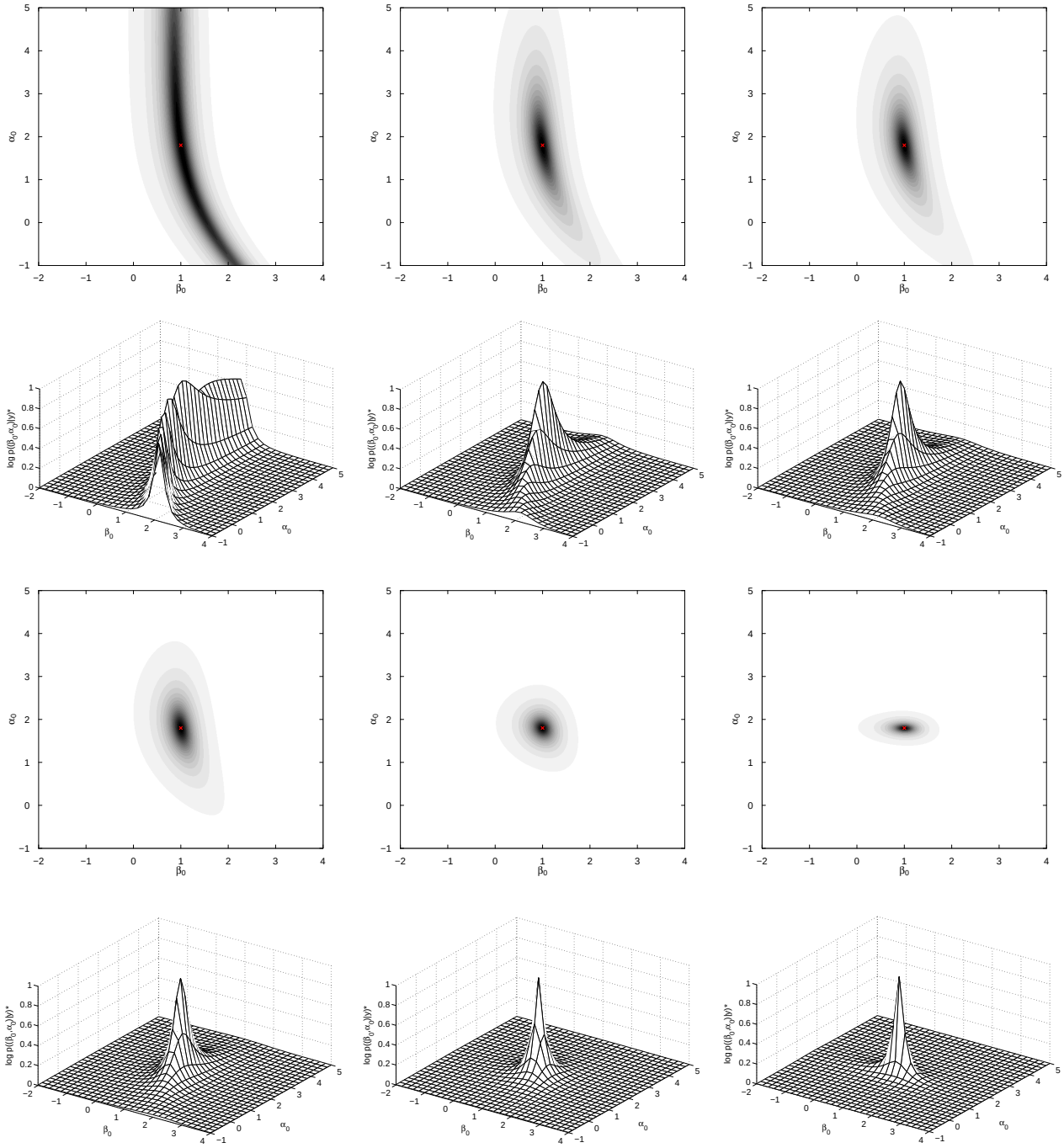

Figure A1: Contour and surface plots of the (scaled) log-posterior $\log p(\beta_0, \alpha_0|\mathbf{y})^*$ for a Pogit intercept model ($\beta_0^{\text{true}} = 1$, $\alpha_0^{\text{true}} = 1.8$) with varying prior information $p(\alpha_0) \sim \mathcal{N}(\alpha_0^{\text{true}}, M_{\alpha,0})$; top: $M_{\alpha,0} = 0.05^2$ (left), $M_{\alpha,0} = 0.01^2$ (middle) and $M_{\alpha,0} = 0.0075^2$ (right); bottom: $M_{\alpha,0} = 0.005^2$ (left), $M_{\alpha,0} = 0.0025^2$ (middle) and $M_{\alpha,0} = 0.001^2$ (right).

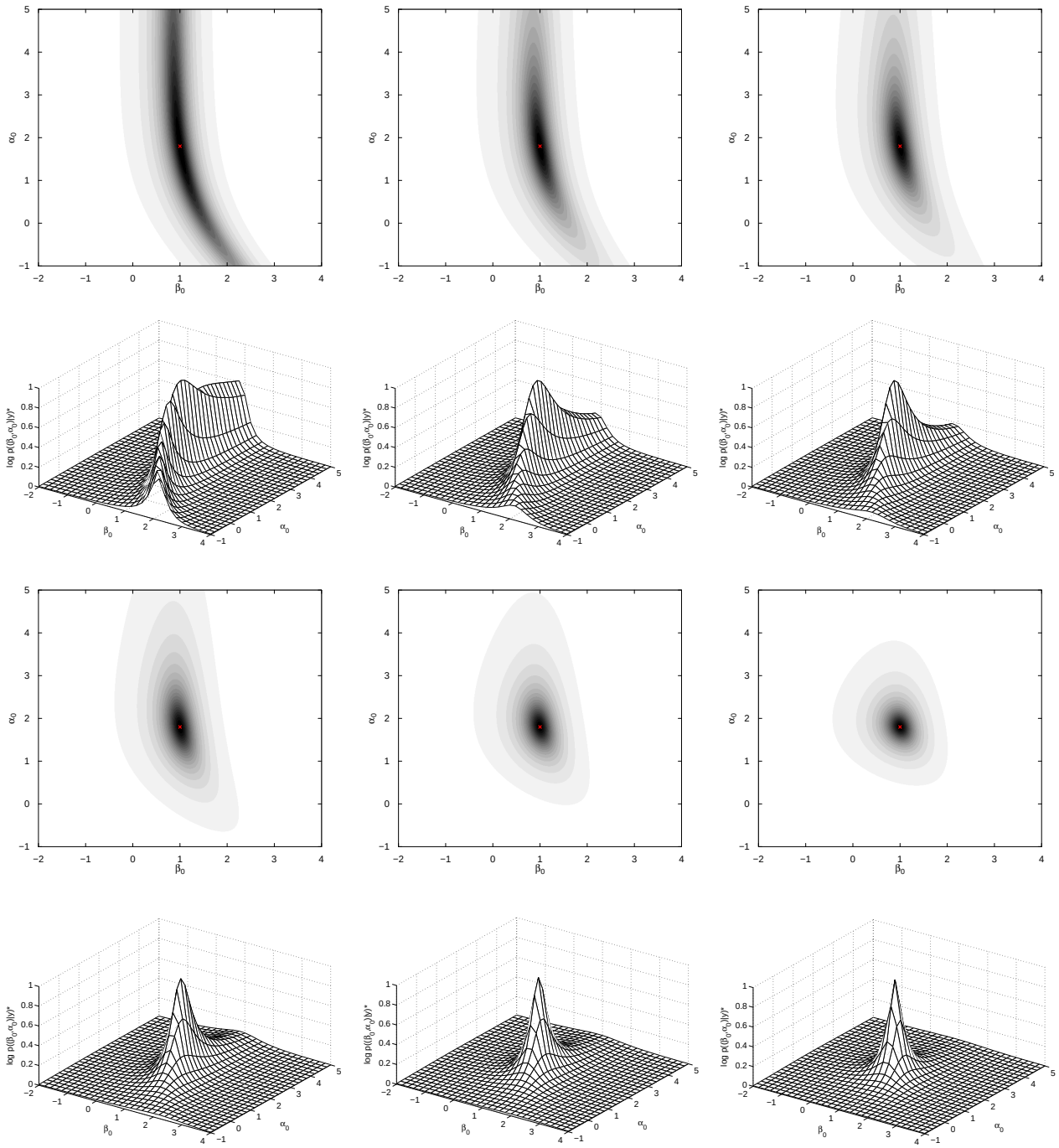

Figure A2: Contour and surface plots of the (scaled) log-posterior $\log p(\beta_0, \alpha_0|\mathbf{y})^*$ for a Pogit intercept model ($\beta_0^{\text{true}} = 1$, $\alpha_0^{\text{true}} = 1.8$) with varying sample size m of validation data; top: $m = 10$ (left), $m = 50$ (middle) and $m = 100$ (right); bottom: $m = 250$ (left) $m = 500$ (middle) and $m = 1000$ (right).

B Further simulation studies

B.1 Sensitivity to the prior distribution (Simulation I)

To investigate how sensitive inference on model parameters is with respect to the prior distribution, we vary the slab variance for the regression effects in simulation I (see Section 4.1 of the main paper) by choosing different values for the hyper-parameter Q. We fix the shape parameter to $\nu = 5$, but instead of our standard choice $Q^* = 20$, we use $Q_1 = 5$, $Q_2 = 80$ and $Q_3 = 10000$. These prior settings correspond to slab variances of $V_1 = 1.25$, $V^* = 5$, $V_2 = 20$ and $V_3 = 2500$, respectively. With these choices, the marginal prior probability for a regression effect to be in the interval $[-2, 2]$ is 0.96 (for Q_1), 0.83 (for Q^*), 0.69 (for Q_2) and 0.52 (for Q_3), respectively.

Table B1 reports the averaged posterior means and the nominal coverage of 95%-HPD intervals of the regression parameters in both sub-models over 300 data sets for different values of Q. To compare model selection performance, we report averaged posterior inclusion probabilities and the number of data sets where a regressor is included in the final model in Table B2.

Table B1: Simulation I. Averaged posterior means (Av. Pm.) and nominal coverage of (model averaged) 95%-HPD intervals of the regression effects over 300 data sets for different hyper-parameters Q of slab variances.

		Av. Pm.				Cov. of 95%-HPD intervals			
	True	$Q_1 = 5$	$Q^* = 20$	$Q_2 = 80$	$Q_3 = 10000$	$Q_1 = 5$	$Q^* = 20$	$Q_2 = 80$	$Q_3 = 10000$
α_0	2.2	2.182	2.237	2.256	2.258	0.95	0.95	0.92	0.92
α_1	-1.9	-1.838	-1.922	-1.954	-1.956	0.95	0.94	0.94	0.93
α_2	0	-0.021	-0.010	-0.006	-0.001	1.00	1.00	1.00	1.00
α_3	0	0.004	0.003	0.002	0.000	1.00	1.00	1.00	1.00
α_4	0	0.007	0.005	0.003	0.001	1.00	1.00	1.00	1.00
β_0	0.75	0.755	0.750	0.748	0.749	0.92	0.92	0.90	0.89
β_1	0.5	0.478	0.496	0.503	0.503	0.95	0.96	0.96	0.94
β_2	-2.0	-1.991	-2.000	-2.003	-2.005	0.97	0.97	0.96	0.96
β_3	0	-0.001	0.000	0.000	0.000	1.00	1.00	1.00	1.00
β_4	0	-0.002	-0.001	-0.001	0.000	1.00	1.00	1.00	1.00

We find that the prior has a minor effect on the averaged posterior means as long as the variance of the slab is not too small (i.e. $V \geq 5$). However, the nominal coverage of 95%-HPD intervals slightly decreases as the slab variance increases.

Figure B3 shows the estimated posterior inclusion probabilities for the different prior settings. We observe that the posterior inclusion probabilities are slightly affected by the prior and decrease with increasing variance of the slab component, which is the expected behaviour, see Malsiner-Walli and Wagner (2011). Thus, as shown in Table B2, classification is less perfect for a small slab variance (e.g. for Q_1), where the false discovery rate of zero effects is considerably higher than for large slab variances. In contrast, if the slab variance gets rather large (e.g. for Q_3), non-zero effects are more likely not to be detected due to the generally smaller posterior inclusion probabilities: Based on posterior inclusion probabilities above 0.5, in two data sets in the logit sub-model and in three data sets in the Poisson sub-model one of the non-zero effects is misclassified as zero.

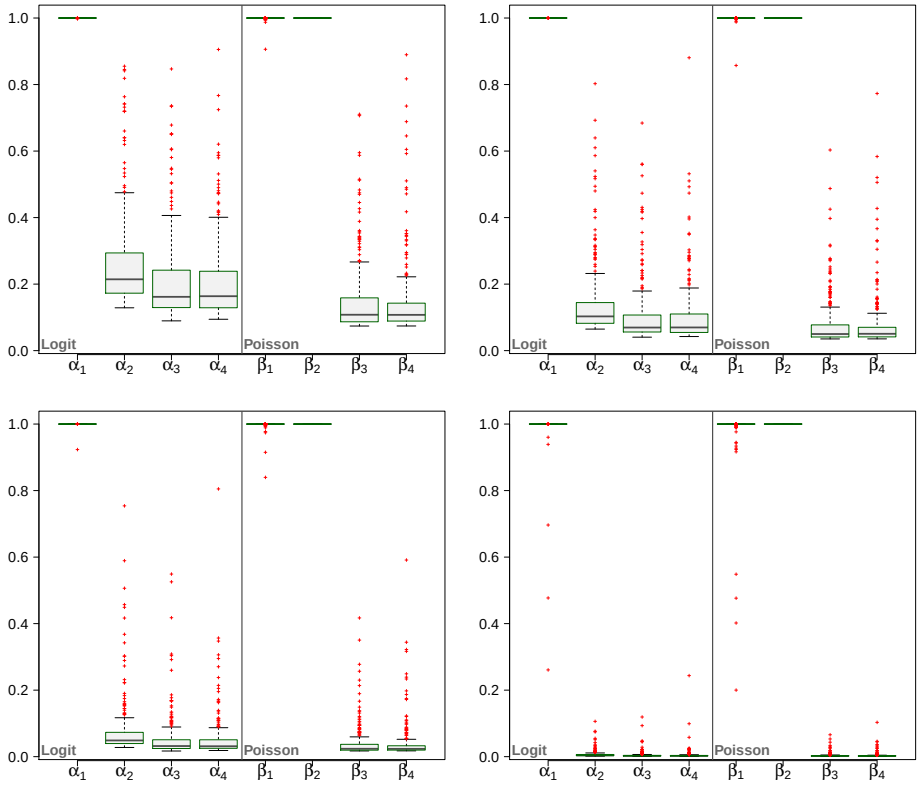

Figure B3: Simulation I. Posterior inclusion probabilities over 300 data sets for different hyper-parameters Q of slab variances; top: $Q_1 = 5$ (left-hand side), $Q^* = 20$ (right-hand side); bottom: $Q_2 = 80$ (left-hand side), $Q_3 = 10000$ (right-hand side); ν fixed.

Table B2: Simulation I. Averaged posterior inclusion probabilities (Av. $\hat{p}_{*,j}$) and number of models ($\mathrm{N_{mod}}$) where a regressor is included in the final model over 300 data sets for different hyper-parameters Q of slab variances. Results for non-zero observations are given in bold.

	Av. $\hat{p}_{*,j}$				$\mathrm{N_{mod}}$			
	$Q_1 = 5$	$Q^* = 20$	$Q_2 = 80$	$Q_3 = 10000$	$Q_1 = 5$	$Q^* = 20$	$Q_2 = 80$	$Q_3 = 10000$
α_1	**1.000**	**1.000**	**1.000**	**0.994**	**300**	**300**	**300**	**298**
α_2	0.263	0.140	0.075	0.007	20	8	3	0
α_3	0.210	0.103	0.051	0.005	15	4	2	0
α_4	0.207	0.102	0.051	0.005	11	3	1	0
β_1	**1.000**	**0.999**	**0.999**	**0.990**	**300**	**300**	**300**	**297**
β_2	**1.000**	**1.000**	**1.000**	**1.000**	**300**	**300**	**300**	**300**
β_3	0.147	0.077	0.039	0.004	6	1	0	0
β_4	0.145	0.076	0.039	0.004	8	4	1	0

As the misclassification of a truly zero effect as non-zero affects only efficiency of the corresponding parameter, whereas falsely classifying a non-zero effect as zero leads to heavy shrinkage of this effect to zero and hence, might result in biased estimates, we do not recommend very large values of Q. From these simulation results, we conclude that our prior choice is reasonable with respect to parameter estimation, nominal coverage and model selection performance. Moreover, similar priors

are used in Fahrmeir et al. (2010) and Wagner and Duller (2012).

B.2 Sensitivity to the prior distribution (Simulation II)

Similar to the first simulation study in Section B.1, we here examine sensitivity of inference with respect to the prior distribution for simulation II (see Section 4.2 of the main paper). The simulation setup is the same as in simulation II and considers clustered observations. The slab variance for the regression effects is again modified in terms of different hyper-parameters Q, while ν is fixed to 5. Additionally to our standard choice $Q^* = 20$ ($V^* = 5$), we use $Q_1 = 5$ ($V_1 = 1.25$), $Q_2 = 80$ ($V_2 = 20$) and $Q_3 = 10000$ ($V_3 = 2500$).

Table B3 reports the averages of the posterior means and the nominal coverage of 95%-HPD intervals over 50 data sets for each regression effect as well as for the random intercept parameter θ under the different prior specifications. The respective model selection performance is summarized in Table B4, showing the averages of the posterior inclusion probabilities and the number of models where a regressor is included in the final model over all data sets.

Table B3: Simulation II. Averaged posterior means (Av. Pm.) and nominal coverage of (model averaged) 95%-HPD intervals of the regression effects over 50 data sets for different hyper-parameters Q of slab variances.

		Av. Pm.				Cov. of 95%-HPD intervals			
	True	$Q_1 = 5$	$Q^* = 20$	$Q_2 = 80$	$Q_3 = 10000$	$Q_1 = 5$	$Q^* = 20$	$Q_2 = 80$	$Q_3 = 10000$
α_0	2.2	2.202	2.201	2.201	2.202	0.98	0.96	0.96	0.96
α_1	-0.3	-0.296	-0.296	-0.296	-0.296	0.96	0.96	0.96	0.96
α_2	0	0.000	0.001	0.001	0.000	1.00	1.00	1.00	1.00
α_3	-0.3	-0.304	-0.304	-0.305	-0.306	0.96	0.94	0.96	0.94
α_4	0	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00
θ_α	0.3	0.299	0.299	0.299	0.299	0.98	0.98	0.98	0.98
β_0	0.75	0.750	0.750	0.750	0.750	0.94	0.94	0.94	0.94
β_1	0.1	0.100	0.100	0.100	0.100	0.96	0.98	0.98	0.96
β_2	0.1	0.101	0.101	0.101	0.101	0.96	0.94	0.94	0.96
β_3	0	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00
β_4	0	0.000	0.000	0.000	0.000	1.00	1.00	1.00	1.00
θ_β	0.1	0.103	0.102	0.102	0.103	0.98	0.96	0.98	0.96

From Table B3, we conclude that compared to simulation I, the prior has only a minor effect in this simulation study: The averaged posterior means as well as the coverage of 95%-HPD intervals are hardly affected by the choice of the slab variance. This is not unexpected as the number of observations (though clustered) is much higher than in simulation I.

Considering model selection performance, the results in Table B4 and Figure B4 in turn demonstrate that the posterior inclusion probabilities decrease with increasing variance of the slab component. Consequently, misclassifications of zero effects are more likely for small slab variances but still, the differences between the prior settings are negligible. Furthermore, non-zero effects are perfectly classified as being non-zero for all data sets and slab variances, even for the very large variance Q_3.

Table B4: Simulation II. Averaged posterior inclusion probabilities (Av. $\hat{p}_{*,j}$) and number of models ($\mathrm{N_{mod}}$) where a regressor is included in the final model over 50 data sets for different hyper-parameters Q. Results for non-zero observations are given in bold.

	Av. $\hat{p}_{*,j}$				$\mathrm{N_{mod}}$			
	$Q_1 = 5$	$Q^* = 20$	$Q_2 = 80$	$Q_3 = 10000$	$Q_1 = 5$	$Q^* = 20$	$Q_2 = 80$	$Q_3 = 10000$
α_1	**1.000**	**1.000**	**1.000**	**1.000**	**50**	**50**	**50**	**50**
α_2	0.134	0.083	0.046	0.004	3	2	0	0
α_3	**1.000**	**1.000**	**1.000**	**1.000**	**50**	**50**	**50**	**50**
α_4	0.078	0.039	0.019	0.002	0	0	0	0
θ_α	**1.000**	**1.000**	**1.000**	**1.000**	**50**	**50**	**50**	**50**
β_1	**1.000**	**1.000**	**1.000**	**1.000**	**50**	**50**	**50**	**50**
β_2	**1.000**	**1.000**	**1.000**	**1.000**	**50**	**50**	**50**	**50**
β_3	0.039	0.025	0.016	0.002	1	1	0	0
β_4	0.020	0.009	0.005	0.000	0	0	0	0
θ_β	**1.000**	**1.000**	**1.000**	**1.000**	**50**	**50**	**50**	**50**

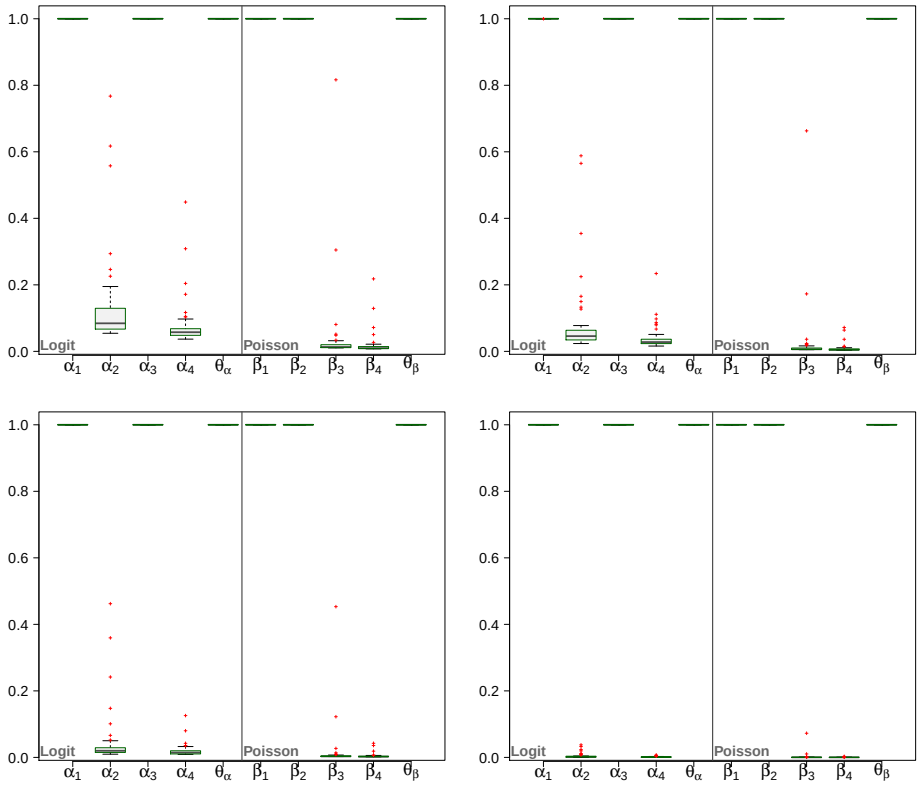

Figure B4: Simulation II. Posterior inclusion probabilities over 50 data sets for different hyper-parameters Q of slab variances; top: $Q_1 = 5$ (left-hand side), $Q^* = 20$ (right-hand side); bottom: $Q_2 = 80$ (left-hand side), $Q_3 = 10000$ (right-hand side); ν fixed.

B.3 Simulation II with partial validation data

Whereas in the simulation study in Section B.2 we assumed that validation data are available for every cluster, we now follow the suggestion of a referee and investigate how parameter estimation is affected if validation data are provided only for certain clusters. We use the same setup as in simulation II (see Section 4.2 of the main paper) except that we consider a different validation sample and assume that validation data are available only for the first 10 clusters but not for the remaining 40 clusters. To have a validation sample of the same size, we assume that $m_i = 125$ for each of the covariate patterns in the first 10 clusters, i.e. for $i = 1, \ldots, 160$. In contrast, no validation data are available for the remaining clusters, i.e. $m_i = 0$ for $i = 161, \ldots, 800$.
Table B5 shows the averages of the posterior estimates and the averaged posterior inclusion probabilities (over 50 data sets) for each regressor and the respective random intercept parameter. Comparing the results to those in Table 2 of Section 4.2 in the main paper, we find that the resulting parameter estimates are impaired only slightly if validation data are not available for every cluster. Similarly, the differences regarding model selection performance are negligible. However, the nominal coverage of 95%-HPD intervals for the intercepts α_0 and β_0 as well as for the random intercept parameter θ_α suffers somehow from the lack of information in the remaining clusters.

Table B5: Simulation II with (partial) validation data for clusters $c = 1, \ldots, 10$. Averaged posterior means (Av. Pm.), averaged estimated posterior inclusion probabilities (Av. $\hat{p}_{*,j}$), nominal coverage of (model averaged) 95%-HPD intervals (Cov.) as well as number of models ($\mathrm{N_{mod}}$) where a regressor is included in the final model over 50 data sets. Results for non-zero effects are given in bold.

Model	Covariate	True value	Av. Pm.	Av. $\hat{p}_{*,j}$	Cov.	$\mathrm{N_{mod}}$
Logit	α_0	2.2	2.202	-	0.92	-
	α_1	-0.3	-0.293	**1.000**	1.00	**50**
	α_2	0	0.000	0.046	1.00	0
	α_3	-0.3	-0.300	**1.000**	0.96	**50**
	α_4	0	0.000	0.036	1.00	0
	θ_α	0.3	0.310	**1.000**	0.92	**50**
Poisson	β_0	0.75	0.751	-	0.90	-
	β_1	0.1	0.100	**1.000**	1.00	**50**
	β_2	0.1	0.101	**1.000**	0.94	**50**
	β_3	0	-0.001	0.030	1.00	1
	β_4	0	0.000	0.008	1.00	0
	θ_β	0.1	0.100	**1.000**	0.96	**50**

Figure B5 illustrates the posterior estimates and 95%-HPD intervals of the random intercepts a_c, $c = 1, \ldots, 50$, in the logit sub-model for one data set with partial validation data (top panel) and complete validation data (bottom panel). As expected, intervals for clusters $c = 1, \ldots, 10$ with validation data are considerably narrower than those of the remaining clusters. Particularly, the latter contain zero for all random intercepts of clusters without validation information. Intervals are of similar length when validation data are complete, see the bottom panel of Figure B5. The respective results for the Poisson sub-model are shown in Figure B6.

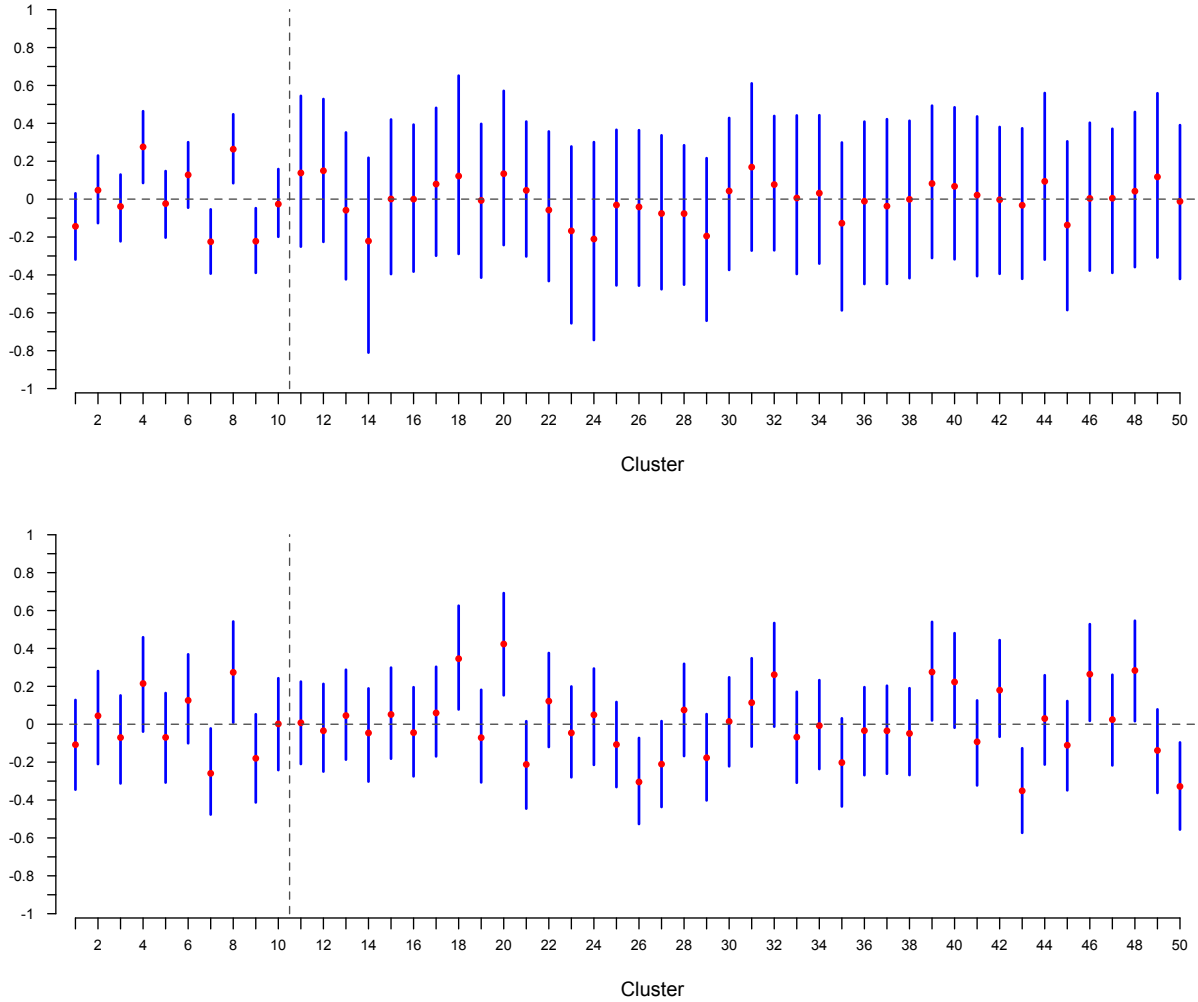

Figure B5: Simulation II. Posterior means and 95%-HPD intervals for random intercepts a_c, $c = 1, \ldots, 50$, in the logit part of the joint model for one data set; top: validation data for clusters $c = 1, \ldots, 10$ (*partial validation*); bottom: validation data for all $C = 50$ clusters (*complete validation*) based on the results of simulation II in Section 4.2 of the main paper.

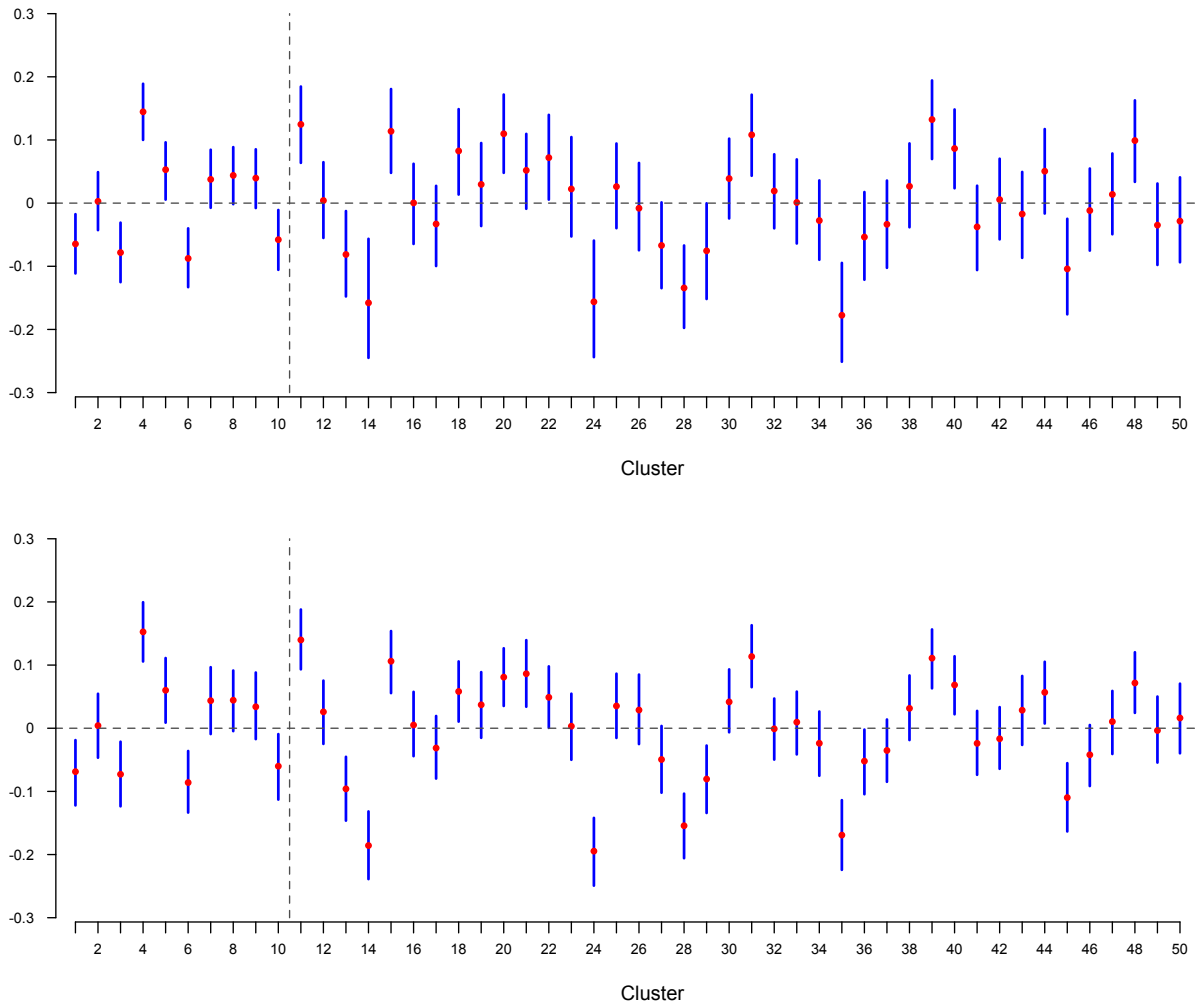

Figure B6: Simulation II. Posterior means and 95%-HPD intervals for random intercepts b_c, $c = 1, \ldots, 50$, in the Poisson part of the joint model for one data set; top: validation data for clusters $c = 1, \ldots, 10$ (*partial validation*); bottom: validation data for all $C = 50$ clusters (*complete validation*) based on the results of simulation II in Section 4.2 of the main paper.

C Supplementary material for the cervical cancer data (Section 5)

C.1 Sampling efficiency

To assess mixing and efficiency of MCMC sampling for the cervical cancer data, we report the effective sample size (ESS) and the integrated autocorrelation time (IAT). ESS estimates the equivalent number of independent draws corresponding to the dependent MCMC draws and is defined as $\text{ESS} = \text{M}/\tau_{\text{IAT}}$, where τ_{IAT} is the integrated autocorrelation time (IAT) and M is the number of MCMC iterations after the burn-in phase. IAT is computed as $\tau_{\text{IAT}} = 1 + 2\sum_{k=1}^{\text{K}} \rho(k)$ using the initial monotone sequence estimator (Geyer, 1992) for K, and $\rho(k)$ is the empirical autocorrelation at lag k.

Table C6 reports the IAT and the ESS for each regression effect and posterior inclusion probability in the Poisson as well as the logit sub-model with Belgium as the reference country. IATs and ESSs cannot be computed for posterior inclusion probabilities in clear-cut situations, i.e. where the posterior inclusion probability is equal to one in all M iterations (marked with a star in Table C6).

Table C6: Cervical cancer data. Integrated autocorrelation times (IAT) and effective sample sizes (ESS) of parameter estimates and inclusion probabilities in the Poisson and logit sub-model (with reference country Belgium).

		Effect		Inclusion	
Model		IAT	ESS	IAT	ESS
Poisson	β_0	7.4	4027.8	-	-
	β_1	12.9	2332.5	1.0	30002.8
	β_2	8.8	3393.7	11.2	2671.7
	β_3	13.6	2206.6	1.0	30002.0
	β_4	1.1	28073.1	$*$	$*$
	β_5	1.6	18334.8	$*$	$*$
	β_6	3.9	7701.7	$*$	$*$
	β_7	1.1	26361.0	1.4	21265.2
	β_8	1.2	26004.4	1.4	22213.1
	β_9	3.3	9039.4	3.5	8606.4
	β_{10}	4.1	7309.2	6.0	5008.3
	β_{11}	16.7	1795.2	11.0	2734.6
	β_{12}	20.2	1488.2	17.5	1716.9
	β_{13}	1.0	29119.9	1.3	23889.6
	β_{14}	1.2	24563.5	1.4	21800.1
	β_{15}	3.0	9851.7	3.9	7646.6
Logit	α_0 (Belgium)	24.4	1227.2	-	-
	α_1 (England)	946.6	31.7	39.1	767.3
	α_2 (France)	48.6	616.9	53.0	566.0
	α_3 (Italy)	23.1	1296.4	13.1	2291.3

The results are based on M= 30000 iterations (keeping only every 10th iterate of 300000 draws) after a burn-in phase of 100000 draws. The ESSs are highest for regression effects with estimated posterior inclusion probabilities close to the boundaries zero and one. Hence, sampling efficiency in the Poisson sub-model is lowest (in terms of the largest values for IAT) for regression effects β_{11}

and β_{12} with posterior inclusion probabilities close to 0.5. As expected, IATs are higher in the logit model, especially for the effect α_1 (England), where separation is present due to perfect reporting in this country.

Sampling efficiency for the logit sub-model with reference category Belgium is compared to the respective results with reference category England in Table C7. We observe that the IATs are lower for all country effects (except for the baseline England) when Belgium is used as the reference category and conclude that it is preferable to use the country Belgium as the reference category.

Table C7: Cervical cancer data. Integrated autocorrelation times (IAT) and effective sample sizes (ESS) of parameter estimates and inclusion probabilities in the logit sub-model (with reference category England).

		Effect		Inclusion	
Model		IAT	ESS	IAT	ESS
Logit	α_0 (England)	442.8	67.8	-	-
	α_1 (Belgium)	195.2	153.7	*	*
	α_2 (France)	389.0	77.1	10.0	3002.0
	α_3 (Italy)	361.4	83.0	35.9	836.0

C.2 Results for the selected model

Based on posterior inclusion probabilities above 0.5, the selected model contains the covariates x_1, x_3 - x_6, x_{11}, x_{12} and x_{15} in the Poisson and w_1 and w_2 in the logit sub-model. Table C8 reports the posterior means and 95%-HPD intervals of this model and Table C9 shows integrated autocorrelation times and effective sample sizes of its parameters.

Table C8: Cervical cancer data. Posterior means with 95%-HPD intervals for the selected model (with reference country Belgium).

		Covariate	Post. mean	95%-HPD
Poisson	β_0	Intercept (Belgium; 25-34)	2.104	(1.964, 2.248)
	β_1	England	0.396	(0.288, 0.501)
	β_3	Italy	-0.980	(-1.078,-0.878)
	β_4	35-44	1.609	(1.493, 1.726)
	β_5	45-54	2.699	(2.586, 2.814)
	β_6	55-64	2.815	(2.700, 2.931)
	β_{11}	France 45-54	-0.246	(-0.357,-0.133)
	β_{12}	France 55-64	-0.161	(-0.271,-0.048)
	β_{15}	Italy 55-64	0.238	(0.134, 0.342)
Logit	α_0	Intercept (Belgium)	1.715	(1.210, 2.215)
	α_1	England	3.400	(0.561, 6.552)
	α_2	France	-0.728	(-1.217,-0.241)

Table C9: Cervical cancer data. Integrated autocorrelation times (IAT) and effective sample sizes (ESS) of parameter estimates in the Poisson and logit sub-model for the selected model (with reference country Belgium).

Model		IAT	ESS
Poisson	β_0	10.5	2857.0
	β_1	26.5	1130.4
	β_3	3.4	8949.4
	β_4	1.0	29207.2
	β_5	1.1	28443.4
	β_6	1.1	28063.4
	β_{11}	3.9	7755.4
	β_{12}	4.1	7325.6
	β_{15}	1.1	27021.6
Logit	α_0	25.2	1188.4
	α_1	1456.8	20.6
	α_2	10.4	2876.9

References

Fahrmeir, L., Kneib, T. and Konrath, S. (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, **20**, 203–219.

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science*, **7**, 473–483.

Malsiner-Walli, G. and Wagner, H. (2011). Comparing spike and slab priors for Bayesian variable selection. *Austrian Journal of Statistics*, **40**, 241–264.

Wagner, H. and Duller, C. (2012). Bayesian model selection for logistic regression models with random intercept. *Computational Statistics and Data Analysis*, **56**, 1256–1274.