Probabilistic Machine Learning

Bayesian (Generalized) Linear Models

Department of Statistics
Ludwig-Maximilians-Universität München

Lona Koers

Munich, 04. July 2025

Submitted as a seminar paper for the seminar on Probabilistic Machine Learning.
Supervised by Dr. Ludwig Bothmann

Abstract

Bayesian generalized linear models (GLMs) offer a framework for incorporating uncertainty and prior knowledge into regression models. By placing prior distributions over parameters, they enable posterior-based uncertainty quantification and regularization.

Especially in high-dimensional or low-information settings, regularization priors stabilize inference and improve generalization. However, the posterior distribution in Bayesian GLMs is often analytically intractable, which makes approximate inference methods necessary.

This paper introduces the Bayesian view on linear and logistic regression while highlighting the role of regularization priors and comparing Laplace approximation and Markov chain Monte Carlo for posterior inference. Using synthetic data, we evaluated predictive performance and variable selection accuracy in low-information scenarios under different priors and compared different methods for posterior inference.

Contents

1 Introduction **1**

2 Bayesian Linear Model **2**
2.1 Model Definition 2
2.2 Prior Choice 2
2.3 Bayesian Inference with Closed Form Priors 4

3 Bayesian Logistic Model **6**
3.1 Bayesian Generalized Linear Model 6
3.2 Bayesian Logistic Model 6
3.3 Approximate Bayesian Inference 7

4 Illustrative Examples **9**
4.1 Regularization and Variable Selection 9
4.2 Performance of Approximate Inference Algorithms 11

5 Conclusion and Outlook **12**

A Appendix **V**

B Electronic Appendix **VII**

1 Introduction

Generalized linear models (GLMs) are a fundamental tool in statistics and machine learning and are widely applied across various domains. Their appeal lies in their simplicity, interpretability, and extensibility (Nelder and Wedderburn, 1972). However, GLMs also come with limitations: They rely on maximum likelihood estimation (MLE), offer no mechanism for incorporating prior information or stabilizing inference (Gelman et al., 2008), and most importantly, only produce point estimates and thus cannot capture the full range of uncertainty in predictions and parameter estimates (Tyralis and Papacharalampous, 2024).

Bayesian GLMs take on a different viewpoint to addresses these shortcomings. They offer a natural way to quantify uncertainty with posterior distributions, which is especially useful in data-scarce scenarios. For instance, Sondhi et al. (2021) demonstrate this in precision oncology, where Bayesian inference compensates for small sample sizes and stabilizes confidence estimation in effect sizes. Recent work has also shown that Bayesian regularization techniques can perform on par with or even outperform classic regularization, while also offering greater flexibility and interpretability (van Erp et al., 2019, Celeux et al., 2012). Additionally, a Bayesian framework allows for the incorporation of domain knowledge through informative priors. For example, Chien et al. (2023) outline a framework for constructing priors directly from expert knowledge or prior experiments.

This paper explores Bayesian GLMs as an alternative to classical frequentist approaches. In Section 2, we introduce Bayesian linear regression as a familiar starting point within the Bayesian framework. Section 3 extends this foundation to generalized models, focusing on logistic regression as the most-used GLM. Section 4 illustrates the application of regularization and approximate inference methods in Bayesian GLMs using synthetic data experiments.

2 Bayesian Linear Model

The (frequentist) linear model is probably the most widely used model in statistics and machine learning. The frequentist and the Bayesian linear models are described in many introductory texts on statistical modeling, such as Fahrmeir et al. (2021) or Gelman, Carlin, Stern, Dunson, Vehtari and Rubin (2013).

2.1 Model Definition

We observe an independent and identically distributed (i.i.d.) sample $\boldsymbol{D} = (\boldsymbol{y}, \boldsymbol{X})$ with n observations and p covariates and assume a linear relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$.[1] A condition on $\boldsymbol{X}$ is always implicit. The frequentist linear regression model then assumes

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}), \tag{1}$$

where the weight parameter $\boldsymbol{\theta}$ and the variance σ^2 are estimated to obtain the fitted model.

To view linear regression from a Bayesian perspective, we simply reinterpret the parameters as random variables. Conditioning on $\boldsymbol{\theta}$ and σ^2, the likelihood takes the same form as in Equation 1:

$$\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}), \tag{2}$$

Note that to predict multiple outputs, an extension to multivariate linear regression is possible.

2.2 Prior Choice

Normal (Inverse Gamma) Prior

To complete the Bayesian linear model specification, we place conjugate priors on both $\boldsymbol{\theta}$ and σ^2.

$$\begin{aligned} \boldsymbol{\theta} \mid \sigma^2 &\sim \mathcal{N}(\breve{\boldsymbol{\mu}}, \sigma^2 \breve{\Sigma}) \\ \sigma^2 &\sim \text{IG}(\breve{a}, \breve{b}), \end{aligned} \tag{3}$$

where $\breve{\boldsymbol{\mu}}, \breve{\Sigma}, \breve{a}$ and $\breve{b}$ are the prior parameters. We choose a Gaussian prior on $\boldsymbol{\theta}$ because it is conjugate to the Gaussian likelihood of $\boldsymbol{y}$. Since the inverse gamma (IG) distribution of σ^2 is conjugate to the Gaussian conditional distribution of $\boldsymbol{\theta}$, the joint prior of $\boldsymbol{\theta}$ and σ^2

$$p(\boldsymbol{\theta}, \sigma^2) \overset{\text{Bayes' rule}}{=} p(\boldsymbol{\theta} \mid \sigma^2) p(\sigma^2)$$

follows a normal-inverse-gamma (NIG) distribution. We can then use Bayes' rule once again to derive the unconditional prior distribution of $\boldsymbol{\theta}$ as a multivariate Student t-distribution.

$$\boldsymbol{\theta} \sim \mathcal{T}(2\breve{a}, \breve{\boldsymbol{\mu}}, \frac{\breve{a}}{\breve{b}} \breve{\Sigma})$$

[1] For a detailed explanation of the notation see Appendix A.

Uninformative Prior

The idea of an uninformative prior is to maximize the influence of the data on the posterior in the absence of prior knowledge. Especially when little to no prior information is available, we can reflect this in the model by flattening the NIG prior. We set

$$\breve{\boldsymbol{\mu}} = \mathbf{0}, \quad \breve{\Sigma}^{-1} = \mathbf{0} \text{ i.e., } \breve{\Sigma} \to \infty$$

and choose $\breve{a} = -\frac{p}{2}$ and $\breve{b} = 0$, where p is the number of features in the model.
We can easily see that with this assumption, the prior for $\boldsymbol{\theta}$ becomes very flat while still retaining the useful qualities from the setup described in Equation 3.
The prior distributional assumptions would then be:

$$\begin{aligned} \boldsymbol{\theta} \mid \sigma^2 &\overset{a}{\sim} \mathcal{N}(\breve{\boldsymbol{\mu}}, \sigma^2 \infty)^2, \qquad & p(\boldsymbol{\theta} \mid \sigma^2) &\propto 1 \\ \sigma^2 &\sim \text{IG}(-\frac{p}{2}, 0), & p(\sigma^2) &\propto \frac{1}{\sigma^2} \end{aligned} \tag{4}$$

Note that we generally have to be careful with completely flat priors because they can result in improper posteriors. It is generally necessary to check if the resulting posterior is proper, which is the case here (Fahrmeir et al., 2021). Another good solution for use cases with little prior knowledge that still require a proper posterior is Zellner's g-prior (Zellner, 1986).

Regularization Priors

Regularization (or penalization) regulates the trade-off between model complexity and out-of-sample performance, or equivalently bias and variance. In frequentist statistics, we minimize the penalized least squares criterion (PLS)

$$\text{PLS}(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) + \lambda \text{ pen}(\boldsymbol{\theta}).$$

where $\lambda > 0$ controls the balance of the tradeoff and therefore the strength of regularization.

In the Bayesian view, we introduce a regularization prior on $\boldsymbol{\theta}$. Concretely:

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}),^3 \\ \boldsymbol{\theta} &\sim \text{regularization prior} \\ \sigma^2 &\sim \text{IG}(\breve{a}, \breve{b}). \end{aligned}$$

Although there are many options for regularization priors, we are going to focus on regularization priors that align directly with familiar frequentist penalties.

[2] Informally stated for demonstration purposes.

[3] Usually, it does not make sense to regularize the intercept. To be completely accurate, we would need to separate the intercept from $\boldsymbol{\theta}$, i.e., split $\boldsymbol{\theta}$ into $(\theta_0, \boldsymbol{\theta'}^\top)$ and consequently set $\boldsymbol{X'}$ as the design matrix without a column for the intercept. We would then specify the model as $\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}(\theta_0 \boldsymbol{I} + \boldsymbol{X'}\boldsymbol{\theta'}, \sigma^2 \boldsymbol{I})$. We chose to simplify this and stick to the previously established definitions because we aim for an understandable explanation of the basic concept of Bayesian regularization.

Ridge regularization (Hoerl and Kennard, 1970a,b) uses $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$. The Bayesian analogue (e.g. Hsiang, 1975, MacKay, 1992) specifies

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}), \tag{5}$$

with τ^2 controlling the degree of regularization akin to the role of λ. In contrast to λ, τ^2 does not need to be set in advance or be optimized as a hyperparameter. We can simply embed it in a hierarchical model by specifying a prior for τ^2, e.g. $\tau^2 \sim \text{IG}(\breve{a}_\tau, \breve{b}_\tau)$, and estimate it alongside $\boldsymbol{\theta}$ and σ^2.

LASSO (least absolute shrinkage and selection operator) **regularization** (Tibshirani, 1996) uses $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ to perform variable selection by setting elements θ_j of $\boldsymbol{\theta}$ to 0 during estimation. This means that LASSO regularization promotes a *sparse* solution. The Bayesian LASSO specifies a Laplace prior on $\boldsymbol{\theta}$ via the scale mixture representation (Park and Casella, 2008)

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\tau}^2 &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\tau}^2 \boldsymbol{I}) \\ \tau_j^2 &\overset{\text{i.i.d.}}{\sim} \text{Exp}(0.5\lambda^2), \quad j = 1, \ldots, p, \end{aligned} \tag{6}$$

where the regularization parameter λ^2 is often given a (hyper-) prior, e.g. $\lambda^2 \sim \text{G}(\breve{a}_\lambda, \breve{b}_\lambda)$.

Because the Bayesian LASSO does not promote a sparse solution, discrete-mixture Spike-and-Slab priors (Mitchell and Beauchamp, 1988) (which are necessary for categorical covariates) or the heavy-tailed horseshoe prior (Carvalho et al., 2010) are preferred for variable selection.

2.3 Bayesian Inference with Closed Form Priors

Parameter Posterior Distribution

In a frequentist linear model, we use least-squares (LS) estimation to obtain the estimate

$$\hat{\boldsymbol{\theta}}_{LS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{7}$$

for $\boldsymbol{\theta}$. Under Gaussian errors, this satisfies

$$\hat{\boldsymbol{\theta}}_{LS} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}).$$

To quantify the uncertainty in the estimation, we can compute confidence intervals for $\boldsymbol{\theta}$, but these reflect only the variability in the estimator, not the uncertainty about the true parameter itself.

In contrast, the Bayesian approach yields a full posterior distribution on $\boldsymbol{\theta}$ by updating the prior distribution with observed data using Bayes' rule. With the NIG prior introduced in Equation 3, conjugacy implies for the joint posterior $p(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y})$ that

$$\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y} \sim \text{NIG}(\hat{\boldsymbol{\mu}}, \hat{\Sigma}, \hat{a}, \hat{b})$$

with posterior mean and variance [4]

$$\hat{\boldsymbol{\mu}} = \hat{\Sigma}(\breve{\Sigma}^{-1}\breve{\boldsymbol{\mu}} + \boldsymbol{X}^\top \boldsymbol{y}), \quad \hat{\Sigma} = (\boldsymbol{X}^\top \boldsymbol{X} + \breve{\Sigma}^{-1})^{-1}. \tag{8}$$

Integrating out σ^2 yields $\boldsymbol{\theta} \mid \boldsymbol{y} \sim \mathcal{T}(2\hat{a}, \hat{\boldsymbol{\mu}}, \hat{b}/\hat{a}\hat{\Sigma})$ and Bayesian credibility intervals can be derived directly from this distribution (see e.g. Held and Sabanés Bové, 2020).

Since we defined the non-information prior (Equation 4) as a special case of the NIG-distributed prior, we can use Equation 8 to directly calculate the posterior mean and variance for a non-informative setting as

$$\hat{\boldsymbol{\mu}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}, \quad \hat{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}.$$

The posterior mean $\hat{\boldsymbol{\mu}}$ coincides with $\hat{\boldsymbol{\theta}}_{LS}$ (Equation 7), which means that a Bayesian linear model with a non-informative prior converges to the frequentist solution. More generally, as the prior variance $\breve{\Sigma}$ grows, $\hat{\boldsymbol{\mu}}$ approaches $\hat{\boldsymbol{\theta}}_{LS}$, since the likelihood (and thus the data) dominates the posterior.

Because Bayesian ridge regression is simply the NIG case of Equation 3 with finite $\breve{\Sigma}$, inference results in the same posterior update as in Equation 8. By contrast, the Bayesian LASSO's Laplace prior has no closed form posterior, but we can easily sample from it using Gibbs sampling (Park and Casella, 2008). We will go more into approximate inference for Bayesian regression models in more depth in Section 3.3.

Posterior Predictive Distribution

In many applications, we do not care so much about $\boldsymbol{\theta}$ itself. Instead, we are interested in predictions $\tilde{\boldsymbol{y}}$ for new, unseen inputs $\tilde{\boldsymbol{X}}$ (or test data $(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}})$), independent of the training data $\boldsymbol{D}$. The Bayesian answer to this is the *posterior predictive distribution* (PPD) (see e.g Box, 1980, Barbieri, 2015)

$$p(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) = \int p(\tilde{\boldsymbol{y}}, \boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta} = \int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \overset{\tilde{\boldsymbol{y}} \perp \boldsymbol{y} \mid \boldsymbol{\theta}}{=} \int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

which is an average of conditional probabilities over the posterior distribution of $\boldsymbol{\theta}$.[5] For the NIG prior in Equation 3, one can show that

$$\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \sigma^2, \boldsymbol{y} \sim \mathcal{T}(2\hat{a}, \tilde{\boldsymbol{X}}\boldsymbol{\theta}, \frac{\hat{b}}{\hat{a}}(\boldsymbol{I} + \tilde{\boldsymbol{X}}\hat{\Sigma}\tilde{\boldsymbol{X}}^\top)).^6$$

Interestingly, the posterior predictive mean $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{X}}\boldsymbol{\theta}$ of the t-distribution coincides with the least squares prediction and its scale matrix reflects both observational noise and posterior uncertainty. Bayesian inference with the Gaussian conjugate is more thoroughly described by Murphy (2007).

If no closed form exists, the PPD can also be simulated (see Section 3.3).

[4] For the full calculation see Appendix A

[5] A note on intuition: In essence, the PPD is the marginal distribution of $\tilde{\boldsymbol{y}}$, conditioned on the data $\boldsymbol{y}$. We recognize the marginal distribution of $\boldsymbol{y}$ from Bayes' rule as the normalization constant, i.e., $p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$.

[6] see Appendix A for more detail.

3 Bayesian Logistic Model

3.1 Bayesian Generalized Linear Model

Bayesian generalized linear models extend the familiar Bayesian linear regression framework by replacing the Gaussian distributional assumption on $\boldsymbol{y}$ with an arbitrary exponential-family distribution (Nelder and Wedderburn, 1972, West et al., 1985). In their most general form, we assume

$$\boldsymbol{y} \mid \boldsymbol{\theta} \sim F(g^{-1}(\boldsymbol{X}\boldsymbol{\theta})),$$

where F is any exponential-family distribution (e.g. Binomial, Poisson, Gamma) and g^{-1} is the inverse link function. Priors for the parameter $\boldsymbol{\theta}$ can be set in the same way as for the Bayesian linear model. However, in practice, the prior choice also depends on the link function, since the link transforms the linear predictor and thereby influences the prior's effect on the response scale (West et al., 1985, Hosack et al., 2017).

3.2 Bayesian Logistic Model

We are going to illustrate Bayesian GLMs with the example of logistic regression models, which have a wide variety of applications in statistics, from text classification to medicine and genetic modeling (see e.g. Dayanik et al., 2006, Sondhi et al., 2021, for interesting applications).

Model Definition

The Bayesian logistic regression model is defined as

$$\begin{aligned} \boldsymbol{y}_i \mid \boldsymbol{\theta} &\sim \text{Bin}(1, g^{-1}(\boldsymbol{x}_i\boldsymbol{\theta})), \quad i = 1, \dots, n \\ g^{-1}(\boldsymbol{x}_i\boldsymbol{\theta}) &= \sigma(\boldsymbol{x}_i\boldsymbol{\theta}). \end{aligned} \tag{9}$$

for observations $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{ip})^\top$ and where $\sigma(y) = \frac{\exp(y)}{1+\exp(y)}$ is the logistic (sigmoid) function. Other choices like the probit link can also be used (see e.g. Albert and Chib, 1993).

Prior Choice

Unlike the Gaussian linear model, the logistic likelihood breaks conjugacy. Nevertheless, we can use a Gaussian prior (Equation 3) or an (improper) flat prior (Equation 4) for $\boldsymbol{\theta}$, but both require approximate inference (see Section 3.3).

To address separation (i.e. perfect prediction) and to induce shrinkage, heavier-tailed priors are commonly employed. Gelman et al. (2008) introduced the t-distribution as a prior for low-information settings and mentions the Cauchy distribution as another possibility, which is elaborated on by Ghosh et al. (2017).

Regularization can be achieved using the same prior distributions as introduced for Bayesian linear regression in Section 2.2 (see e.g. van Erp et al., 2019, Fahrmeir et al., 2010, O'Hara and Sillanpää, 2009).

3.3 Approximate Bayesian Inference

Unlike for the linear model, Bayesian inference with a closed form posterior is not possible in most cases (see e.g. Polson et al., 2013). To sample from the posterior and PPD, we need to use approximate Bayesian inference methods.

Sampling from the Posterior with MCMC Methods

Markov chain Monte Carlo (MCMC) generates samples from the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ without making any (explicit) assumptions about the form of the posterior, although MCMC performs best if the parameter posterior is known up to a constant. The Metropolis–Hastings algorithm (Hastings, 1970) for K samples[7] proceeds as follows:

1. Initialize $\boldsymbol{\theta}^{(1)}$
2. For $k = 1, \dots, K$
   (a) Draw $\boldsymbol{\theta}^{(*)}$ from the *proposal distribution* $q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(k)})$
   (b) calculate the *acceptance probably*

$$\alpha = \min\Big(1, \frac{p(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{y})\; p(\boldsymbol{\theta}^{(*)})\; q(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{\theta}^{(*)})}{p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})\; p(\boldsymbol{\theta}^{(k)})\; q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(k)})}\Big)$$

   (c) Accept or discard the proposal $\boldsymbol{\theta}^{(*)}$ (for $u \sim \text{Uni}[0, 1]$)

$$\begin{cases} u \leq \alpha & \boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(*)} \\ u > \alpha & \boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} \end{cases}$$

The efficiency of Metropolis-Hastings depends critically on the proposal distribution q. A common choice is a Gaussian distribution centered at the current state with the covariance given by the (estimated) negative inverse Hessian of the log–posterior, often obtained via IWLS (Gamerman, 1998, Lenk and DeSarbo, 2000, Scott, 2011):[8]

$$q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(k)}) \sim \mathcal{N}(\boldsymbol{\theta}^{(k)} \mid -H^{-1}(\boldsymbol{\theta}^{(k)})), \quad H(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} \log\Big(p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})\; p(\boldsymbol{\theta}^{(k)})\Big)$$

(Scott, 2011) argues that using heavier-tailed proposals (e.g. Student–t) can improve mixing, which means that the algorithm converges faster and a shorter burn-in period is necessary.

Beyond Metropolis–Hastings, several advanced samplers are popular:

- Gibbs sampling for models with conditional conjugacy (Dellaportas and Smith, 1993).

[7] Note that by construction, the samples are (sometimes heavily) correlated and that the number of repetitions necessary until convergence depends on $\boldsymbol{\theta}^{(1)}$.

[8] The symmetry of the Gaussian distribution simplifies the algorithm to the Metropolis algorithm, where the acceptance probability can be calculated only using $p(\boldsymbol{\theta} \mid \boldsymbol{y})\; p(\boldsymbol{\theta})$.

- Hamiltonian Monte Carlo, which exploits gradient information to explore high-dimensional posteriors efficiently (Neal, 1993).
- Data augmentation (Albert and Chib, 1993), using Gaussian scale mixtures and introducing auxiliary latent variables to restore conjugacy in logistic models (Holmes and Knorr-Held, 2003, Frühwirth-Schnatter and Frühwirth, 2007, Scott, 2011).

Full Bayes with Laplace Approximation

In contrast to MCMC methods, Laplace approximation (LA) approximates the full posterior with a Gaussian distribution (Tierney et al., 1986):

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \approx \mathcal{N}(\hat{\boldsymbol{\theta}}_{MAP}, H^{-1}(\hat{\boldsymbol{\theta}}_{MAP})),$$

where $\hat{\boldsymbol{\theta}}_{MAP}$ is the maximum posterior estimate, obtained by maximizing the (real) posterior with standard optimization methods.

In the case of the Bayesian logistic model with a simple parameter prior $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, this results in

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{MAP} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \overset{\text{Bayes' rule}}{=} \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log\Big(\sigma(y_i \, \boldsymbol{x}_i \boldsymbol{\theta})\Big) - \frac{1}{2\sigma^2} \boldsymbol{\theta}^\top \boldsymbol{\theta} \\
H(\boldsymbol{\theta}) &= -\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{1}{\sigma^2} \boldsymbol{I} + \sum_{i=1}^{n} \sigma(y_i \, \boldsymbol{x}_i \boldsymbol{\theta}) \Big(1 - \sigma(y_i \, \boldsymbol{x}_i \boldsymbol{\theta})\Big) \boldsymbol{x}_i \boldsymbol{x}_i^\top.
\end{aligned}$$

For hierarchical models, Rue et al. (2009) proposed an extended algorithm based on integrated nested Laplace approximations (INLA).

Posterior Predictive Distribution

In logistic regression, which is a binary classification setting, we obtain the PPD by calculating the distribution of the positive class[9] $p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{\theta}, \boldsymbol{y})$ and inferring the negative class.

As MCMC results in samples from the posterior, we can use the samples $\boldsymbol{\theta}^{(k)}$ to approximate the PPD with

$$p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{\theta}, \boldsymbol{y}) \approx \frac{1}{K} \sum_{k=1}^{K} \sigma(\tilde{\boldsymbol{X}} \boldsymbol{\theta}^{(k)}). \tag{10}$$

Under Laplace approximation, we may either

- draw samples $\boldsymbol{\theta}^{(s)} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_{MAP}, H^{-1}(\hat{\boldsymbol{\theta}}_{MAP}))$ with $s = 1, \ldots, S$ and compute Equation 10 or
- use the LA-approximated PPD and compute

$$p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{\theta}, \boldsymbol{y}) = \int \sigma(\tilde{\boldsymbol{X}} \boldsymbol{\theta}) \, \mathcal{N}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_{MAP}, H^{-1}(\hat{\boldsymbol{\beta}}_{MAP})) \, d\boldsymbol{\theta}.$$

[9] Encoded here with $y_i \in \{0 \text{ (negative)}, 1 \text{ (positive)}\}$

4 Illustrative Examples

4.1 Regularization and Variable Selection

We now apply three of the previously discussed priors to linear and logistic models using synthetic data (consisting of training and test sets) under two settings:

- **Scenario A**: A well-behaved setting without collinearity to examine shrinkage:

$$
\begin{aligned}
&n = 150,\ n_{train} = 100,\ n_{test} = 50,\\
&\boldsymbol{\theta} = (2, 1.5, 0, 0, 0),\\
&\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}),\\
&\text{linear: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{I}), \quad \text{logistic: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim \text{Ber}(\sigma(\boldsymbol{X}\boldsymbol{\theta})).
\end{aligned}
$$

- **Scenario B**: A low-information setting where $n \approx p$ with collinearity between informative and non-informative covariates:

$$
\begin{aligned}
&n = 150,\ n_{train} = 30,\ n_{test} = 120,\\
&\boldsymbol{\theta} = (2, 1.5, 0, \overset{26 \text{ times}}{\ldots}, 0),\\
&\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & & \\ & S_3 & 0 \\ & & I_{26} \\ & 0 & \end{pmatrix}, \qquad S_3 = \begin{pmatrix} 1 & 0.8 & 0.8 \\ 0.8 & 1 & 0.8 \\ 0.8 & 0.8 & 1 \end{pmatrix},\\
&\text{linear: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{I}), \quad \text{logistic: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim \text{Ber}(\sigma(\boldsymbol{X}\boldsymbol{\theta})).
\end{aligned}
$$

For each scenario, we fit three linear and three logistic regression models using the following priors: A flat prior as a benchmark, using the conjugate setting described in Equation 4 with the R-package `brms` (Bürkner, 2017).[10] We fit a ridge prior (see Equation 5) and a LASSO prior (see Equation 6) via the `bayesreg` package (Makalic and Schmidt, 2016) with automated optimization of the regularization parameters τ^2 and λ^2. For all models, we ran MCMC with 20,000 iterations, a 1,000 burn-in, and a thinning interval of 10.

Our evaluation focused on two aspects: Firstly, *variable selection accuracy*, i.e. the number of correctly identified influential covariates (Hits) and falsely as influential declared covariates (FP). Although ridge and LASSO shrink coefficients, they do not perform variable selection by themselves. Thus, we used Bayesian credibility intervals as a criterion to decide whether a parameter is credibly nonzero as described in van Erp et al. (2019). Secondly, *predictive accuracy*. This was measured on the test set by the mean log posterior predictive density (MLPPD) proposed by Gelman, Hwang and Vehtari (2013). Since the log-likelihood is a proper scoring rule, it is well-suited to evaluate Bayesian models.

Table 1 shows numerical results, while Figure 1 visualizes parameter estimates and uncertainty. For linear regression, all priors correctly identified the influential variables and

[10] To obtain a virtually flat prior, the Gaussian prior variance was set to 10^6 and the Gamma prior parameters were set to $a = 0.001, b = 0.001$, which results in an uninformative prior.

Model	Prior	Scenario A			Scenario B		
		Hits (of 2)	FP (of 3)	MLPPD	Hits (of 2)	FP (of 28)	MLPPD
Linear	flat	2	0	-1.425	2	1	-1.605
Linear	LASSO	2	0	-1.424	2	0	-1.464
Linear	ridge	2	0	-1.427	2	0	-1.575
Logit	flat	2	1	-0.390	0	21	$-\infty$
Logit	LASSO	1	0	-0.463	1	0	-0.485
Logit	ridge	1	0	-0.455	1	0	-0.493

Table 1: Evaluation metrics of Bayesian linear and logistic regression, each with a flat, ridge, and LASSO prior, under Scenarios A and B.

produced few or no false positives. The effect of regularization is more pronounced in logistic regression: in both scenarios, the regularized models (LASSO and ridge) declared fewer coefficients as influential and reduced false positives. In the low-information Scenario B, regularization priors improved both variable selection and predictive accuracy (MLPPD). Except under the flat prior, Bayesian logistic regression achieved slightly better predictive performance than linear regression, although differences in MLPPD between priors were minimal. Conversely, linear models provided more accurate estimates with narrower CIs than logistic models. In Scenario B, the uncertainty of the unregularized logistic model becomes especially evident, highlighting the importance of regularization in high-dimensional, low-information settings.

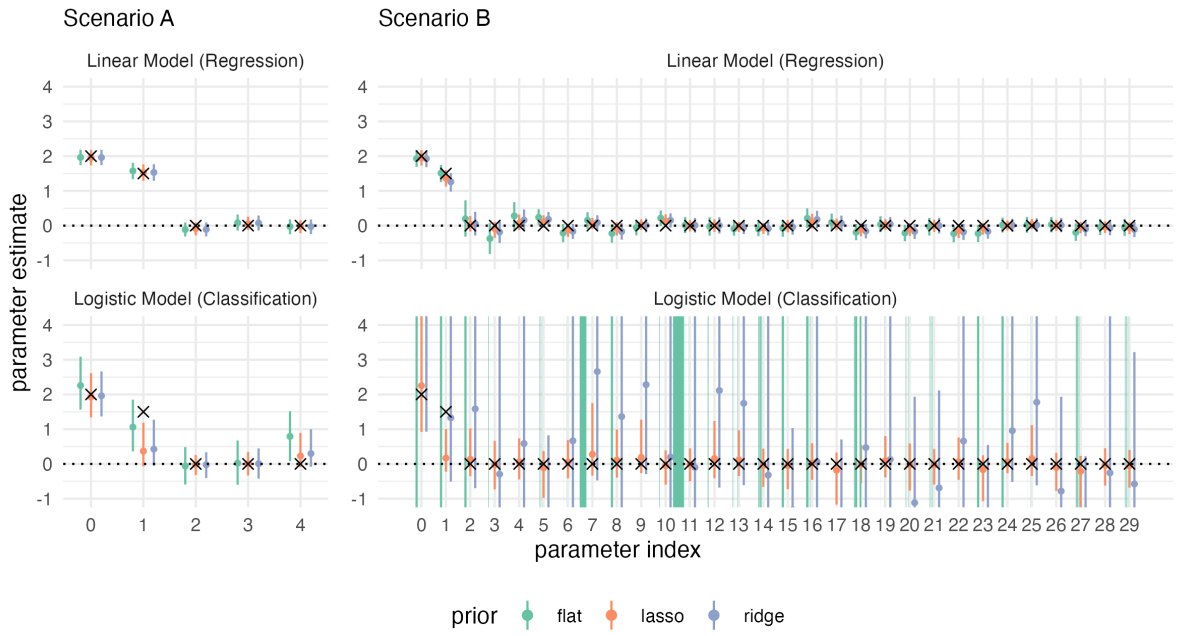

Figure 1: Estimated model parameters with 95% credibility intervals (CI). True parameter values are black crosses.

4.2 Performance of Approximate Inference Algorithms

In this second experiment, we compared LA and Metropolis-Hastings MCMC in Bayesian linear and logistic regression.

We generate 1,000 synthetic data sets with $n = 100$:

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{\theta} = (-0.5, 2, 1)$$
$$\text{linear: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{X\theta}, \boldsymbol{I}), \quad \text{logistic: } \boldsymbol{y} \mid \boldsymbol{\theta} \sim \text{Ber}(\sigma(\boldsymbol{X\theta}))$$

Each data set was used to fit one linear and one logistic model, using both inference approaches. We assumed $\boldsymbol{\theta} \sim \mathcal{N}(0, 10 \cdot \boldsymbol{I})$ and fixed residual variance at $\sigma^2 = 10$. For LA, we used `r-INLA` (Rue et al., 2009, `www.r-inla.org`) with settings to default from INLA to simple LA. For MCMC, we used the `MCMCglmm` package (Hadfield, 2010) with settings specifically to use Metropolis-Hastings, a relatively small sample size of 5,000, a burn-in period of 500, and a thinning interval of 10. For each method, we recorded CPU runtime, posterior means, and posterior standard deviations of the estimated parameters.

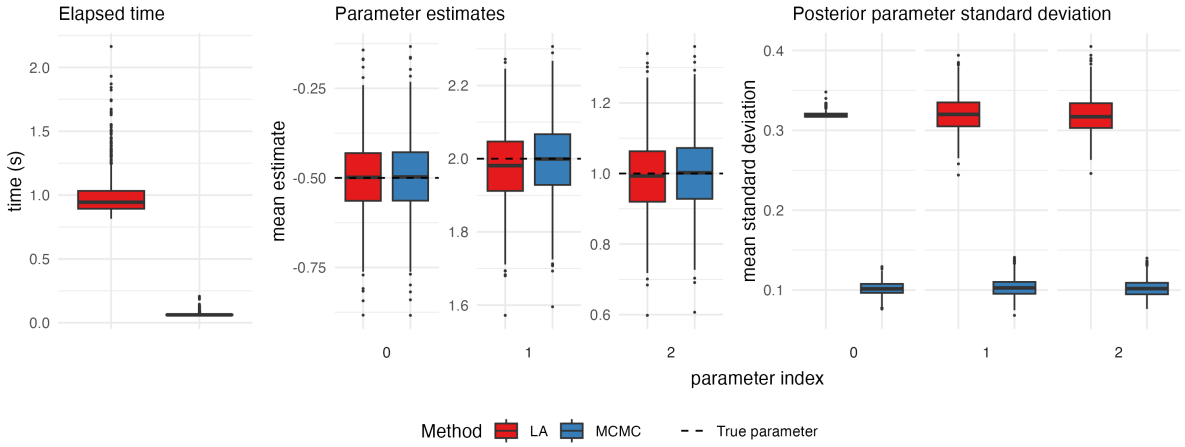

Figure 2: Bayesian **linear** regression: Comparison of LA (red) and MCMC (blue) across 1,000 simulations. MCMC is faster in this case. Both methods estimate the parameters accurately, though LA yields higher posterior uncertainty.

Results for the linear and logistic model can be seen in Figure 2 and Figure 3 respectively. In our experiment, inference with LA was generally slower than with MCMC methods. There was not much difference in the posterior parameter estimates in the case of the Bayesian linear model (Figure 2), although LA leads to a much higher standard deviation of parameters. In contrast, LA showed clear advantages in logistic regression (Figure 3): it yielded more accurate parameter estimates and lower posterior uncertainty than MCMC. These results highlight that the performance of approximate inference methods can differ significantly depending on the model and likelihood.

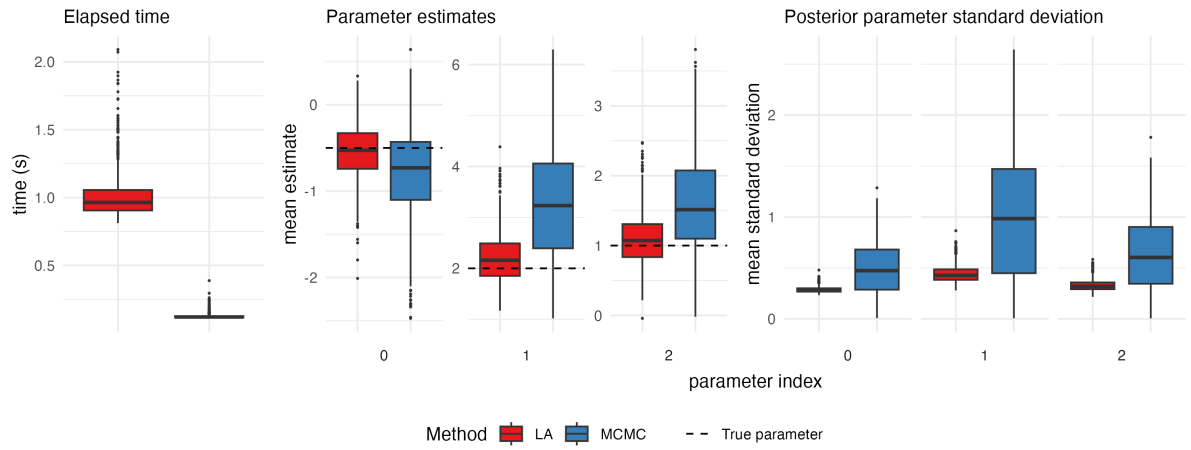

Figure 3: Bayesian **logistic** regression: Comparison of LA (red) and MCMC (blue) across 1,000 simulations. LA outperforms MCMC in both accuracy and precision of parameter estimates. While LA is slower to compute, it provides more stable estimates.

5 Conclusion and Outlook

In this paper, we reviewed the model specification, prior choice, and approximate inference methods for Bayesian generalized linear models. We detailed how priors can be used to stabilize estimation and how to implement regularization in a Bayesian setting. Using the example of logistic regression, we demonstrated the necessity of numerical methods for inference. We explained LA and MCMC, which form the basis of modern approximate inference methods for Bayesian GLMs.

We examined regularized Bayesian models in an applied example under challenging (synthetic) data conditions. We found that regularization can improve predictive performance and reduce the number of covariates falsely declared as informative, particularly in logistic regression. Comparing MCMC and LA revealed that LA can yield more precise estimates than MCMC despite its lower computational speed. However, this could be mitigated by the specific implementation used in the experiment. As these examples are meant to be illustrative, we caution against overgeneralizing and note that more efficient regularization and inference tools are readily available.

Some limitations of the described methods have been addressed above, such as the need for more complex prior distributions for real variable selection in Section 2.2. For scenarios that require more flexibility, the Bayesian framework can be extended in plenty of ways. For example, hierarchical (multilevel) GLMs introduce group-specific random effects and can be used for longitudinal data (Gelman et al., 2008), and structured additive regression extends Bayesian GLMs to work with nonlinear effects (Fahrmeir et al., 2010). Nevertheless, uncertainty quantification with Bayesian models sill relies on parametric distributional assumptions and can be sensitive to prior misspecification (Piironen and Vehtari, 2017). In contrast, methods like conformal prediction can be applied more generally while guaranteeing finite-sample and distribution-free coverage of prediction intervals (see e.g. Angelopoulos and Bates, 2022).

A Appendix

Notation

We denote prior parameters with $\breve{}$ and posterior parameters with $\hat{}$. Vectors are written in bold-face like so $\boldsymbol{x}$ and matrices are bold capital letters $\boldsymbol{X}$. In general, we assume n observations and p covariates. The intercept θ_0 is always included in $\boldsymbol{\theta}$ and thus

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Proofs and Derivations

Posterior of the Normal-Inverse-Gamma prior

For the model described in (3), the posterior distribution is calculated according to Fahrmeir et al. (2021) as

$$\begin{aligned} p(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y}) &\overset{\text{Bayes' rule}}{\propto} \mathcal{L}(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y}) p(\boldsymbol{\theta}, \sigma^2) \\ &= \mathcal{L}(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y}) p(\boldsymbol{\theta} \mid \sigma^2) p(\sigma^2) \\ &= \frac{1}{(\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\Big) \\ &\quad \frac{1}{(\sigma^2)^{p/2}} \exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{\theta} - \breve{\boldsymbol{\mu}})^\top \breve{\Sigma}^{-1} (\boldsymbol{\theta} - \breve{\boldsymbol{\mu}})\Big) \\ &\quad \frac{1}{(\sigma^2)^{\hat{a}+1}} \exp\Big(-\frac{\breve{b}}{\sigma^2}\Big), \end{aligned}$$

which can be shown to be NIG-distributed

$$\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y} \sim \text{NIG}(\hat{\boldsymbol{\mu}}, \hat{\Sigma}, \hat{a}, \hat{b})$$

with parameters

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \hat{\Sigma}(\breve{\Sigma}^{-1}\breve{\boldsymbol{\mu}} + \boldsymbol{X}^\top \boldsymbol{y}) \\ \hat{\Sigma} &= (\boldsymbol{X}^\top \boldsymbol{X} + \breve{\Sigma}^{-1})^{-1} \\ \hat{a} &= \breve{a} + \frac{n}{2} \\ \hat{b} &= \breve{b} + \frac{1}{2}(\boldsymbol{y}^\top \boldsymbol{y} + \breve{\boldsymbol{\mu}}^\top \breve{\Sigma}^{-1} \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}^\top \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}). \end{aligned}$$

For the conditional posteriors it holds that

$$\begin{aligned} \boldsymbol{\theta} \mid \sigma^2, \boldsymbol{y} &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \sigma^2 \hat{\Sigma}) \\ \boldsymbol{\theta} \mid \boldsymbol{y} &\sim \mathcal{T}(2\hat{a}, \hat{\boldsymbol{\mu}}, \hat{b}/\hat{a}\hat{\Sigma}). \end{aligned}$$

Posterior predictive distribution of the Normal-Inverse-Gamma prior

In the case of (3), the posterior predictive distribution is calculated as

$$\begin{aligned} p(\tilde{\boldsymbol{y}} \mid \boldsymbol{y}) &= \int\int p(\tilde{\boldsymbol{y}}, \boldsymbol{\theta}, \sigma^2) d\boldsymbol{\theta} d\sigma^2 \\ &= \int\int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta}, \sigma^2) d\boldsymbol{\theta} d\sigma^2 \\ &= \int\int \mathcal{N}(\tilde{\boldsymbol{y}} \mid \boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}) \mathrm{NIG}(\boldsymbol{\theta}, \sigma^2 \mid \hat{\boldsymbol{\mu}}, \hat{\Sigma}, \hat{a}, \hat{b}). \end{aligned}$$

According to e.g. Murphy (2007), the result is

$$\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \sigma^2, \boldsymbol{y} \sim \mathcal{T}(2\hat{a}, \tilde{\boldsymbol{X}}\boldsymbol{\theta}, \frac{\hat{b}}{\hat{a}}(\boldsymbol{I} + \tilde{\boldsymbol{X}}\hat{\Sigma}\tilde{\boldsymbol{X}}^\top))$$

with posterior predictive mean

$$\mathbb{E}_{\boldsymbol{\theta}}(\mathbb{E}_{\tilde{\boldsymbol{y}}}(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta}, \sigma^2, \boldsymbol{y}) \mid \sigma^2, \boldsymbol{y}) = \mathbb{E}(\tilde{\boldsymbol{X}}\boldsymbol{\theta} \mid \sigma^2, \boldsymbol{y}) = \tilde{\boldsymbol{X}}\boldsymbol{\theta},$$

as stated by Gelman, Carlin, Stern, Dunson, Vehtari and Rubin (2013). The posterior predictive variance $\frac{\hat{b}}{\hat{a}}\boldsymbol{I} + \frac{\hat{b}}{\hat{a}}\tilde{\boldsymbol{X}}\hat{\Sigma}\tilde{\boldsymbol{X}}^\top$ consists of measurement noise in the prior from $\frac{\hat{b}}{\hat{a}}$ and uncertainty in the parameter $\boldsymbol{\theta}$ from $\frac{\hat{b}}{\hat{a}}\tilde{\boldsymbol{X}}\hat{\Sigma}\tilde{\boldsymbol{X}}^\top$.

B Electronic Appendix

Data, code, and figures are provided in electronic format. All figures and scripts can be accessed at `https://github.com/lona-k/bayesian-GLMs-seminar`.

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association* **88**(422): 669–679.

Angelopoulos, A. N. and Bates, S. (2022). A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv:2107.07511 [cs].

Barbieri, M. M. (2015). Posterior Predictive Distribution, *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, pp. 1–6.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness, *Journal of the Royal Statistical Society. Series A (General)* **143**(4): 383–430.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan, *Journal of Statistical Software* **80**(1): 1–28.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals, *BIOMETRIKA* **97**(2): 465–480.

Celeux, G., Anbari, M. E., Marin, J.-M. and Robert, C. P. (2012). Regularization in Regression: Comparing Bayesian and Frequentist Methods in a Poorly Informative Situation, *Bayesian Analysis* **7**(2): 477–502.

Chien, Y.-F., Zhou, H., Hanson, T. and Lystig, T. (2023). Informative g-Priors for Mixed Models, *Stats* **6**(1): 169–191.

Dayanik, A., Lewis, D. D., Madigan, D., Menkov, V. and Genkin, A. (2006). Constructing informative prior distributions from domain knowledge in text classification, *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, Association for Computing Machinery, New York, NY, USA, pp. 493–500.

Dellaportas, P. and Smith, A. F. M. (1993). Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **42**(3): 443–459.

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**(2): 203–219.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. D. (2021). *Regression: Models, Methods and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Frühwirth-Schnatter, S. and Frühwirth, R. (2007). Auxiliary mixture sampling with applications to logistic models, *Computational Statistics & Data Analysis* **51**(7): 3509–3528.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalised linear models, *Biometrika* **85**(1): 215–227.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, 3 edn, Chapman and Hall/CRC, New York.

Gelman, A., Hwang, J. and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models, *The Annals of Applied Statistics* **2**(4): 1360–1383.

Ghosh, J., Li, Y. and Mitra, R. (2017). On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression.

Grammarly Inc. (2025). Grammarly (free version, 1.124), URL: `https://www.grammarly.com`. Last accessed: 02.07.2025.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package, *Journal of Statistical Software* **33**(2): 1–22.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika* **57**(1): 97–109.

Held, L. and Sabanés Bové, D. (2020). *Likelihood and Bayesian Inference: With Applications in Biology and Medicine*, Statistics for Biology and Health, Springer, Berlin, Heidelberg.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge Regression: Applications to Nonorthogonal Problems, *Technometrics* **12**(1): 69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics* **12**(1): 55–67.

Holmes, C. and Knorr-Held, L. (2003). Efficient simulation of bayesian logistic regression models.

Hosack, G. R., Hayes, K. R. and Barry, S. C. (2017). Prior elicitation for Bayesian generalised linear models with application to risk control option assessment, *Reliability Engineering & System Safety* **167**: 351–361.

Hsiang, T. C. (1975). A Bayesian View on Ridge Regression, *Journal of the Royal Statistical Society. Series D (The Statistician)* **24**(4): 267–268.

Lenk, P. J. and DeSarbo, W. S. (2000). Bayesian Inference for Finite Mixtures of Generalized Linear Models with Random Effects, *Psychometrika* **65**(1): 93–119.

MacKay, D. J. C. (1992). Bayesian Interpolation, *Neural Computation* **4**(3): 415–447.

Makalic, E. and Schmidt, D. F. (2016). High-dimensional bayesian regularised regression with the bayesreg package.

Mitchell, T. and Beauchamp, J. (1988). Bayesian Variable Selection in Linear-Regression, *Journal of the American Statistical Association* **83**(404): 1023–1032.

Murphy, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution.

Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods, *Department of Computer Science, University of Toronto Toronto, Ontario, Canada* .

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models, *Journal of the Royal Statistical Society. Series A (General)* **135**(3): 370–384.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which, *Bayesian Analysis* **4**(1): 85–117.

OpenAI (2025). ChatGPT o4-mini-high, `https://chat.openai.com/chat`. Last accessed: 19.06.2025.

Park, T. and Casella, G. (2008). The Bayesian Lasso, *Journal of the American Statistical Association* **103**(482): 681–686.

Piironen, J. and Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection, *Statistics and Computing* **27**(3): 711–735.

Polson, N. G., , James G., S., and Windle, J. (2013). Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables, *Journal of the American Statistical Association* **108**(504): 1339–1349.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian models by using Integrated Nested Laplace Approximations, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **71**(2): 319–392.

Scott, S. L. (2011). Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models, *Statistical Papers* **52**(1): 87–109.

Sondhi, A., Segal, B., Snider, J., Humblet, O. and McCusker, M. (2021). Bayesian additional evidence for decision making under small sample uncertainty, *BMC Medical Research Methodology* **21**(1): 221.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1): 267–288.

Tierney, L., and Kadane, J. B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association* **81**(393): 82–86.

Tyralis, H. and Papacharalampous, G. (2024). A review of predictive uncertainty estimation with machine learning, *Artificial Intelligence Review* **57**(4): 94.

van Erp, S., Oberski, D. L. and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression, *Journal of Mathematical Psychology* **89**: 31–50.

West, M., , P. Jeff, H., and Migon, H. S. (1985). Dynamic Generalized Linear Models and Bayesian Forecasting, *Journal of the American Statistical Association* **80**(389): 73–83.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions, *Bayesian Inference and Decision techniques* .

Research Instruments

All thoughts presented in this paper are my own, unless otherwise noted. Neither ChatGPT nor any other LLM directly wrote any part of this paper. Throughout this work, I used Grammarly (Grammarly Inc., 2025) to improve and proofread the text. ChatGPT (OpenAI, 2025) was used, along with traditional research methods, as a starting point to find additional relevant literature (such as the original author and paper that established a method, e.g. Tibshirani (1996) for the LASSO method). Despite this, I have read and checked all listed sources before using and citing them in this work. ChatGPT was also used as a conversation partner to help me understand difficult parts of papers and theories, and to come to terms with the philosophical ideas behind Bayesian statistics. I consulted ChatGPT, although sparingly and with limited helpful results, about some errors in the code for the examples in this paper. More abstractly, I used ChatGPT to format this document and to assist me in solving difficult LaTeX errors, such as issues with automated citations and rendering.

Declaration of Authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid, and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources, I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

München, 03.07.2025
Location, date

Name