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

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

2 Linear Bayesian Model

The (frequentist) Linear Regression Model is probably the most well-known and most used model in statistics. Both the frequentist and the Bayesian model are described in many introductory texts into statistical modelling, such as (**?**) or (**?**).

2.1 Model definition

Given a random sample $\boldsymbol{D} = ((y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n))$ of size n, we assume a linear relationship between the input observations $\boldsymbol{X}$ (also sometimes called features or covariates) and the target variable $\boldsymbol{y}$. The model is defined by the following distribution:

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

where the weight parameter $\boldsymbol{\theta}$ and the variance σ^2 are estimated to get the fitted model. A condition on the data $\boldsymbol{X}$ is always implicit.

To view Linear Regression from a Bayesian perspective, we simply change our perception of the parameters: instead of viewing them as scalars, we now see them as random variables. This means that all we need to change about (1) is conditioning on the parameters:

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

Note that to predict multiple outputs, an extension of both the frequentist and the Bayesian model to Multivariate Linear Regression is possible.

2.2 Prior choice

Gaussian (Inverse Gamma) Prior

To fully estimate a Bayesian model, we need to specify prior distributions for the regression parameters. Since $\boldsymbol{y}$ follows a Gaussian distribution, it seems natural at first to also set the distribution of the regression weights $\boldsymbol{\theta}$ as a Gaussian distribution in order to make use of the Gaussian-Gaussian conjugate for prior distribution and likelihood.

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

where $\breve{\boldsymbol{\mu}}, \breve{\Sigma}, \breve{a}$ and $\breve{b}$ are the prior parameters and we set the prior distribution of σ^2 as an Inverse Gamma distribution. 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 because of the conjugate between the Gaussian and Inverse Gamma distributions. 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 problem with this prior distribution setup is that we would need to specify four prior parameters, which is difficult in the case of having little to no prior knowledge. Especially $\breve{\boldsymbol{\mu}}$ and $\breve{\Sigma}$ are normally chosen based on past results. This is why we will try to construct a non-informative prior for the Bayesian Linear Model for such cases. The idea of an uninformative prior is to maximize the influence of the data on the posterior in absence of prior knowledge.

We set $\breve{\boldsymbol{\mu}} = \mathbf{0}$ and $\breve{\Sigma}^{-1} = \mathbf{0}$, which is roughly equivalent to assuming infinite prior variance. 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). For the distribution of σ^2, we set $\breve{a} = -\frac{p}{2}$ and $\breve{b} = 0$, where p is the number of features in the model. The prior distributional assumptions would then be:

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

Note that we generally have to be careful with completely flat priors, because they can result in improper priors. Generally, we need to check if the resulting posterior is proper, which is the case here.

There are many other ways to motivate an uninformative prior, such as using Jeffrey's prior. Another good solution for use-cases with little prior knowledge that still require a proper posterior is Zellner's g-prior (**?**).

Regularization Priors

Regularization (also called penalization) is a technique to regulate the tradeoff between model complexity and adjustment to the data. It can also be regarded as regulating the bias-variance-tradeoff. In frequentist statistics, penalized (linear) regression estimates the regression weights $\boldsymbol{\theta}$ by minimizing the Penalized Least Squares criterion (PLS)

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

The balance of the tradeoff and therefore the strength of regularization is controlled by the hyperparameter $\lambda > 0$.

To regularize a Bayesian model, we need to specify a so-called regularization prior for $\boldsymbol{\theta}$. We assume

$$\begin{aligned} \boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2 &\sim \mathcal{N}(\boldsymbol{X\theta}, \sigma^2\boldsymbol{I}),^1 \\ \boldsymbol{\theta} &\sim \text{regularization prior} \\ \sigma^2 &\sim \mathrm{IG}(\breve{a}, \breve{b}), \end{aligned} \tag{5}$$

There are many options for regularization priors in Bayesian statistics, but in the following we will focus on the cases where there is an equivalent in frequentist Statistics.

Ridge regularization uses $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$ and is also called L2 regularization. Bayesian Ridge Regression (REFS) specifies the prior distribution of the weights $\boldsymbol{\theta}$ as

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

where τ^2 is a hyperparameter that regulates the degree of regularization akin to the role of λ. In contrary to λ, τ^2 does not need to be set in advanced or optimized as a hyperparameter. We can simply build a hierarchical model by specifying a prior for τ^2 and estimate it directly. A common choice is $\tau^2 \sim \text{IG}(\breve{a}_\tau, \breve{b}_\tau)$.

Lasso regularization uses $\text{pen}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ and is also calles L1 regularization. In contrary to Ridge regularization, Lasso can perform real variable selection by setting elements θ_j of $\boldsymbol{\theta}$ to 0 during estimation. We say that Lasso regularization promotes a *sparse* solution.

Bayesian Lasso regularization uses conditional Laplace priors for $\boldsymbol{\theta} \mid \sigma^2$. As **?** point out, it can also be represented as a hierarchical scale-mixture model, which specifies the priors as

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

where λ^2 is the regularization parameter akin to frequentist regularization. Similarly to Bayesian Ridge regression, we can set a (hyper-) prior for λ. (REF) propose e.g. $\lambda^2 \sim \text{G}(\breve{a}_\lambda, \breve{b}_\lambda)$.

Unfortunately, the Bayesian Lasso does not promote a sparse solution, which makes its possibilities for application rather limited. There are however a multitude of other regularization priors that can be used for variable selection with a sparse solution. Popular choices are Spike and Slab priors (**?**) and the horseshoe prior (**?**).

2.3 Bayesian inference with closed form priors

Parameter posterior distribution

In a frequentist linear model, we estimate $\boldsymbol{\theta}$ via LS-estimation (or Maximum Likelihood Estimation) by solving the optimization problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}).$$

[1] 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 previsously established definitions because we aim for an understandable explanation of the basic concept of Bayesian regularization.

The solution is the LS-estimator $\hat{\boldsymbol{\theta}}_{LS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$. To quantify the uncertainty of the estimation, we use the Law of Large Numbers to estimate that

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

and calculate confidence intervals for $\hat{\boldsymbol{\theta}}_{LS}$. Although this is useful, we can only use it to gain a sense of uncertainty of our estimation and have gained no more information about the distribution of the *real* parameter $\boldsymbol{\theta}$.

In Bayesian statistics on the other hand, we can calulate the posterior distribution of $\boldsymbol{\theta}$ by updating the prior distribution with observed data using Bayes' rule:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{8}$$

To make this more clear, we will show this on the example of the (marginal) NIG prior distribution for $\boldsymbol{\theta}$ introduced in (3). In this case, we estimate two parameters, $\boldsymbol{\theta}$ and σ^2. We are interested in their joint posterior.

$$p(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y}) \overset{(8)}{\propto} \mathcal{L}(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y})p(\boldsymbol{\theta}, \sigma^2)\ ^2$$

The result is a NIG distribution[3] with posterior mean and variance

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \hat{\Sigma}(\breve{\Sigma}^{-1}\breve{\boldsymbol{\mu}} + \boldsymbol{X}^\top \boldsymbol{y}) \\ \hat{\Sigma} &= (\boldsymbol{X}^\top \boldsymbol{X} + \breve{\Sigma}^{-1})^{-1}. \end{aligned} \tag{9}$$

The posterior mean $\hat{\boldsymbol{\mu}}$ can be used as a point estimate for $\boldsymbol{\theta}$. Alternatives would be the posterior mode, which in the case of the NIG-distribution is equal to the posterior mean. To quantify uncertainty about $\boldsymbol{\theta}$, we can derive Credibility Intervals directly form $p(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y})$ (**?**).

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

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} \\ \hat{\Sigma} &= \boldsymbol{X}^\top \boldsymbol{X}. \end{aligned}$$

We can see that the posterior mean $\hat{\boldsymbol{\mu}}$ is equivalent to (**??**). This means that a Bayesian linear regression model with a non-informative prior is equivalent to the frequentist linear regression model. Intuitively, this makes a lot of sense: if we include (close to) no prior information into the model, the posterior distribution is dominated by the likelihood and only the information drawn directly from the data influences the posterior distribution.

In general, one can use this construct to show that $\hat{\boldsymbol{\mu}}$ becomes more similar to $\hat{\boldsymbol{\theta}}_{LS}$ if we have less (certain) prior information about θ, i.e. if the prior variance $\breve{\Sigma}$ is large.

[2] Where $\mathcal{L}$ is the likelihood of $p(\boldsymbol{\theta}, \sigma^2)$, i.e. the likelihood of the NIG distribution

[3] For the full calculation see Appendix A

Since Bayesian Ridge regression is setup as another special case of (3), the posterior distribution of $\boldsymbol{\theta}$ is estimated in exactly the same manner.
Note that in Lasso regression, the posterior parameter distribution has no analytical solution, but we can easily sampled from it using a Gibbs sampling algorithm as described by **?**. We will go more into depth on approximate inference for Bayesian regression models in Section 3.

Posterior predictive distribution

Traditionally, Bayesian statistics is focused on deriving properties of the posterior parameter distributions. But especially in a Machine Learning context, we are interested in making predictions based on new unseen data $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{y}})$.
Rather than using a single weight vector to make predictions (as we would do in the frequentist case with $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}}$), we use the posterior marginal distribution of $\boldsymbol{y}$:

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

which is called the *posterior predictive distribution*. It is the posterior equivalent to the prior marginal distribution of $\boldsymbol{y}$ and we recognize it from Baye's rule as the normalization constant.

If we want to make a prediction for ty, this equates to calculating

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

where $\int p(\tilde{\boldsymbol{y}} \mid \boldsymbol{\theta})$ is the Likelihood for the new data $\tilde{\boldsymbol{y}}$. Generally speaking, the posterior predictive distribution is an average of conditional probabilities over the posterior distribution of $\boldsymbol{\theta}$.

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

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

The result is

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

Interestingly, the posterior predictive mean of the t-distribution $\tilde{\boldsymbol{X}}\boldsymbol{\theta}$ is equivalent to the calculation for a prediction in the frequentist case.

If there is no analytical solution avaivable, the posterior predictive distribution can also be simulated, as will be described in Section 3.

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 $\boldsymbol{X}$.

$\boldsymbol{\theta}$ 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 **?** as

$$
\begin{aligned}
p(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y}) &\overset{(8)}{\propto} \mathcal{L}(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y}) p(\boldsymbol{\theta}, \sigma^2) \\
&= \mathcal{L}(\boldsymbol{\theta}, \sigma^2 \mid \boldsymbol{y}) p(\boldsymbol{\theta} \mid \sigma^2) p(\sigma^2) \\
&= \frac{1}{(\sigma^2)^{n/2}} \exp\big(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\theta})^\top (\boldsymbol{y} - \boldsymbol{X\theta})\big) \\
&= \frac{1}{(\sigma^2)^{p/2}} \exp\big(-\frac{1}{2\sigma^2}(\boldsymbol{\theta} - \breve{\boldsymbol{\mu}})^\top \breve{\Sigma}^{-1} (\boldsymbol{\theta} - \breve{\boldsymbol{\mu}})\big) \\
&= \frac{1}{(\sigma^2)^{\breve{a}+1}} \exp\big(-\frac{\breve{b}}{\sigma^2}\big),
\end{aligned}
$$

which is NIG-distributed

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

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

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

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.

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