

Appendix

Appendix

The following paragraphs contain additional information about the **simulation studies** that were covered in chapter ??.

Section ??: Correlated Predictor Variables

The simulation design was chosen in the following way:

- The design matrix $\mathbf{X} = (\mathbf{x}_1 \ \mathbf{x}_2 \ \mathbf{x}_3)$ is simulated from a three dimensional normal distribution $\mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu} = (-5 \ 2 \ 0)^T$ and covariance matrix $\begin{pmatrix} 1 & \rho & \rho \\ \rho & 3 & \rho \\ \rho & \rho & 5 \end{pmatrix}$. Hence, the dependence among the regressors is fully determined by the parameter ρ .
- The design matrix $\mathbf{Z} = (\mathbf{z}_1 \ \mathbf{z}_2)$ consists of linear combinations of the regressors \mathbf{x}_1 up to \mathbf{x}_3 , more specifically $\mathbf{z}_1 = 0.8 \cdot \mathbf{x}_1 + 0.2 \cdot \mathbf{x}_2$ and $\mathbf{z}_2 = \mathbf{x}_2 - 0.5 \cdot \mathbf{x}_3$. In both design matrices intercept columns are added for estimation purposes.
- The true coefficient vectors are given by $\boldsymbol{\beta} = (\beta_0 \ \beta_1 \ \beta_2 \ \beta_3)^T = (0 \ 3 \ -1 \ 1)^T$ and $\boldsymbol{\gamma} = (\gamma_0 \ \gamma_1 \ \gamma_2)^T = (0 \ 2 \ 0)^T$.
- The outcome variable \mathbf{y} is generated according to the correctly specified location-scale model $y_i \stackrel{iid}{\sim} \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta}, \exp(\mathbf{z}_i^T \boldsymbol{\gamma})^2)$ for $i = 1, \dots, n$ with sample size $n = 50$. Before data generation all columns in \mathbf{X} and \mathbf{Z} are standardized, i.e. mean-centered around 0 and scaled to unit variance.
- Three different values were chosen for $\rho \in \{0, -0.5, 0.9\}$ to compare the ‘nice’ case of uncorrelated predictors with the performance for negative and positive dependence. For each covariance structure the three models `mcmc_ridge()`, `mcmc()` and `lmls()` are fitted to the standardized covariates, where each Posterior Mean estimate from both of the Markov Chain Monte Carlo samplers is based on 10.000 samples.

Section ??: Sample Size

Section ??: Redundant Covariates

We again state the conditions that the simulation study is based on:

- The design matrix $\mathbf{X} = (\mathbf{1}_n \ \mathbf{x}_1 \ \dots \ \mathbf{x}_{20})$ consists of one intercept column plus 10 *pairs* of successive regressors, starting with the pair $(\mathbf{x}_1, \mathbf{x}_2)$. Each pair $(\mathbf{x}_i, \mathbf{x}_{i+1})$ for $i \in \{1, 3, \dots, 19\}$ is (independently from all remaining pairs) drawn from a bivariate normal distribution with mean vector $\boldsymbol{\mu} = (0 \ 0)^T$ and correlation matrix $\begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$.
- The design matrix $\mathbf{Z} = (\mathbf{1}_n \ \mathbf{x}_1 \ \mathbf{x}_3)$ is of minor interest in this case and consists of an intercept column plus two uncorrelated columns chosen from \mathbf{X} .
- The true coefficients of $\boldsymbol{\beta}$ are determined by the pattern $\beta_i = 0$, if i is even and $\beta_i = 1$, if i is odd. Thus, all covariates with even subscript are redundant, whereas those with odd subscript contribute to

\mathbf{y} . The true γ , again of minor interest here, is given by $\gamma = (0 \quad 1 \quad 1)^T$.

- The outcome variable \mathbf{y} is generated according to the correctly specified location-scale model $y_i \stackrel{iid}{\sim} \mathcal{N}(\mathbf{x}_i^T \beta, \exp(\mathbf{z}_i^T \gamma)^2)$ for $i = 1, \dots, n$, where the covariates are used on their original scale.
- The sample size $n = 50$ is deliberately chosen small compared to the number of regressors. Before fitting each of the three models, all columns of \mathbf{X} except the intercept column is standardized to zero mean and unit variance. Both of the Bayesian models generate 10.000 values for each coefficient.

Section ??: Challenging the Model Assumptions

The data for this simulation study is generated by the following conventions:

- The design matrix $\mathbf{X} = (\mathbf{1}_n \quad \mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3 \quad \mathbf{x}_4)$ contains four independently sampled regressor variables plus one intercept column:
 - $\mathbf{x}_1 \stackrel{iid}{\sim} \mathcal{N}(5, 16)$,
 - $\mathbf{x}_2 \stackrel{iid}{\sim} \text{Exp}(5)$,
 - $\mathbf{x}_3 \stackrel{iid}{\sim} \mathcal{U}([-2, 12])$,
 - $\mathbf{x}_4 \stackrel{iid}{\sim} \text{Ber}(0.3)$.
- The design matrix $\mathbf{Z} = (\mathbf{1}_n \quad \mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{z}_3)$ contains the additional regressor variable $\mathbf{z}_3 \stackrel{iid}{\sim} t_{10}$, which is independently sampled from all other columns.
- The true coefficient vectors are given by $\beta = (\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4)^T = (0 \quad -10 \quad -5 \quad -3 \quad -1)^T$ and $\gamma = (\gamma_0 \quad \gamma_1 \quad \gamma_2 \quad \gamma_3)^T = (0 \quad 3 \quad 5 \quad 10)^T$.
- Three different specifications for the outcome distribution were chosen:
 - $y_i \sim \mathcal{N}(\mu, \sigma^2)$,
 - $y_i \sim \mu + \left(\sigma \cdot \sqrt{\frac{3}{5}}\right) T$, where $T \sim t_5$,
 - $y_i \sim \mu + \sigma \cdot U$, where $U \sim \mathcal{U}([0, 1])$. In all cases, the outcome vectors are generated with the covariates on their original (unstandardized) scale.
- In order to isolate the impact of the different shapes of the three probability distributions from the effect of varying moment structures, the mean $\mu = \mathbf{x}_i^T \beta$ and the variance $\sigma^2 = \exp(\mathbf{z}_i^T \gamma)^2$ are held constant across the models.
- All three models `mcmc_ridge()`, `mcmc()` and `lmls()` are fitted with standardized covariates. The sample size is set to $n = 50$ and the result of both Bayesian samplers are based on 10.000 simulations.

Section ??: Hyperparameters - Impact on Estimation Accuracy

Section ??: Hyperparameters - Impact on Ridge Penalty

This simulation study is conducted in the following way:

- The column vectors of the design matrices $\mathbf{X} = (\mathbf{x}_1 \quad \mathbf{x}_2)$ and $\mathbf{Z} = (\mathbf{z}_1 \quad \mathbf{z}_2)$ are independently drawn from normal distributions with variance $\sigma^2 = 1$ and varying means $\mu \in (1, 2, 5, 3)$. Then both design matrices are standardized and intercept columns are added.
- The true coefficient vectors are given by $\beta = (\beta_0 \quad \beta_1 \quad \beta_2)^T = (0 \quad 8 \quad 2)^T$ and $\gamma = (\gamma_0 \quad \gamma_1 \quad \gamma_2)^T = (0 \quad 3 \quad 3)^T$.
- The outcome variable \mathbf{y} is generated based on the standardized covariates according to the correctly specified location-scale model $y_i \stackrel{iid}{\sim} \mathcal{N}(\mathbf{x}_i^T \beta, \exp(\mathbf{z}_i^T \gamma)^2)$ for $i = 1, \dots, n$ with sample size $n = 50$.

- The pairs of hyperparameters (a_τ, b_τ) and (a_ξ, b_ξ) take values on a grid, which is constructed by all combinations of the sequence $(\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16, 32, 64, 128, 256)$. While one pair is varied, the other pair is fixed on the values $(1, 1)$.
- The `mcmc_ridge()` function does not use the `lmls()` function as basis in this case, but rather use the standardized data directly as input. The number of simulations `num_sim` is chosen as 1000 and the starting values `beta_start` and `gamma_start` are set to $(1 \ 1)$, respectively.