Probabilistic Machine Learning

Bayesian (Generalised) 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

This should be an abstract

Contents

1 Introduction

Bishop (2019) introduced this and that. Another statement that needs a reference, but the authors are not named directly (Bishop, 2019). Another statement where the reference is just one possible source (see, e.g., Bishop, 2019).

2 Linear Bayesian Model

The (frequentist) Linear Regression Model is probably the most widely used model in statistics and machine learning. Both the frequentist and the Bayesian Linear Models are described in many introductory texts on statistical modelling, such as Fahrmeir et al. (2021) or Gelman et al. (2013).

2.1 Model definition

We observe an i.i.d. sample $\boldsymbol{D} = ((y_1, \boldsymbol{x}_1), \dots, (y_n, \boldsymbol{x}_n)) = (\boldsymbol{y}, \boldsymbol{X})$ and assume a linear relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$. 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. A condition on $\boldsymbol{X}$ is always implicit.

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 (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.

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

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 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.

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

Uninformative Prior

The idea of an uninformative (or flat) 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 flatten the NIG prior by setting

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

and choosing $\breve{a} = -\frac{p}{2}$ 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 (3).
The prior distributional assumptions would then be:

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

Note that we generally have to be careful with completely flat priors; it is necessary to check if the resulting posterior is proper (which is the case here).

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 vs. variance. In frequentist statistics, we minimize the Penalized Least Squares criterion (PLS)

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

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:

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2\boldsymbol{I}),^{2} \\ \boldsymbol{\theta} &\sim \text{regularization prior} \\ \sigma^2 &\sim \mathrm{IG}(\breve{a}, \breve{b}), \end{aligned} \tag{5}$$

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

[1] Informally stated for demonstational purposes.

[2] 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 $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$ and the Bayesian analogue (?MacKay, 1992, ?, e.g.) specifies

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

with τ^2 controlling the degree of regularization akin to the role of λ. In constrast to λ, τ^2 does not need to be set in advance or 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 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)

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

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 Bayesian Lasso does not promote a sparse solution, discrete-mixture Spike-and-Slab priors (Mitchell and Beauchamp, 1988) (which are necessary for categorical coraviates) 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 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 (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 [3]

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

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 (Held and Sabanés Bové, 2020).

Since we defined the non-information prior (4) as a special case of the NIG-distributed prior, we can use (8) to directly calculate the posterior mean and variance 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 $\hat{\boldsymbol{\beta}}_{LS}$ (7), so 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 $\hat{\boldsymbol{\beta}}_{LS}$, since the likelihood (and thus the data) dominates the posterior.

Bayesian Ridge regression is simply the NIG case in (3) with finite $\hat{\Sigma}$, resulting in the same posterior update in (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 depth on approximate inference for Bayesian regression models in Section 3.3.

Posterior predictive distribution

In many applications, we care more about 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 $\boldsymbol{D}$, than about $\boldsymbol{\theta}$ itself. The Bayesian answer to this is the *posterior predictive distribution* (Barbieri, 2015, see e.g)

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

which is an average of conditional probabilities over the posterior distribution of $\boldsymbol{\theta}$.[4]
For the NIG prior in (3), one can show[5] that

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

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 (n.d.).

If no closed form exists, the posterior predictive distribution can also be simulated (see Section 3.3)

[3] For the full calculation see Appendix A

[4] A note on intuition: In essence, the posterior predictive distribution 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. $p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$.

[5] see Appendix A

3 Logistic Bayesian Model

3.1 Bayesian Generalized Linear Regression Model

Just as linear regression models are easily generalized to a multitude of distributional assumptions (Nelder and Wedderburn, 1972), the same can be done with Bayesian linear models. Bayesian generalized linear models (GLM) (see e.g. West et al., 1985) generally assume

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

where F is any distribution from the exponential family 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, but note that in practice modelling might be complicated, because the choice of prior also depends on the Link function (West et al., 1985).

3.2 Bayesian Logistic Regression Model

We will illustrate how to work with Bayesian GLMs at the example of Logistic regression models, which have a wide variety of applications in statistics, from text classification to genetic modelling.

Model definition

The Bayesian Logistic model is defined as

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

What makes this Binomial model logistic is the link function. We recognize it from machine learning settings as the logistic or sigmoid functions. Other link function, such as the probit functions, can also be used.

Prior choice

As introduced for the Bayesian linear model (3), we can use a Gaussian prior for $\boldsymbol{\theta}$, but the solution to this is not analytically available as it is for the linear model. An uninformative prior can be set analogue to (4). It is an improper posterior and results in a proper posterior, although without any known distribution type. This makes the use of approximate Bayesian inference methods necessary in both cases (see Section 3.3).

A common issue in logistic regression is separation (i.e. perfect classification), which leads instable models. Heavier-tailed prior distributions have been proposed to mitigate this issue in Bayesian logistic regression. Prominent choices are the Student t-distribution, which was introduced by Gelman et al. (2008) as a prior for low-information settings that results in higher model stability, or the Cauchy distribution (Ghosh et al., 2017, Gelman et al., 2008).

In general, **regularization** can be achieved with 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). Note that the Student t-distribution also has a regularizing effect (Gelman et al., 2008).

3.3 Approximate Bayesian inference

Unlike for the linear model, Bayesian inference with closed-form posteriors is in most cases not possible for the logistic model. In order to sample from the parameter and predictive posterior distributions, we need to use approximate Bayesian methods.

Sampling from the posterior with MCMC methods

Although Markov Chain Monte Carlo (MCMC) methods make no (explicit) assumption about the form of the posterior distribution, they perform best if the parameter posterior is known up to a constant (which in most cases is the normalization constant).

In the simplest form of MCMC, the Metropolis-Hastings algorithm can be used to generate K samples from the posterior parameter disribution as follows:

1. Initialize $\boldsymbol{\theta}^{(0)}$ and set $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(0)}$
2. For $k = 1, \dots, K$
   (a) Draw $\boldsymbol{\theta}^{(*)}$ from the *proposal distribution* $q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(k)})$
   (b) calculate the *accceptance 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) Draw $u \sim \text{Uni}[0, 1]$
   (d) Make a decision

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

Note that by construction, the samples are (sometimes heavily) correlated and that the number of repetitions necessary until convergence depends on $\boldsymbol{\theta}^{(0)}$.

As we can see, the acceptance probability strongly depends on the quality of the proposal distribution. If it is a close approximation of the posterior, i.e. $p(\boldsymbol{\theta} \mid \boldsymbol{y})p(\boldsymbol{\theta}) \approx q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$, it follows that $\alpha \approx 1$ and the proposal is accepted in most cases. Otherwise, the Metropolis-Hastings rejections correct for approximation errors as needed (Scott, 2011). A common choice for the proposal distribution is to draw from a Gaussian distribution using iteratively weighted least squares (IWLS) (see e.g. Gamerman, 1998, Lenk and DeSarbo, 2000, Scott, 2011):

$$q(\boldsymbol{\theta}^{(*)} \mid \boldsymbol{\theta}^{(k)}) \sim \mathcal{N}(\boldsymbol{\theta}^{(k)} \mid -H^{-1}),$$

where $-H^{-1}$ is the Hessian of $\log p(\boldsymbol{\theta}^{(k)} \mid \boldsymbol{y})\ p(\boldsymbol{\theta}^{(k)})$.[6] Scott (2011) also argues that t-distributed proposal densities are generally preferred for their heavier tails and thus increased randomness of the algorithm.

Several other MCMC algorithms, such as Hamiltonian Monte Carlo, Gibbs sampling, or combined Metropolis-Gibbs sampling, have also been proposed (see e.g. Dellaportas and Smith, 1993). Polson et al. (2013) introduced a specialized algorithm for inference in regularized Bayesian generalized models.

Another interesting solution to inference in Bayesian GLMs without conjugate priors is to use data augmentation (Albert and Chib, 1993) with auxiliary variables. This was refined by Holmes (n.d.) for Bayesian logistic regression. They use a Gaussian scale mixture and additional auxiliary variables to reformulate the Logistic model, which results in being able to make use of the Gaussian-Gaussian conjugate for inference. While Holmes (n.d.), Frühwirth-Schnatter and Frühwirth (2007) use Gibbs sampling to estimate the posterior distribution of hyperparameters of the scale mixture model, we can also use data augmentation with the MH algorithm (Scott, 2011).

Full Bayes with Laplace Approximation

In contrast to MCMC methods, Laplace Approximation (LA) approximates the full posterior distribution. The general idea of LA is represent posterior with a Gaussian using Taylor expansion, i.e.

$$\boldsymbol{\theta} \mid \boldsymbol{y} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\Sigma}). \tag{10}$$

To estimate the mean $\hat{\boldsymbol{\mu}}$ and variance $\hat{\Sigma}$ of this distribution, we find the maximum posterior estimate $\hat{\boldsymbol{\beta}}_{MAP}$ by maximizing the (real) posterior with standard optimization methods. We then set

$$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\theta}}_{MAP}, \quad \hat{\Sigma} = H^{-1}(\hat{\boldsymbol{\theta}}_{MAP}).$$

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

[6] Note that the symmetry of the Gaussian distribution simplifies the Metropolis-Hastings algorithm to the Metropolis algorithm, where the acceptance probability can be calculated with

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

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

As we can see by the assumed prior, simple LA is not applicable for hierarchical models. Rue et al. (2009) proposed an altered algorithm based on INLA, for LA in latent Gaussian models. Although MCMC could also be used in this setting, Rue et al. (2009) showed their approach to be faster and more generally applicable.

Posterior predictive distribution

In a binary classification setting such as Logistic regression, we obtain the posterior predictive distribution by calculating the distribution for the positive class[7] $p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{\theta}, \boldsymbol{y})$ and inferring the negative class by computing $p(\tilde{\boldsymbol{y}} = 0 \mid \boldsymbol{\theta}, \boldsymbol{y}) = 1 - p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{\theta}, \boldsymbol{y})$.

As MCMC results in samples from the posterior, we can simply use the samples $\boldsymbol{\theta}_k$ to approximate the posterior predictive distribution with

$$p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{\theta}, \boldsymbol{y}) \approx \frac{1}{K} \sum_{k=1}^{K} \frac{\exp(\tilde{\boldsymbol{X}}\boldsymbol{\theta})}{1+\exp(\tilde{\boldsymbol{X}}\boldsymbol{\theta})}. \tag{11}$$

In the case of Laplace Approximation, we can use the approximation of the posterior parameter distribution (QUOTE EQUATION) to calculate the posterior predictive distribution as

$$\begin{aligned}p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{\theta}, \boldsymbol{y}) &= \int p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{\theta})\; p(\boldsymbol{\theta} \mid \boldsymbol{y})\; d\boldsymbol{\theta} \\ &= \int \frac{\exp(\tilde{\boldsymbol{X}}\boldsymbol{\theta})}{1+\exp(\tilde{\boldsymbol{X}}\boldsymbol{\theta})}\; \mathcal{N}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\beta}}_{MAP}, H^{-1}(\hat{\boldsymbol{\beta}}_{MAP}))\; d\boldsymbol{\theta}.\end{aligned}$$

A simpler way for approximating the PPD with LA would be to draw samples $\boldsymbol{\theta}_s$ with $s = 1, \ldots, S$ from the approximated posterior and use (11) analogue to MCMC PPD approximation.

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

4 Simulation Study

4.1 Regularization and variable selection

Experiment Setup

- Linear Regression - Logistic Regression

Results and Evaluation

4.2 Performance of approximate inference algorithms in Bayesian regression

setup with LA vs. MCMC (MH, HM, Gibbs, etc.?) in a regression environment Evaluation

5 Conclusion

A concise summary of contents and results

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 $\boldsymbol{X}$.

n observations

p covariates

$\boldsymbol{\theta}$ regression weights

Distributions

When deriving equations, we assume the following probability density functions and parameter placements:

$\mathcal{N}(\mu, \sigma^2)$ Gaussian distribution with mean μ and variance σ^2

Gamma distribution

$IG(a, b)$ Inverse Gamma distribution with scale parameter a and location parameter b

(multivariate) Student t-distribution

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

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

which is 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

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

According to e.g. Murphy (n.d.), 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 et al. (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 form. All figures and scripts are avaivable from `https://github.com/lona-k/probML_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. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Barbieri, M. M. (2015). Posterior Predictive Distribution, *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, pp. 1–6. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat07839.

Bishop, C. M. (2019). *Pattern recognition and machine learning*, Information Science and Statistics, Springer Science+Business Media, LLC, New York, NY.

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

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. Publisher: [Royal Statistical Society, Oxford University Press].

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., 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. Publisher: Institute of Mathematical Statistics.

Ghosh, J., Li, Y. and Mitra, R. (2017). On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression. arXiv:1507.07170 [stat].

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. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics* **12**(1): 55–67. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

Holmes, K.-H. (n.d.). Efficient simulation of Bayesian logistic regression models.

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.

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. (n.d.). Conjugate Bayesian analysis of the Gaussian distribution.

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. Publisher: [Royal Statistical Society, Oxford University Press].

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. Publisher: International Society for Bayesian Analysis.

Park, T. and Casella, G. (2008). The Bayesian Lasso, *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION* **103**(482): 681–686.

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. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.2013.829001.

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.

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

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* . Publisher: Elsevier Science.

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 unauthorised 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.

Location, date

Name