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, Carlin, Stern, Dunson, Vehtari and Rubin (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

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

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):

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

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.

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

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:

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

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, 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)

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

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:

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

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

[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 (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 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 (see e.g. 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* (PPD) (Box, 1980, 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 PPD can also be simulated (see Section 3.3)

[3] For the full calculation see Appendix A

[4] 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. $p(\boldsymbol{y}) = \int p(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$.

[5] see Appendix A

3 Logistic Bayesian Model

3.1 Bayesian Generalized Linear Regression 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

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

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 (West et al., 1985).

3.2 Bayesian Logistic Regression 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 modelling. (SOURCE)

Model definition

The Bayesian logistic regression model is defined as

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

where $\sigma(\boldsymbol{x}_i\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}_i\boldsymbol{\theta})}$ is the logistic (sigmoid) function. Other choices like the probit link can also be used.

Prior choice

Unlike the Gaussian linear model, the logistic likelihood breaks conjugacy. Nevertheless, we can use a Gaussian prior (3) or an (improper) flat prior (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 also 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).

3.3 Approximate Bayesian inference

Unlike for the linear model, Bayesian inference with closed-form posteriors is not possible in most cases. 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[6]proceeds as follows:

1. Initialize $\boldsymbol{\theta}^{(1)}$
2. For $k = 1, \ldots, K$
   (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) 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 centered at the current state with covariance given by the (estimated) negative inverse Hessian of the log–posterior, often obtained via IWLS (Gamerman, 1998, Lenk and DeSarbo, 2000, Scott, 2011):[7]

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

(Scott, 2011) argues that using heavier-tailed proposals (e.g. Student–t) can improve mixing by allowing larger moves.

Beyond Metropolis–Hastings, several advanced samplers are popular:

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

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

[7] 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, n.d., Frühwirth-Schnatter and Frühwirth, 2007, Scott, 2011).

Full Bayes with Laplace Approximation

In contrast to MCMC methods, Laplace Approximation (LA) approximates the full posterior distribution by assuming 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 Approximation.

Posterior predictive distribution

In a binary classification setting, we obtain the PPD by calculating the distribution of the positive class[8] $p(\tilde{\boldsymbol{y}} = 1 \mid \boldsymbol{\theta}, \boldsymbol{y})$ 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 (10) or
- use the LA-approximated PPD and compute

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

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

4 Simulation Study

4.1 Regularization and variable selection

In this section, we apply thee of the previously discussed prior distributions to Linear and Logistic regression.
We create a synthetic data sets (consisting of training and test data) from each of two different setting, referred to as scenario A and B.

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

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

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

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

For each scenario, we fit three linear and logistic regression models. We set the Markov chain length as 20000, a burnin period of 1000 and a thinning interval of 10. We used three different prior distributions: A flat prior as a benchmark, using the conjugate setting described in Equation 4 with the R-package `brms`. The Gaussian prior variance was set to 10^6 and for the Gamma prior we set $a = 0.001, b = 0.001$ which results in an uninformative prior. To examine regularization, we used a Ridge prior (see Equation 5) and a Lasso prior (see Equation 6) with the `bayesreg` package and automatic optimization of the regularization parameters τ^2 and λ^2.

Since Ridge and Lasso regularization cannot perform variable selection, we included the coefficients based on the Bayesian credibility interval criterion described in van Erp et al. (2019). We were interested in whether the models could accurately select relevant variables (1), and in the correct prediction of the outcome $\boldsymbol{y}$ (2).

To evaluate (1), we calculated correctly identified influential coefficients (Hits) and the falsely as influential declared coefficients (FP) For evaluation of predictive accuracy (2), we used the test data to calculate the mean (or: expected) log predictive posterior density (MLPPD), which was proposed by Gelman, Hwang and Vehtari (2013) and used for similar purposes by van Erp et al. (2019). Since the log-likelihood is a proper scoring rule, it is well-suited to evaluate Bayesian regression models. Results for these metrics can be seen in Table 1. In Figure 1, we assessed the uncertainty of the model parameter estimates.

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 under scenario A and B. Linear Regression regardless of prior identified influential coefficients more accurately (Hits) and misclassified less coefficients as influential (FP). We see the effect of regularization more strongly in Logistic regression, where regardless of the scenario, less coefficients were declared influential. Regularization priors performed better in the non-informative scenario B. Except for the flat prior, Bayesian logistic regression made more accurate predictions than linear regression, but the MLPPD is not very different between regularization and no regularization.

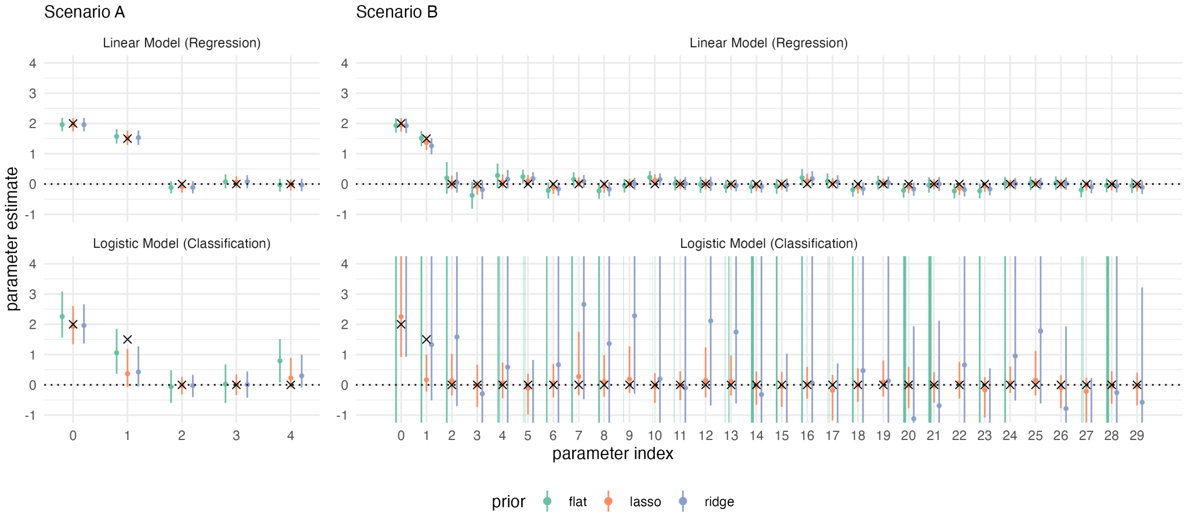

Figure 1: Estimated model parameters with 95% credibility interval (CI). The true parameters are notes as black crosses. Linear models produced more accurate estimates and smaller parameter CIs than logistic models. The necessity of regularization becomes apparent in scenario B, where the parameter estimates of the unregularized logistic model are very uncertain.

4.2 Performance of approximate inference algorithms in Bayesian regression

In a second example, we compared the performance of LA and Metropolis-Hastings algorithms in Bayesian linear and logistic models.

We generate 1000 synthetic data sets with $n = 100$:

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

We then fit one linear and one logistic model for each data set, using both LA and MCMC inference methods. We assumed $\boldsymbol{\theta} \sim \mathcal{N}(0, 10\boldsymbol{I})$ and residual variance fixed at $\sigma^2 = 10$. For LA, we used `r-INLA` with settings to default from INLA to simple LA. For MCMC, we used the R-package `MCMCglmm` with settings specifically to use Metropolis Hastings, a relatively small sample size of 5000, a burnin periord of 500 and a thinning interval of 10. For both algorithms, we recorded the CPU runtime and the posterior parameter estimates and standard deviation.

Results for the linear and logistic model can be seen in Figure 2 and Figure 3 respectively. In our experiment, LA was generally slower and thus more computationally expensive than MCMC methods, but this could also be mitigated by the specific implementation used in the experiment. In the case of the Bayesian linear model (Figure 2), there is not much difference in the posterior parameter estimates, although LA leads to much higher standard deviation of parameters. In Logistic regression (Figure 3) on the other hand, LA outperforms MCMC in both parameter estimation and standard deviation.

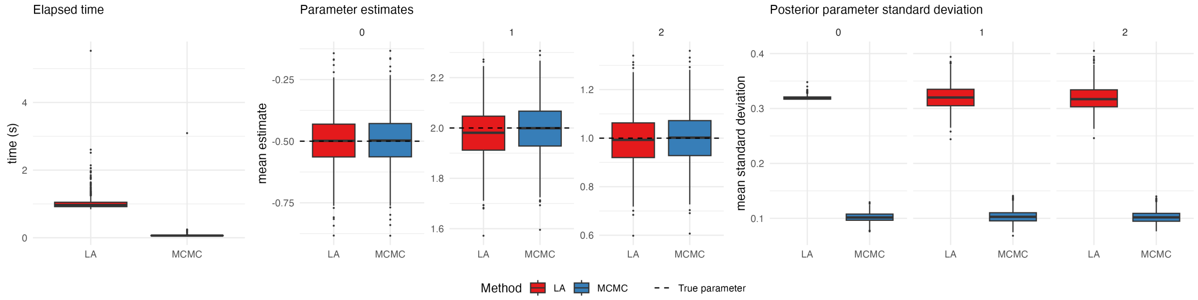

Figure 2: Linear regression: mean time, parameter estimates and standard deviation in LA and MCMC.

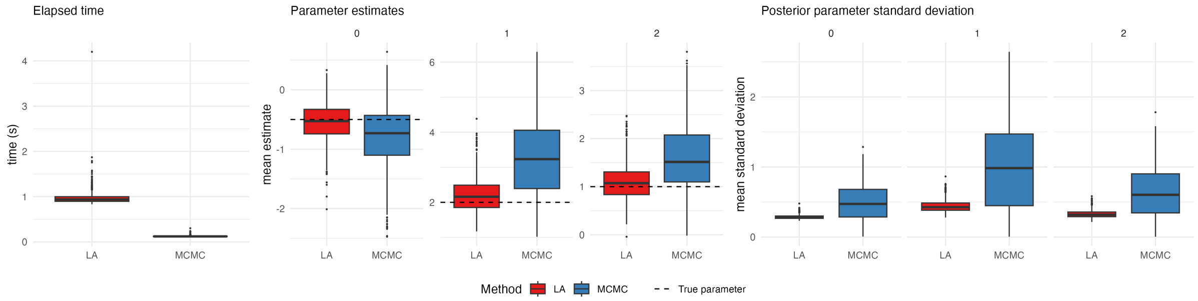

Figure 3: Logistic regression: mean time, parameter estimates and standard deviation in LA and MCMC.

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 $\boldsymbol{X}$. In general, we assume n observations and $p-1$ covariates (which means that the intercept θ_0 is always included in $\boldsymbol{\theta}$ and that $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{p-1} \end{pmatrix}$)

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

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

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

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

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, 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 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.

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

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., Hwang, J. and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. arXiv:1307.5928 [stat].

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].

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika* **57**(1): 97–109. Publisher: [Oxford University Press, Biometrika Trust].

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.

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

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.

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. 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.

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].

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.

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