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 (2006) introduced this and that. Another statement that needs a reference, but the authors are not named directly (Bishop, 2006). Another statement where the reference is just one possible source (see, e.g., Bishop, 2006).

# 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., n.d.) or (Gelman et al., n.d.).

## 2.1 Model definition

We observe an i.i.d. sample $\boldsymbol{D} = ((y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)) = (\boldsymbol{y}, \boldsymbol{X})$ 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 $\sigma^2$ 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 $\sigma^2$, 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 $\sigma^2$.

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

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 $\sigma^2$ is conjugate to the Gaussian conditional distribution of $\boldsymbol{\theta}$, the joint prior of $\boldsymbol{\theta}$ and $\sigma^2$

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

follows a Normal Inverse Gamma (NIG). 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:

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

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, n.d.).

### Regularization Priors

Regularization (or penalization) regulates the trade off between model fit and complexity, or equivalently bias vs. 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:

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

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** uses $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$ and the Bayesian analogue specifies

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

with $\tau^2$ controlling the degree of regularization akin to the role of $\lambda$. In constrast to $\lambda$, $\tau^2$ 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 $\tau^2$, e.g. $\tau^2 \sim \text{IG}(\breve{a}_\tau, \breve{b}_\tau)$, and estimate it alongside $\boldsymbol{\theta}$ and $\sigma^2$.

**Lasso regularization** uses $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ to perform variable selection by setting elements $\theta_j$ 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 (n.d.)

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

where the regularization parameter $\lambda^2$ 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, n.d.) or the heavy-tailed horseshoe prior (Carvalho et al., n.d.) 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

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

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

$$\hat{\boldsymbol{\theta}}_{LS} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}). \tag{7}$$

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 $\sigma^2$ 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, n.d.).

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 $\hat{\boldsymbol{\beta}}_{LS}$ ((**??**)), 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 (n.d.). We will go more into depth on approximate inference for Bayesian regression models in Section 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*

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

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.

If no closed form exists, the posterior predictive distribution can also be simulated.

[3] For the full calculation see Appendix A

[4] A note on intuition: In essence, the posterior predictive distribution is the marginal distribution of $\tilde{\boldsymbol{y}}$, only for new response values $\tilde{\boldsymbol{y}}$, i.e. $p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. We recognize it from Bayes' rule as the normalization constant.

[5] see Appendix A

# 3 Logistic Bayesian Model

# 4 Simulation Study

# 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 $\mu$ and variance $\sigma^2$

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. (n.d.) 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

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

According to Kevin P. Murphy (n.d.) 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. (n.d.). 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

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Springer.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (n.d.). The horseshoe estimator for sparse signals, **97**(2): 465–480. Num Pages: 16 Place: Oxford Publisher: Oxford Univ Press Web of Science ID: WOS:000280559700015.

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

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

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

Kevin P. Murphy (n.d.). *Probabilistic Machine Learning: Advanced Topics*, MIT Press.

Mitchell, T. and Beauchamp, J. (n.d.). Bayesian variable selection in linear-regression, **83**(404): 1023–1032. Num Pages: 10 Place: Alexandria Publisher: Amer Statistical Assoc Web of Science ID: WOS:A1988R852000013.

Park, T. and Casella, G. (n.d.). The bayesian lasso, **103**(482): 681–686. Num Pages: 6 Place: Alexandria Publisher: Amer Statistical Assoc Web of Science ID: WOS:000257897500025.

ZELLNER, A. (n.d.). On assessing prior distributions and bayesian regression analysis with g-prior distributions. 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