Mathematical Model

1 Mathematical Model

This chapter is dedicated to the underlying theoretical model of Bayesian Ridge Regression in the context of the Gaussian Location-Scale Regression model that the `lmls` package is based on. First, the distributional assumptions for the *location* parameter $\boldsymbol{\beta}$, the *scale* parameter $\boldsymbol{\gamma}$ and the prior variances τ^2 and ξ^2, which are specific to Bayesian models, are clearly stated. Based on these prior distributions, the full conditional distributions of each parameter given all of the remaining model components are derived. Throughout this rather theoretical exposition we will build connections from the derived equations to practical consequences that have to be kept in mind for the code implementation discussed in sections **??** and **??**.

1.1 Prior Distributions

Assuming conditional independence among the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as well as flat priors for the intercept parameters

$$f(\beta_0) \propto const \qquad \text{and} \qquad f(\gamma_0) \propto const,$$

first note that

$$\begin{aligned} f(\boldsymbol{\beta} \mid \tau^2) &= f(\beta_0) f(\tilde{\boldsymbol{\beta}} \mid \tau^2) \propto f(\tilde{\boldsymbol{\beta}} \mid \tau^2) \quad \text{and} \\ f(\boldsymbol{\gamma} \mid \xi^2) &= f(\gamma_0) f(\tilde{\boldsymbol{\gamma}} \mid \xi^2) \propto f(\tilde{\boldsymbol{\gamma}} \mid \xi^2), \end{aligned}$$

where the notation $\boldsymbol{\beta} = (\beta_0, ..., \beta_K) \in \mathbb{R}^{K+1}$, $\tilde{\boldsymbol{\beta}} = (\beta_1, ..., \beta_K) \in \mathbb{R}^K$ and, analogously, $\boldsymbol{\gamma} = (\gamma_0, ..., \gamma_J) \in \mathbb{R}^{J+1}$, $\tilde{\boldsymbol{\gamma}} = (\gamma_1, ..., \gamma_J) \in \mathbb{R}^J$ is used.

Thus, the Prior distributions of the parameters in the Bayesian Ridge Regression model, which are responsible for the regularizing effect compared to models without penalty, are given by

- $\tilde{\boldsymbol{\beta}} \mid \tau^2 \sim \mathcal{N}\left(\mathbf{0}, \, \tau^2 \cdot \mathbf{I}_K\right)$,
- $\tilde{\boldsymbol{\gamma}} \mid \xi^2 \sim \mathcal{N}\left(\mathbf{0}, \, \xi^2 \cdot \mathbf{I}_J\right)$,
- $\tau^2 \sim IG(a_\tau, \, b_\tau)$, with fixed hyperparameters a_τ and b_τ,
- $\xi^2 \sim IG(a_\xi, \, b_\xi)$, with fixed hyperparameters a_ξ and b_ξ.

1.2 Full Posterior Distribution

The starting point for deriving the Full Conditional distributions, which will majorly impact the implementation of the Metropolis-Hastings sampling process, is always the Full Posterior distribution. As usual in the Bayesian literature, we write the Full Posterior $f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2 \mid \mathbf{y})$ in terms of the Likelihood function / observation model $f(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2)$ and the *joint* Prior distribution $f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2)$, i.e.

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2 \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2) \cdot f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2).$$

The specific form of the Likelihood function is given by the Location-Scale Regression model that the `lmls` package is built upon:

$$y_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2 = y_i \mid \boldsymbol{\beta}, \boldsymbol{\gamma} \sim \mathcal{N}\left(\mathbf{x}_i^T \boldsymbol{\beta}, \, \exp\left(\mathbf{z}_i^T \boldsymbol{\gamma}\right)^2\right).$$

Taking the independence structure into account, the joint Prior distribution can be written as

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2) &= f(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \tau^2, \xi^2) \cdot f(\boldsymbol{\gamma}, \tau^2, \xi^2) \\ &= f(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \tau^2, \xi^2) \cdot f(\boldsymbol{\gamma} \mid \tau^2, \xi^2) \cdot f(\tau^2, \xi^2) \\ &= f(\boldsymbol{\beta} \mid \tau^2) \cdot f(\boldsymbol{\gamma} \mid \xi^2) \cdot f(\tau^2) \cdot f(\xi^2). \end{aligned}$$

Combining these results yields for the Full Posterior distribution the general form

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2 \mid \mathbf{y}) &\propto f(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2) \cdot f(\boldsymbol{\beta} \mid \tau^2) \cdot f(\boldsymbol{\gamma} \mid \xi^2) \cdot f(\tau^2) \cdot f(\xi^2) \\ &\propto f(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}) \cdot f(\tilde{\boldsymbol{\beta}} \mid \tau^2) \cdot f(\tilde{\boldsymbol{\gamma}} \mid \xi^2) \cdot f(\tau^2) \cdot f(\xi^2), \end{aligned}$$

in which the corresponding densities for the observation model and the individual prior distributions can be inserted. These are given by

- $f(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \exp\left(\mathbf{z}_i^T \boldsymbol{\gamma}\right)^2}} \cdot \exp\left(-\frac{1}{2\exp\left(\mathbf{z}_i^T \boldsymbol{\gamma}\right)^2} \cdot \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2\right)$,
- $f(\boldsymbol{\beta} \mid \tau^2) \propto \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\tau^2}} \cdot \exp\left(-\frac{1}{2\tau^2} \cdot \beta_k^2\right) = (2\pi)^{-\frac{K}{2}}\, \tau^{-K} \exp\left(-\frac{1}{2\tau^2} \cdot \tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{\beta}}\right)$,
- $f(\boldsymbol{\gamma} \mid \xi^2) \propto \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\xi^2}} \cdot \exp\left(-\frac{1}{2\xi^2} \cdot \gamma_j^2\right) = (2\pi)^{-\frac{J}{2}}\, \xi^{-J} \exp\left(-\frac{1}{2\xi^2} \cdot \tilde{\boldsymbol{\gamma}}^T \tilde{\boldsymbol{\gamma}}\right)$,
- $f(\tau^2) = \frac{b_\tau}{\Gamma(a_\tau)} \left(\frac{1}{\tau^2}\right)^{a_\tau + 1} \exp\left(-\frac{b_\tau}{\tau^2}\right)$,
- $f(\xi^2) = \frac{b_\xi}{\Gamma(a_\xi)} \left(\frac{1}{\xi^2}\right)^{a_\xi + 1} \exp\left(-\frac{b_\xi}{\xi^2}\right)$,

leading to the Full Posterior

$$\begin{aligned} f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau^2, \xi^2 \mid \mathbf{y}) \propto & \prod_{i=1}^{n} \frac{1}{\exp\left(\mathbf{z}_i^T \boldsymbol{\gamma}\right)} \cdot \exp\left(-\frac{1}{2\exp\left(\mathbf{z}_i^T \boldsymbol{\gamma}\right)^2} \cdot \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2\right) \\ & \cdot \tau^{-K} \exp\left(-\frac{1}{2\tau^2} \tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{\beta}}\right) \cdot \xi^{-J} \exp\left(-\frac{1}{2\xi^2} \tilde{\boldsymbol{\gamma}}^T \tilde{\boldsymbol{\gamma}}\right) \\ & \cdot \left(\frac{1}{\tau^2}\right)^{a_\tau + 1} \exp\left(-\frac{b_\tau}{\tau^2}\right) \cdot \left(\frac{1}{\xi^2}\right)^{a_\xi + 1} \exp\left(-\frac{b_\xi}{\xi^2}\right). \end{aligned}$$

1.3 Full Conditional Distributions

The Full Posterior distribution contains complete information about the statistical model. For our purposes, we are mostly interested in the Full Conditional Distribution of each model parameter. These can be obtained by simply neglecting all factors of the Full Posterior that do not depend on the parameter in consideration.

The Full Conditional distribution can then be recovered by the resulting density kernel, either by recognizing a known distribution or by adding a normalization constant (which, however, is not needed for Markov Chain Monte Carlo sampling).

1.3.1 Full Conditional of τ^2:

$$\begin{aligned} f(\tau^2 \mid \cdot) &\propto \tau^{-K} \cdot \exp\left(-\frac{1}{2\tau^2} \cdot \tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{\beta}}\right) \cdot \left(\frac{1}{\tau^2}\right)^{a_\tau + 1} \exp\left(-\frac{b_\tau}{\tau^2}\right) \\ &\propto \left(\frac{1}{\tau^2}\right)^{a_\tau + \frac{K}{2} + 1} \cdot \exp\left(-\frac{1}{\tau^2}\left(b_\tau + \frac{1}{2}\tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{\beta}}\right)\right). \end{aligned}$$

This is the kernel of an Inverse-Gamma distribution parameterized by

$$\tau^2 \mid \cdot \sim IG(a_\tau + \frac{K}{2},\, b_\tau + \frac{1}{2}\tilde{\boldsymbol{\beta}}^T \tilde{\boldsymbol{\beta}}).$$

1.3.2 Full Conditional of ξ^2:

$$\begin{aligned} f(\xi^2 \mid \cdot) &\propto \xi^{-J} \cdot \exp\left(-\frac{1}{2\xi^2} \cdot \tilde{\boldsymbol{\gamma}}^T \tilde{\boldsymbol{\gamma}}\right) \cdot \left(\frac{1}{\xi^2}\right)^{a_\xi + 1} \exp\left(-\frac{b_\xi}{\xi^2}\right) \\ &\propto \left(\frac{1}{\xi^2}\right)^{a_\xi + \frac{J}{2} + 1} \cdot \exp\left(-\frac{1}{\xi^2}\left(b_\xi + \frac{1}{2}\tilde{\boldsymbol{\gamma}}^T \tilde{\boldsymbol{\gamma}}\right)\right). \end{aligned}$$

Thus, the Full Conditional of ξ^2 follows an Inverse-Gamma distribution as well:

$$\xi^2 \mid \cdot \sim IG(a_\xi + \frac{J}{2},\, b_\xi + \frac{1}{2}\tilde{\boldsymbol{\gamma}}^T\tilde{\boldsymbol{\gamma}}).$$

1.3.3 Full Conditional of $\boldsymbol{\beta}$:

Here, the derivation is more involved. In order to keep the calculations structured, we introduce the following notation:

$$\mathbf{w}_i := \frac{\mathbf{x}_i}{\exp\left(\mathbf{z}_i^T\boldsymbol{\gamma}\right)} \in \mathbb{R}^{K+1}, \qquad \mathbf{W} := \begin{pmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_n^T \end{pmatrix} \in \mathbb{R}^{n \times (K+1)}$$

and

$$u_i := \frac{y_i}{\exp\left(\mathbf{z}_i^T\boldsymbol{\gamma}\right)} \in \mathbb{R}, \qquad \mathbf{u} := \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \in \mathbb{R}^n,$$

yielding $\sum_{i=1}^n u_i\mathbf{w}_i = \mathbf{W}^T\mathbf{u}$ and $\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i^T = \mathbf{W}^T\mathbf{W}$.

Therefore the Full Conditional distribution of $\boldsymbol{\beta}$ can be written as

$$\begin{aligned}
f(\boldsymbol{\beta} \mid \cdot) &\propto \exp\left(-\frac{1}{2}\cdot\left[\frac{1}{\tau^2}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}} + \sum_{i=1}^n \frac{1}{\exp\left(\mathbf{z}_i^T\boldsymbol{\gamma}\right)^2}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2\right]\right) \\
&= \exp\left(-\frac{1}{2}\cdot\left[\frac{1}{\tau^2}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}} + \sum_{i=1}^n\left(\frac{y_i^2}{\exp\left(\mathbf{z}_i^T\boldsymbol{\gamma}\right)^2} - \frac{2y_i\mathbf{x}_i^T}{\exp\left(\mathbf{z}_i^T\boldsymbol{\gamma}\right)^2}\boldsymbol{\beta} + \boldsymbol{\beta}^T\frac{\mathbf{x}_i}{\exp\left(\mathbf{z}_i^T\boldsymbol{\gamma}\right)}\frac{\mathbf{x}_i^T}{\exp\left(\mathbf{z}_i^T\boldsymbol{\gamma}\right)}\boldsymbol{\beta}\right)\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\cdot\left[\frac{1}{\tau^2}\tilde{\boldsymbol{\beta}}^T\tilde{\boldsymbol{\beta}} - 2\cdot\sum_{i=1}^n u_i\mathbf{w}_i^T\boldsymbol{\beta} + \sum_{i=1}^n \boldsymbol{\beta}^T\mathbf{w}_i\mathbf{w}_i^T\boldsymbol{\beta}\right]\right) \\
&= \exp\left(-\frac{1}{2}\cdot\left[\boldsymbol{\beta}^T\left(\sum_{i=1}^n \mathbf{w}_i\mathbf{w}_i^T + \frac{1}{\tau^2}\begin{pmatrix} 0 & 0 \\ 0 & \mathbf{I}_K \end{pmatrix}\right)\boldsymbol{\beta} - 2\cdot\sum_{i=1}^n \boldsymbol{\beta}^T u_i\mathbf{w}_i\right]\right) \\
&= \exp\left(-\frac{1}{2}\cdot\left[\boldsymbol{\beta}^T\left(\mathbf{W}^T\mathbf{W} + \frac{1}{\tau^2}\begin{pmatrix} 0 & 0 \\ 0 & \mathbf{I}_K \end{pmatrix}\right)\boldsymbol{\beta} - 2\cdot\boldsymbol{\beta}^T\mathbf{W}^T\mathbf{u}\right]\right).
\end{aligned}$$

Comparing this representation with the kernel of a multivariate normal distribution leads to the conclusion

$$\boldsymbol{\beta} \mid \cdot \sim \mathcal{N}_{K+1}\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right)$$

with the parameters

$$\boldsymbol{\Sigma}_\beta = \left(\mathbf{W}^T\mathbf{W} + \frac{1}{\tau^2}\begin{pmatrix} 0 & 0 \\ 0 & \mathbf{I}_K \end{pmatrix}\right)^{-1} \qquad \text{and} \qquad \boldsymbol{\mu}_\beta = \boldsymbol{\Sigma}_\beta\mathbf{W}^T\mathbf{u}.$$

1.3.4 Full Conditional of $\boldsymbol{\gamma}$:

Using the notation

$$\mathbf{z}_i \in \mathbb{R}^{J+1}, \qquad \mathbf{Z} = \begin{pmatrix} \mathbf{z}_1^T \\ \vdots \\ \mathbf{z}_n^T \end{pmatrix} \in \mathbb{R}^{n \times (J+1)} \qquad \text{and} \qquad \boldsymbol{\gamma} \in \mathbb{R}^{J+1},$$

the Full Conditional distribution of $\boldsymbol{\gamma}$ is given by

$$\begin{aligned}
f(\boldsymbol{\gamma} \mid \cdot) &\propto \exp\left(-\frac{1}{2}\cdot\left[\frac{1}{\xi^2}\tilde{\boldsymbol{\gamma}}^T\tilde{\boldsymbol{\gamma}} + \sum_{i=1}^n\left(\frac{1}{\exp\left(\mathbf{z}_i^T\boldsymbol{\gamma}\right)^2}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2 + 2\cdot\mathbf{z}_i^T\boldsymbol{\gamma}\right)\right]\right) \\
&= \exp\left(-\frac{1}{2}\cdot\left[\frac{1}{\xi^2}\tilde{\boldsymbol{\gamma}}^T\tilde{\boldsymbol{\gamma}} + 2\cdot\mathbf{1}_n^T\mathbf{Z}\boldsymbol{\gamma} + \sum_{i=1}^n\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}}{\exp\left(\mathbf{z}_i^T\boldsymbol{\gamma}\right)}\right)^2\right]\right).
\end{aligned}$$

In contrast to the Full Conditionals for $\boldsymbol{\beta}$, τ^2 and ξ^2, this kernel cannot be assigned to a known distribution. Thus, for sampling from the Full Posterior distribution, it is not feasible to use a Gibbs Sampler in its purest form. More specifically, we will include a Metropolis Hastings step for sampling the $\boldsymbol{\gamma}$ parameter vector.

Although this 'inconvenience' is not required for $\boldsymbol{\beta}$ (since we can use independent samples from a multivariate normal distribution), we have explored and analyzed the statistical properties of sampling both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ via the Metropolis-Hastings procedure, which will be briefly discussed in the following chapters.