FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of bachelor's thesis

Title:	Approximations in logistic regression
Student:	Eliáš El Frem
Supervisor:	Ing. Kamil Dedecius, Ph.D.
Study program:	Informatics
Branch / specialization:	Knowledge Engineering
Department:	Department of Applied Mathematics
Validity:	until the end of summer semester 2022/2023

Instructions

Logistic regression is a generalized linear model used for dichotomous random variables. Since the inference of its regression coefficients is not analytically tractable, it is based on numerical optimization methods. While the maximum likelihood-based estimators exploit the likelihood function maximized by the Newton's method, the Bayesian estimator relies on the Laplacian approximation of the posterior distribution. The aim of the thesis is to study the impact of the numerical approximations on the estimator convergence to the values of the true coefficients. A particular focus should be given to their sequential estimation from streaming data.

Particular goals:

1. Study and describe the theory of the generalized linear models, particularly the logistic regression.
2. Describe the logistic regression model and its inference - both the maximum likelihood-based and the Bayesian one.
3. Compare the impact of approximations.
4. Experimentally validate obtained results.

Electronically approved by Ing. Karel Klouda, Ph.D. on 17 February 2022 in Prague.

Bachelor's thesis

APPROXIMATIONS IN LOGISTIC REGRESSION

Eliáš El Frem

Faculty of Information Technology
Department of Applied Mathematics
Supervisor: Ing. Kamil Dedecius, Ph.D.
May 12, 2022

Czech Technical University in Prague
Faculty of Information Technology

Citation of this thesis: EL Frem Eliáš. *Approximations in logistic regression.* Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2022.

Contents

List of Figures

List of Tables

I would like to express my gratitude to my supervisor Ing. Kamil Dedecius, Ph.D., for his guidance, support, and patience during the writing of this thesis. I would also like to thank my family and friends for their support and encouragement.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis. I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular that the Czech Technical University in Prague has the right to conclude a license agreement on the utilization of this thesis as a school work under the provisions of Article 60 (1) of the Act.

In Prague on May 12, 2022 ....

Abstract

Logistic regression is a classification method used in machine data processing. This thesis deals with the optimization of the Bayesian estimate of the parameters of the logistic model using Laplace approximation. The state-of-the-art methods usually rely on suboptimal Laplace-type approximations. This thesis goes one step further. It investigates the impact of a more precise approximation on the estimation quality. The results obtained by computationally intensive optimization are compared with the traditional less intensive but also more imprecise.

Keywords logistic regression, Bayesian inference, binary classification, parameter estimation, machine learning, Laplace approximation

Abstrakt

Logistická regrese je klasifikační metoda používaná při strojovém zpracování dat. Tato práce se zabývá optimalizací Bayesovského odhadu parametrů logistického modelu pomocí Laplaceovské aproximace. Dosavadní metody většinou spoléhají na suboptimální Laplaceovské aproximace. Tato práce jde ještě o krok dále. Zkoumá dopad přesnější aproximace na kvalitu odhadu. Výsledky výpočetně náročných optimalizací jsou zde porovnány s tradičními optimalizacemi, jež jsou sice výpočetně jednodušší, ale také méně přesné.

Klíčová slova logistická regrese, Bayesovská inference, binární klasifikace, odhad parametrů, strojové učení, Laplaceovská aproximace

List of abbreviations

GLM	Generalized linear models
MLE	Maximum likelihood estimate
RMSE	Root mean squared error
BGR	Blue green red

Introduction

Nearly everything in this world can be a source of all kinds of data. With their help, we extend our knowledge in various fields of interest or improve existing processes in all areas of human activity. The field of science dedicated to processing data, understanding their structure, and predicting further behavior of their source is called machine learning and features various models designed for this purpose.

General linear models are a class unifying several statistical models that describe the connection between a variable and its regressors. This thesis deals with models used for classification, more specifically binary classification. Such methods are, for example, logistic regression and probit regression. There are other methods, but we will leave them aside for the purposes of this thesis.

The popularity of these models lies in their versatility. They are used in various fields such as medicine, economics, and risk management [1, 2, 3]. Logistic regression is a model that predicts the probability that the predicted variable belongs to a particular class (for example, if the value of the predicted variable is 1 in binary classification).

Parameters of the regression model can be estimated by either frequentist or Bayesian methods. Frequentist methods are based on processing the whole dataset at once and estimating the most likely parameters to generate this dataset. This allows us for very precise estimation on large datasets.

On the other hand, Bayesian estimation works by constantly updating existing information about the subject, which allows us to use prior knowledge on the matter. This also allows us to incorporate the data into the model online without processing the whole dataset again.

Goals

The main goal of this thesis is to examine the impact of approximations on the quality of Bayesian estimators in logistic regression. First, we will introduce the Generalised Linear Models, particularly the logistic regression model. Then we will describe the estimation methods - specifically the MLE estimator and the Bayesian estimator. The practical part of this thesis is dedicated to testing the impact of optimizations on a simulated dataset. After that, the same will be done for a real-world dataset, comparing the results of the optimized Bayes estimator with the one-step Bayes estimator and MLE estimator.

Chapter 1

State of Art

The methods described in this thesis are used for solving the problem of binary classification. We will deal primarily with supervised-learning classification methods, more precisely generalized linear models.

Among the other supervised-learning classification methods using the Bayesian methods belongs the naive Bayes classifier, which is often used in spam filters [4], or Bayesian Networks [5].

One of the most popular classification methods is the decision trees [6]. Their main advantage is their comprehensibility to human observers. They are often used in expert systems, where we understand the data we want to classify. Due to their versatility, they can be used in various fields, such as agriculture, text sentiment classification, and medicine [7, 8, 9].

Another classification method is support vector machines [10]. In this method, the data points are viewed as n-dimensional vectors. We then find the hyperplane, where the margin between the points belonging to a different class is most significant. The new points will then be classified by the side of the gap they belong to. Support vector machines are being used in text categorization [11].

Another popular classification method is k-nearest neighbors (k-NN). Training data are represented as vectors in n-dimensional space. New data samples are then added to the space and assigned class corresponding to the class majority of their nearest neighbors - vectors closest to the one we added according to the selected metric. Further reference in [12]. The advantage of this approach is that we do not need to teach the model - instead, what is computationally intensive is the classification itself. These properties make k-nn useful for image segmentation [13, 14].

Neural networks can also do classification. Neural networks are (as the name suggests) networks composed of one or more layers of neurons inspired by neurons in the human brain. Each neuron has its activation function whose input is either input of the neural network (for the first layer) or the output of one or more neurons from the previous layer (for further layers). They are used for a large variety of tasks in various fields such as medicine, image classification, or text classification [15, 16, 17].

Methods that combine several classifiers are called ensemble methods. The most popular members of this category are bagging (bootstrap aggregating) and boosting. Bagging is based on separating data into several chunks that we call bootstrap datasets. Each bootstrap dataset is then processed with a different model. The final result is then the result given by most of the models. We sequentially process the whole dataset with each model when using boosting methods. After each fitting, we identify the previously misclassified instances and elevate their priority for the next model. Boosting gives us more precise results in some cases, but bagging, on the contrary, is less prone to overfitting.

As said in the introduction, the scope of this thesis is the GLMs. Their class was popularized by Nelder [18] and features a wide variety of models for both regression and classification. In this thesis, we will deal particularly with the logistic model. Logistic model invention is according to Cramer [19] attributed to Verhulst, who in his work *La loi d'accroissement de la population* first used the term "logistic" [20]. It uses the logit link function to deliver the probabilities of data samples belonging to a particular class.

Another popular model for classification from this class is probit model, whose invention is according to [19] often being attributed to Bliss [21] and Gaddum [22]. It works on similar principles as the logistic regression, but instead of using the logit link function defined as:

$$logit(x) = \log\left(\frac{1}{1-x}\right), \tag{1.1}$$

it uses the probit link function which is defined as:

$$probit(x) = \Phi^{-1}(x), \tag{1.2}$$

where Φ denotes the cumulative distribution function of standard normal distribution. It has uses in economics and risk management [2, 3].

This thesis is focused on the estimation of the parameters of the logistic models. Generally, there is two approaches - the Bayesian and the frequentist inference.

Frequentist interference is, in most cases, represented by MLE estimation. MLE is based on the likelihood function, which tells us, how probable are the processed data for our estimated parameters in the selected model. Maximizing this function gives us the maximal likelihood estimate.

While the MLE technique gives us parameters that are most likely to generate the dataset we used it on, the Bayesian approach returns their probability distribution. This distribution is often not analytically tractable. Thus we have to use approximations.

There are many ways to approximate the posterior distribution, such as Monte Carlo methods based on numerical sampling or Laplace approximation. From this class, the algorithms are, for example, Markov Chain Monte Carlo, first introduced in 1949 in [23], sampling importance resampling algorithm, introduced by Rubin in [24] and the weighted likelihood bootstrap introduced by Newton and Raftery in [25]. The weighted likelihood bootstrap is an extension of the Bayesian bootstrap algorithm introduced by Rubin in [24].[26] The Markov Chain Monte Carlo methods include Gibbs sampling, introduced in [27] by Geman and Geman, which can be used in both logit and probit regression [28].[29]

Lewis in [26] also reviews methods for estimating the predictive distribution needed for making predictions of the unknown value. The method Lewis states are: the harmonic mean of the output likelihoods and modifications thereof [25], bridge sampling [30] and path sampling [31].

Another way to approximate the analytically intractable distributions is the variational Bayesian methods, which can be used as the alternative to Markov Chain Monte Carlo methods. For further reference see [32]. This thesis deals with optimizations in Laplace approximation. Laplace approximation works by approximating unknown distribution with a normal distribution centered in the global maximum of the approximated distribution.

Finding the maximum of a function can be done in multiple ways. In this thesis, we will use two methods. The first one is Newton-Raphson iterative algorithm described in [33, 29] and the second one is the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm described in [33] used by scipy.optimize.minimize [34].

This subject is loosely related to the research of MacKay, who studied the improvement of quality Bayesian prediction by changing the basis of Laplace approximation in [35]. Our research aims to compare the quality of the parameter estimation based on improving the quality of Laplace approximation by using the real global maximum of the approximated distribution.

Chapter 2

Generalized linear models

In this chapter, generalized linear models will be introduced. Using their definition, we will define the logistic model and then look into the two ways we can make estimates for this model - MLE and Bayes inference.

2.1 GLMs - definition and basic description

First of all, the author will introduce a proper definition of GLMs. The generalized linear models consist of three components[36]:

- Random component.
- Linear predictor
- Link function

The author will now describe and define each component of a GLM.

Let us suppose we are observing a stochastic process. We are interested in a specific quantity that is the outcome of the process. The vector containing these observations is called the random component of a GLM.

▶ **Definition 2.1.1** (Random component of a GLM)**.** *Let $y = (y_1, \ldots, y_n)$ be a vector of n observations of a stochastic process with probability distribution from the exponential family. We then call y the random component of an associated GLM.*

Together with the random component, we have acquired the data we suspect are linked to the observed variable we try to explain. For example temperature of a particular object during the day would be connected to its size, the thermal conductivity of the material it is made from, etc. These explanatory variables are called regressors. Each regressor has its coeficients. The vector of weighted regressors is called the linear predictor of a GLM.

▶ **Definition 2.1.2** (Linear predictor of a GLM)**.** *Let $y_t \in y$ be an observation from the random component of a GLM y. Let $x_t \in R^n$ be its observable regressor and $\theta \in R^n$ a vector of unknown regression coefficients. We then call the product $\theta^T x_t$ a linear predictor of a GLM.*

We then try to link the observed random component to the linear predictors using what is called a link function. The model that links these data together is called the Generalized linear model.

▶ **Definition 2.1.3** (Generalized linear model)**.** *Let $y_t \in y$ be an observation from the random component of a GLM. Let $\theta^T x_t$ be a linear predictor. The generalized linear model then has following form:*

$$\hat{y}_t = \mathbb{E}[y_t | x_t, \theta] = g^{-1}(\theta^T x_t) \tag{2.1}$$

where $\hat{y}_t = \mathbb{E}[y_t | x_t, \theta]$ is called expected value of y_t and $g(\cdot)$ is a link function of the GLM.

From Definition 2.1.3, we can see that

$$\theta^T x_t = g(\mathbb{E}[y_t | x_t, \theta]) = g(\hat{y}_t). \tag{2.2}$$

This is called mean function.

In the following subsection, we will go through a few examples of the link function and present the usage of the models that use them.

2.1.1 Identity link function

Identity link function has the following formula:

$$x_t \theta^T = \hat{y}_t. \tag{2.3}$$

The mean function of the identity link function is

$$\hat{y}_t = x_t \theta^T. \tag{2.4}$$

The identity link function is used when the random component has a normal distribution. The GLM model using this function is called the Linear regression model and is mainly used for regression due to the range of link function ranging $(-\infty, \infty)$.

2.1.2 Negative inverse link fuction

The negative inverse function has the formula:

$$x_t \theta^T = -(\hat{y}_t)^{-1}. \tag{2.5}$$

Its mean function has the formula:

$$\hat{y}_t = -(x_t \theta^T)^{-1}. \tag{2.6}$$

The negative inverse link function is typically used for random components with exponential or gamma distribution. Therefore the range of the link function is in $(0, \infty)$.

2.1.3 Logit link function

The Logit link function can be used with multiple distributions. For the binominal logistic model, the distribution used is the Bernoulli distribution.

The link function for the logistic model has the formula:

$$x_t\theta^T = \log\left(\frac{\hat{y}_t}{1-\hat{y}_t}\right). \tag{2.7}$$

This is the logit function which will be properly defined in the next section, along with the The mean function has the formula:

$$\hat{y}_t = \frac{1}{1+e^{-x_t\theta^T}}. \tag{2.8}$$

The range of the link function is $(0, 1)$ - the probability of the $\hat{y}_t$ being 0.

When the random component is not dichotomic, we call the model a multinomial logistic model. For multinomial logistic regression, the distribution used is the Binominal distribution.

The link function then has the formula:

$$x_t\theta^T = \log\left(\frac{\hat{y}_t}{n-\hat{y}_t}\right). \tag{2.9}$$

The mean function has the same formula as the one for the logistic regression model. Here the output of the model returns the probability of the data sample belonging to the nth class.

2.1.4 Log link function

Log link function is used for Poisson regression and is used for the random component having Poisson distribution. The negative inverse function has the formula:

$$x_t\theta^T = \log(\hat{y}_t). \tag{2.10}$$

Its mean function has the formula:

$$\hat{y}_t = e^{x_t\theta^T}. \tag{2.11}$$

The output of the model is used, for example, to enumerate the rate of occurrences of individual events.

2.2 Logistic regression model

Logistic regression model is used for binary classification, where the observed variable y_t is dichotomous:

$$y_t = \begin{cases} 1 & \text{with probability } p_t, \\ 0 & \text{with probability } 1-p_t. \end{cases} \tag{2.12}$$

Therefore it can be described by Bernoulli distribution with p_t being its parameter:

$$y_t = Bernoulli(p_t) \tag{2.13}$$

with $\mathbb{E}[y_t] = p_t$. Instead of predicting the value of variable y_t, we instead aim to predict the probability:

$$p_t = p(x_t, \theta) = \Pr(y_t = 1 | x_t, \theta). \tag{2.14}$$

The logistic regression model uses logit function as a link function.

▶ **Definition 2.2.1** (Logit function)**.** *Let $x \in R$. Then we define the logit function as:*

$$logit(x) = \log\left(\frac{1}{1-x}\right). \tag{2.15}$$

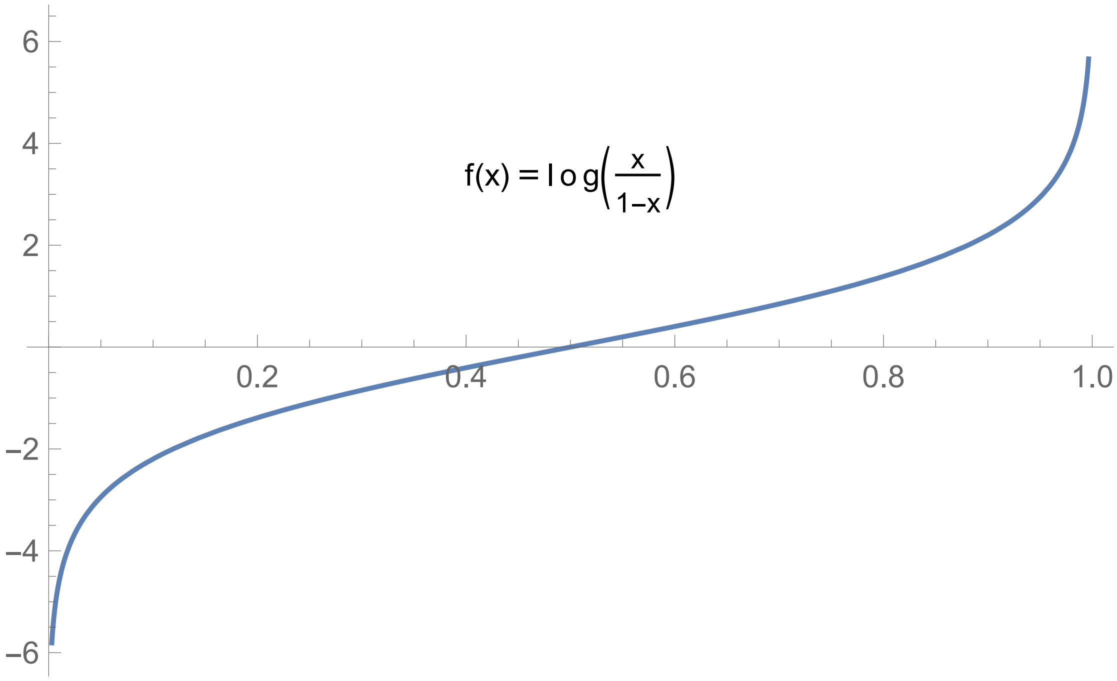


Figure 2.1 Logit function

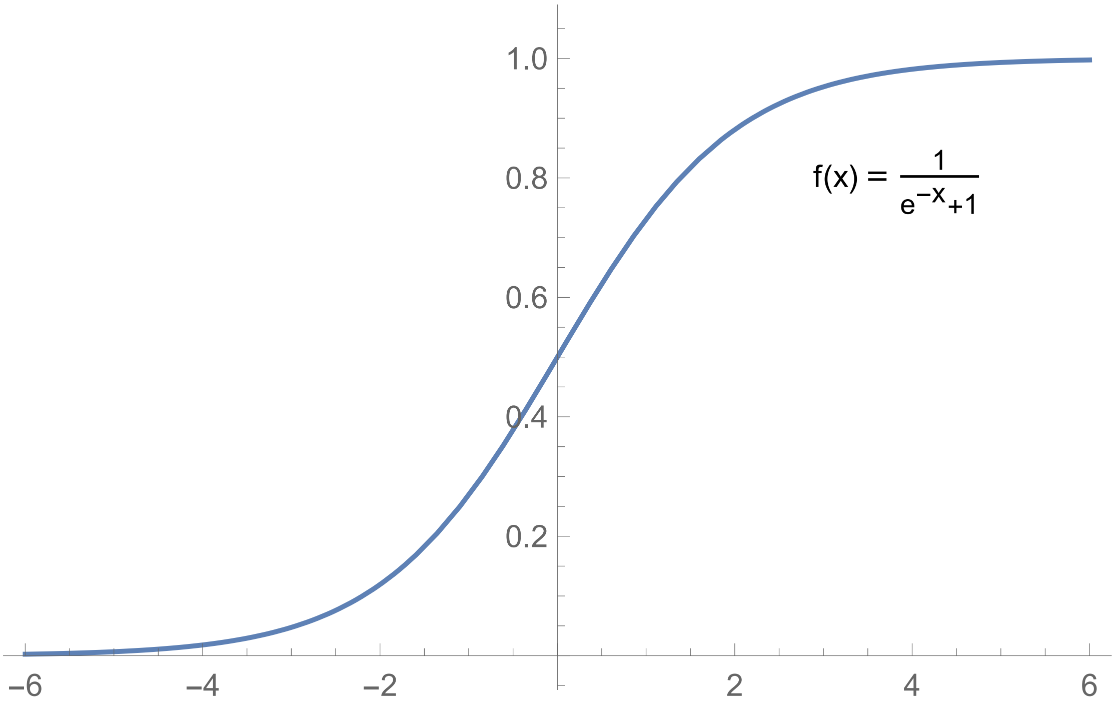
The inversion of the logit function is called the logistic sigmoid with the formula

$$logit^{-1}(x) = \sigma(x) = \frac{1}{1+e^{-x}}. \tag{2.16}$$

Using 2.1.3 we can then describe the logistic regression model as:

$$\hat{y}_t = p_t = p(x_t, \theta) = logit^{-1}(\theta^T x_t) = \sigma(\theta^T x_t) = \frac{1}{1+e^{-\theta^T x_t}}. \tag{2.17}$$

The logistic sigmoid is widely used in statistics and machine learning methods. Besides the logistic regression, it can often be found in neural networks as their activation function. Using the logistic sigmoid, we can assign a probability to variables ranging over $(-\infty, \infty)$.



■ **Figure 2.2** Logistic sigmoid function

2.3 MLE estimators

We search for the parameters that make our model most probable. In the logistic regression model, we want to find θ that would maximize the probability of the observed data. For the dataset with random component y and the matrix of its regressors X with coefficient vector θ, the likelihood $L(\theta, X, y)$ for the logistic regression model is measured as:

$$L(\theta, X, y) = \prod_{i=1} p(x_i, \theta)^{y_i} (1 - p(x_i, \theta))^{1-y_i} \tag{2.18}$$

$$= \prod_{i=1} \left(\frac{1}{1 + e^{-\theta^T x_i}} \right)^{y_i} \left(1 - \frac{1}{1 + e^{-\theta^T x_i}} \right)^{1-y_i}. \tag{2.19}$$

To find the parameters with maximal likelihood, we have to find the maximum of this function. To maximize this function, we set its derivative equal to zero. Because the derivative of 2.18 is computationally intensive, we can instead choose to derivate its logaritm - the log-likelihood function:

$$l(\theta, X, y) = \log \prod_{i=1} p(x_i, \theta)^{y_i}(1 - p(x_i, \theta))^{1-y_i} \tag{2.20}$$

$$= \sum_{i=1} \left(y_i \log \left(\frac{1}{1 + e^{-\theta^T x_i}} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + e^{-\theta^T x_i}} \right) \right) \tag{2.21}$$

$$= \sum_{i=1} \left(y_i \theta^T x_i - \log(1 + e^{\theta^T x_i}) \right). \tag{2.22}$$

Then the maximal likelihood estimate for the parameter θ is the solution of the following equation:

$$\frac{\partial l(\theta)}{\partial \theta} = \sum_{i=1} x_i(y_i - p(x_i, \theta)) = 0. \tag{2.23}$$

The solution of this equation can be computed using numerical optimization.

2.4 Laplace approximation

Before describing the Bayesian inference, we will go through the principles of the Laplace approximation. Note that the approach in this section is inspired by [29] which can be used for further reference.

Let us suppose a random variable z with its distribution $p(z)$ defined as:

$$p(z) = \frac{1}{Z} f(z), \tag{2.24}$$

where

$$Z = \int f(z) dz \tag{2.25}$$

is an unknown normalization coefficient.

Our goal is to approximate the distribution $p(z)$ by placing the normal distribution with the center in the mode of $p(z)$. The mode of $p(z)$ is the point where its first derivation is equal to 0.

After finding the modus, which can be done by the Newton-Raphson algorithm presented later in this thesis, we have yet to identify the variance of the normal distribution. For this, we can utilize the fact that the logarithm of normal distribution is a quadratic function of the variables. Using this, we can consider a Taylor expansion of $\log f(z)$ centered on the mode z_0 so that

$$\log f(z) \simeq \log f(z_0) - \frac{1}{2} A(z - z_0) \tag{2.26}$$

where

$$A = - \left(\frac{\partial^2 \log f(z)}{\partial z^2} \right). \tag{2.27}$$

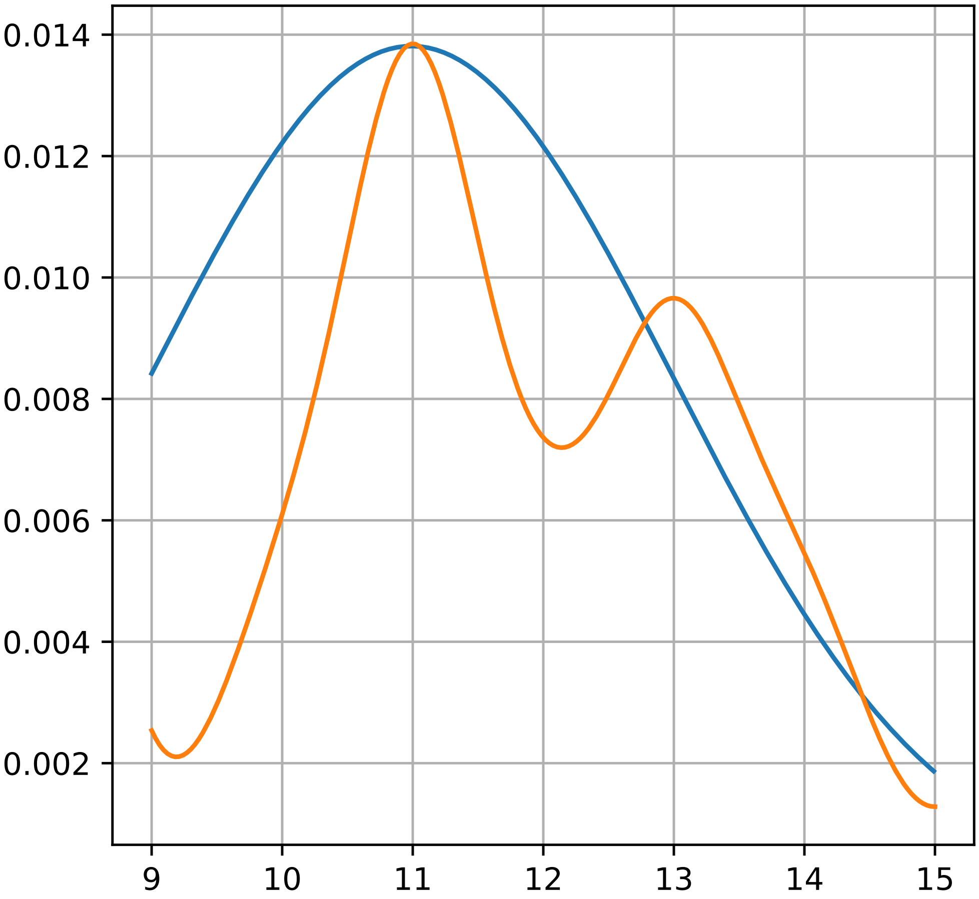
From that, we can obtain normalized distribution $q(z)$ using standard result of normalization of a normal distribution

$$q(z) = \left(\frac{A}{2\pi} \right)^{\frac{1}{2}} e^{-\frac{A}{2}(z - z_0)}. \tag{2.28}$$

From this follows:

$$q(z) = \left(\frac{A}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{A}{2}(z-z_0)} = \mathcal{N}(z|z_0, A^{-1}). \tag{2.29}$$

Illustration image of how the approximation could look like can be seen in Figure 2.3.



■ **Figure 2.3** Example of the Laplace approximation. The distribution (orange) is approximated by normal distribution (blue)

2.5 Bayesian estimators

The Bayesian approach to statistical inference is a more natural process for humans. The idea behind it is to periodically update previous knowledge, similar to human learning. In this section, we will go through the basics of Bayesian inference, and then we will focus specifically on its application in logistic regression.

2.5.1 Basic principles of Bayesian inference

Haugh in [37] sums up four steps for Bayesian inference:

1. begin with some prior belief statement,
2. use the prior belief and a dynamic model to make a prediction,
3. update the prediction using a set of observations and an observation model to obtain a posterior belief, and
4. declare the posterior belief our new prior belief and return to 2.

We will now go through all the steps and further describe each point.

Let y be a random component as defined in 2.1.1.Let us remind that y is dichotomous random variable with the Bernoulli distribution with probability p. The prior belief in our case is our information about the parameter p. Let us consider the following four examples of the prior distribution of parameter p:

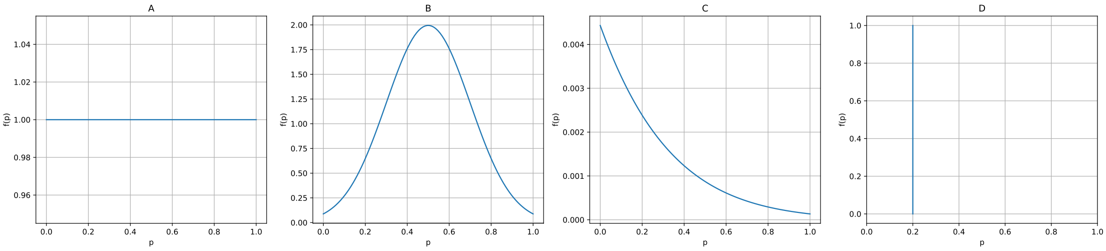


Figure 2.4 Examples of prior distributions

In graph A, the prior distribution tells us nothing about the parameter p. All values in the range $(0, 1)$ are equally probable. Graph B shows the case where we suppose that the value of p is more likely to be around 0.5, and the values on both sides are equally probable. Case C presents the situation where we suspect that the value of p is closer to zero and much less likely to have any value close to one. Graph D is a special case of a prior distribution. Here the value is fixed at 0.2, and no amount of data can change this belief. This distribution is called the Dirac Delta distribution.

For random component y, regressor X and regression coeficients θ, the Bayes theorem gives us following formula for posterior distribution:

$$\pi(\theta|x_{0:t}, y_{0:t}) = \frac{f(y_t|x_t, \theta) \cdot \pi(\theta|x_{0:t-1}, y_{0:t-1})}{f(y_t|x_{0:t}, y_{0:t})}, \tag{2.30}$$

where:

- $f(y_t|x_t, \theta)$ is the data model,
- $\pi(\theta|x_{0:t-1}, y_{0:t-1})$ is the prior distribution,
- $f(y_t|x_{0:t}, y_{0:t})$ is normalizing term.

Note that t is the number of actual steps and notation $\pi(\theta|x_{0:t-1}, y_{0:t-1})$ means that we are using posterior distribution for the data samples $x_0, x_1, \ldots, x_t$ and their respective random component members.

We want to avoid rewriting the normalizing term in each equation adjustment, so instead of equality (=), we use the proportionality sign ($\propto$), which denotes equality up to the normalizing constant. Therefore we can write:

$$\begin{aligned}\pi(\theta|x_{0:t}, y_{0:t}) =& \frac{f(y_t|x_t, \theta) \cdot \pi(\theta|x_{0:t-1}, y_{0:t-1})}{f(y_t|x_{0:t}, y_{0:t})} && (2.31)\\ \propto& f(y_t|x_t, \theta) \cdot \pi(\theta|x_{0:t-1}, y_{0:t-1}). && (2.32)\end{aligned}$$

In the next section, we will apply this estimate to the logistic model.

2.5.2 Bayesian estimation in logistic regression

Now let's look into how the Bayesian estimation works with the logistic model. In Equation 2.17 we have defined the logistic model as:

$$\hat{y} = logit^{-1}(\theta^T x_t). \tag{2.33}$$

We try to predict the distribution of parameter θ based on the prior knowledge and the data likelihood. The probability function for the Bernoulli distribution that the random component has is

$$f(y_t|\theta, x_t) = p^{y_t}(1-p)^{1-y_t}. \tag{2.34}$$

This is the likelihood of the data.

Prior distribution in this case is the posterior distribution from the previous step

$$\pi(\theta|x_{0:t-1}, y_{0:t-1}). \tag{2.35}$$

The new posterior distribution will then be computed as:

$$\begin{aligned}\pi(\theta|x_{0:t}, y_{0:t}) &\propto f(y_t|\theta, x_t) \cdot \pi(\theta|x_{0:t-1}, y_{0:t-1}) && (2.36)\\ &\propto p^{y_t}(1-p)^{1-y_t} \cdot \pi(\theta|x_{0:t-1}, y_{0:t-1}). && (2.37)\end{aligned}$$

The problem with this update is that the posterior distribution is not analytically tractable. We have to estimate the value. In this thesis, it is done by the Laplace approximation.

The Laplace approximation is used in both the estimation of the θ parameters and the final prediction.

The simulation of the posterior distribution is done by a multivariate normal distribution centered on the maximum of the posterior distribution. This means that the mean is our estimate of the θ parameter (denoted $\hat{\theta}$). The variance of the normal distribution, Σ is computed as the inverse of the second derivative of log-likelihood of θ estimate from the previous step ($\hat{\theta}_{t-1}$) with the formula:

$$\Sigma_t = -\left(\frac{\partial^2 l(\hat{\theta}_{t-1})}{\partial\hat{\theta}_{t-1}\partial\hat{\theta}_{t-1}}\right)^{-1}. \tag{2.38}$$

The prior distribution is then approximated by the multivariate normal distribution described above as follows:

$$\pi(\theta|x_{0:t-1}, y_{0:t-1}) \sim \mathcal{N}(\hat{\theta}_{t-1}, \Sigma_{t-1}) \tag{2.39}$$

$$= \frac{|\Sigma_{t-1}^{-1}|^{\frac{1}{2}}}{(2\pi)} e^{-\frac{1}{2}(\theta-\hat{\theta}_{t-1})^T \Sigma_{t-1}^{-1}(\theta-\hat{\theta}_{t-1})}. \tag{2.40}$$

When combining the above equation with the equation 2.36, we get following formula for the posterior distribution:

$$\pi(\theta|x_{0:t}, y_{0:t}) \propto f(y_t|\theta, x_t) \cdot \pi(\theta|x_{0:t-1}, y_{0:t-1}) \tag{2.41}$$

$$\sim f(y_t|\theta, x_t) \cdot \mathcal{N}(\hat{\theta}_{t-1}, \Sigma_{t-1}) \tag{2.42}$$

$$\propto p^{y_t}(1-p)^{1-y_t} \cdot \mathcal{N}(\hat{\theta}_{t-1}, \Sigma_{t-1}) \tag{2.43}$$

$$\propto p^{y_t}(1-p)^{1-y_t} \cdot \frac{|\Sigma_{t-1}^{-1}|^{\frac{1}{2}}}{(2\pi)} e^{-\frac{1}{2}(\theta-\hat{\theta}_{t-1})^T \Sigma_{t-1}^{-1}(\theta-\hat{\theta}_{t-1})}. \tag{2.44}$$

New $\hat{\theta}$ is then found as MAP of $\pi(\theta|x_{0:t}, y_{0:t})$, which is found by numerical optimization methods.

2.5.3 Prediction in Bayesian logistic regression

The Bayesian prediction of the members of the random component is made by the predictive distribution. Unfortunately, like the posterior distribution, the predictive distribution is not analytically tractable. Therefore we will again use the Laplace approximation for the predictive distribution. The predictive distribution for the random component y' with the regressor x' takes the form [26]:

$$f(y'|x', x_{0:t}, y_{0:t}) = \int f(y'|x', x_{0:t}, y_{0:t}, \theta) \cdot \pi(\theta|x_{0:t}, y_{0:t}) d\theta \tag{2.45}$$

$$\approx 2 \cdot \pi \cdot \left(\det \frac{\partial^2 l(\hat{\theta}_t)}{\partial\hat{\theta}_t\partial\hat{\theta}_t}\right) \cdot f(y'|x', \theta) \cdot \pi(\theta|x_{0:t}, y_{0:t}), \tag{2.46}$$

where:

- $\pi(\theta|x_{0:t}, y_{0:t})$ is the posterior distribution from 2.36,
- $f(y'|x', \theta)$ is the likelihood of a random component,
- θ we be approximated by $\hat{\theta}_t$.

After that, we get y' which is the probability of $\hat{y}_t$ being 1. Therefore after receiving the results of predictive distribution, the prediction of $\hat{y}_t$ will be assigned as follows:

$$\hat{y}_t = \begin{cases} 1 & \text{if } y' >= m, \\ 0 & \text{if } y' < m, \end{cases} \tag{2.47}$$

where m is our chosen threshold which usually has the value of 0.5.

2.5.4 Newton-Raphson algorithm

In both MLE and Bayesian estimators, we need to maximize functions; the likelihood function in the MLE estimator and the posterior distribution in Bayesian estimation. This can be made using numerical approximation methods. The method used for experiments in this thesis was one step of the Newton-Raphson algorithm. For a description of the algorithm, we first have to define the Hessian matrix.

▶ **Definition 2.5.1** (Hessian matrix)**.** *Let β be a vector of n parameters $(\beta_1, \beta_2, \dots, \beta_n)$. Let $E(\beta)$ be a function of β. The matrix H has its members defined as*

$$H_{i,j} = \frac{\partial^2 E(\beta)}{\partial \beta_i \partial \beta_j}. \tag{2.48}$$

We then call H the Hessian matrix for $E(\beta)$.

The Newton-Raphson update for minimizing a function $E(\beta)$ than takes the form [29]

$$\beta^{new} = \beta^{old} - H^{-1} \frac{\partial E(\beta)}{\partial \beta}, \tag{2.49}$$

where H is the Hessian matrix for $E(\beta)$.

The details of the Newton-Raphson are beyond the scope of this thesis. They can be found, e.g., in [29].

Chapter 3

Experiments

The state-of-the-art algorithms that approximate the posterior distribution using the Laplace method described earlier usually perform only one step of the Newton-Raphson optimization. They rely on the fact that with an increasing number of measurements, the posterior sufficiently concentrate at a tiny region, and the optimization thus reaches some point very close to the maximum. Alternatively, the methods are stuck in a local minimum, as shown in 3.1.

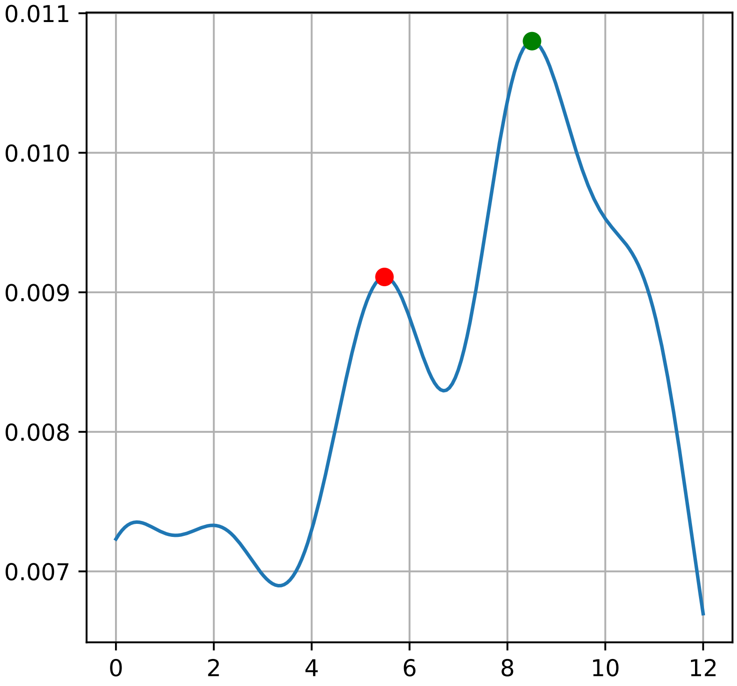


Figure 3.1 The algorithm can get stuck in local maximum (red) instead of the global maximum (green)

This can negatively affect the quality of the parameter estimation. The experimental part of this thesis examines this hypothesis first on simulated data, then on a real-life dataset.

3.1 Used metrics

To see the impact of chosen approximation and inference method, we compare the estimators by whether they visually converge to the real value (or the maximal likelihood estimate for the whole dataset in the case of real-life data) and three metrics:

- RMSE,
- Brier score,
- Confusion matrix.

First, the metrics will be briefly described here; then, we will go through the results of our experiments.

3.1.1 RMSE

RMSE is a metric used to evaluate the quality of an estimator based on the deviation of its parameters from real parameters. The formula for computing the RMSE is

$$RMSE = \sqrt{\sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2}, \tag{3.1}$$

where $\hat{\theta}$ is the estimate of parameter vector and n number of values in the parameter vector.

In the experimental part of this thesis, the RMSE will be recounted sequentially after incorporating each data point into our model and updating the estimate.

3.1.2 Brier score

While RMSE compares the estimators by the parameters they give us for the model, the Brier score compares the model estimates of the random component. For N predictions $\hat{y}$ of the random component y, we define the Brier score as

$$B = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2. \tag{3.2}$$

In this thesis, the Brier score will be computed only once, for the model that uses the final estimate of parameters.

3.1.3 Confusion matrix

The confusion matrix shows us the results of the prediction (the same one we made to compute the Brier score) in the form described in Table 3.1.

	Pred. true	**Pred. false**	**Sum**
True	True positive	False negative	Sum true
False	False positive	True negative	Sum false
Sum	Sum pred. true	Sum pred. false	Total sum

Table 3.1 Confusion metrix have following form in this thesis

Thus, we can see how many data samples were misclassified and whether the model tends to return false positives or negatives.

These metrics will help us fully understand the models we create and explain their behavior.

3.1.4 Used models

For our experiments, we will compare three estimators, which we have labeled:

- MLE estimator,
- One-step Bayes estimator,
- Optimized Bayes estimator.

The MLE estimator uses the MLE algorithm to determine the most probable parameters for a given dataset. For it is impossible to work with small amounts of data, its plot may start slightly later than those of the other two estimators.

One-step Bayes estimator will use Laplace approximation with Newton-Raphson algorithm from Section 2.5.4 for estimation of the MAP estimate of parameter θ. As the name suggests, only one iteration of the Newton-Raphson algorithm will be made.

Optimized Bayes estimator will instead use scipy.optimize.minimize function from the SciPy package [34], which uses BFGS algorithm mentioned in Chapter 1. This function, as the SciPy documentation tells us, is used to find the local minimum of a given function. It is not guaranteed to find the global minimum of a function; however, during our tests, it gave the same results as other methods in the SciPy package that claimed to find the global minimum of a function in a fraction of the time used by these algorithms.

3.2 Simulated datasets experiments

For the purposes of this thesis, we used following set of parameters for generating the linear predictors of our model:

regressor	lowest val	highest val
intercept	1	1
x_1	18	60
x_2	-20	20
x_3	-4	30

As for the θ coeficients, we have used two sets described in Table 3.2. Then, using these linear predictors and (2.4), we have generated the random component y. For the first set of θ coefficients, the values in the random component were distributed to classes in a ratio close to

2:1, whether for the second one, the ratio was closer to 1:1. The number of generated samples was 5,000 and 10,000.

First set of coeficients		Second set of coeficients	
θ_0	-1.6	θ_0	-1.6
θ_1	0.03	θ_1	0.5
θ_2	0.04	θ_2	-1
θ_3	-0.02	θ_3	-1.2

Table 3.2 Two sets of coeficients used in our experiments

The results can be described in the following three cases:

- All estimators' parameters converge to the real values.
- One-step Bayes estimator fails to converge/converges considerably slower than MLE/optimized Bayes estimator.
- All estimators fail to converge in 10,000 steps.

3.2.1 Simulated data case 1: All estimators converge to real values

In this case, all estimators converge successfully to the real values of θ with their RMSE being generally lower than 0.1, which shows a perfect fit to the dataset. With the first set of θ coefficients, this was the case most of the time. In the case presented in Figure 3.2 the final RMSE was lower than 0.01 for the optimized Bayes estimator (as can be seen in Table 3.3). For the second set of θ coefficients, the convergence was slightly worse for both Bayes estimators and sometimes did not happen at all (this will be further discussed in the next subsection). The RMSE was generally a bit higher for this set of coefficients.

In Figures 3.2 and 3.3, we can see that in the beginning, the estimates all deviated significantly from the real coefficients due to the low amount of data samples. After this, the convergence is more or less the same for all models. Final RMSE, as well as the Brier score, are depicted in Table 3.3.

Method	Brier score	RMSE
MLE	0.208971	0.018007
one-step Bayes	0.208969	0.007455
Optimized Bayes	0.208970	0.016361

Table 3.3 Simulated data case 1: Final Brier score and RMSE

We can see that the RMSE is great for all models, with the one using the optimized Bayes estimator slightly lower. This confirms that all estimators give us parameters that converge to the θ vector.

In cases similar to this one, the optimized Bayes estimator converged faster than the one-step Bayes estimator for smaller amounts of data. After receiving enough data samples, they usually merge.

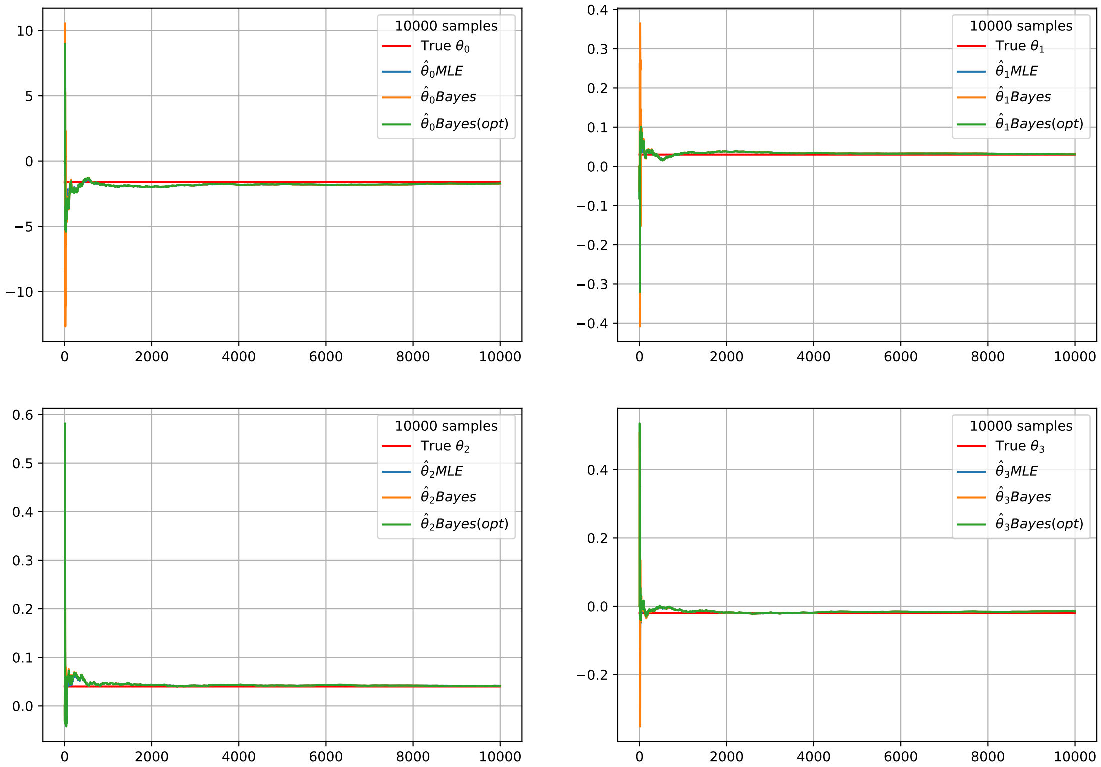


Figure 3.2 Simulated data case 1: All values converge to real values of θ

Last property of the models we will look at are confusion matrices in tables 3.4, 3.5 and 3.6. As the Brier score from Table 3.3 suggests, the number of misclassified samples is similar for all estimators. What may be surprising is that the number is rather big. From the matrices, we can see that 30% of the samples were misclassified. For the less represented class members, this was the case for more than 60% samples. To further examine this phenomenon, we took real values of θ and made the prediction using them - the Brier score was still near the value of 0.2. Thus while making further assumptions about the quality of the estimations, we can consider values around 0.2 to be the best achievable results for the dataset generated using the first θ parameter vector.

	pred 0	pred 1	sum
0	5660	820	6480
1	2500	1020	3520
sum	8160	1840	10000

Table 3.4 Simulated data case 1: Confusion matrix for MLE

	pred 0	pred 1	sum
0	5670	810	6480
1	2511	1009	3520
sum	8181	1819	10000

Table 3.5 Simulated data case 1: Confusion matrix for one-step Bayes

	pred 0	pred 1	sum
0	5661	819	6480
1	2506	1014	3520
sum	8167	1833	10000

Table 3.6 Simulated data case 1: Confusion matrix for optimized Bayes

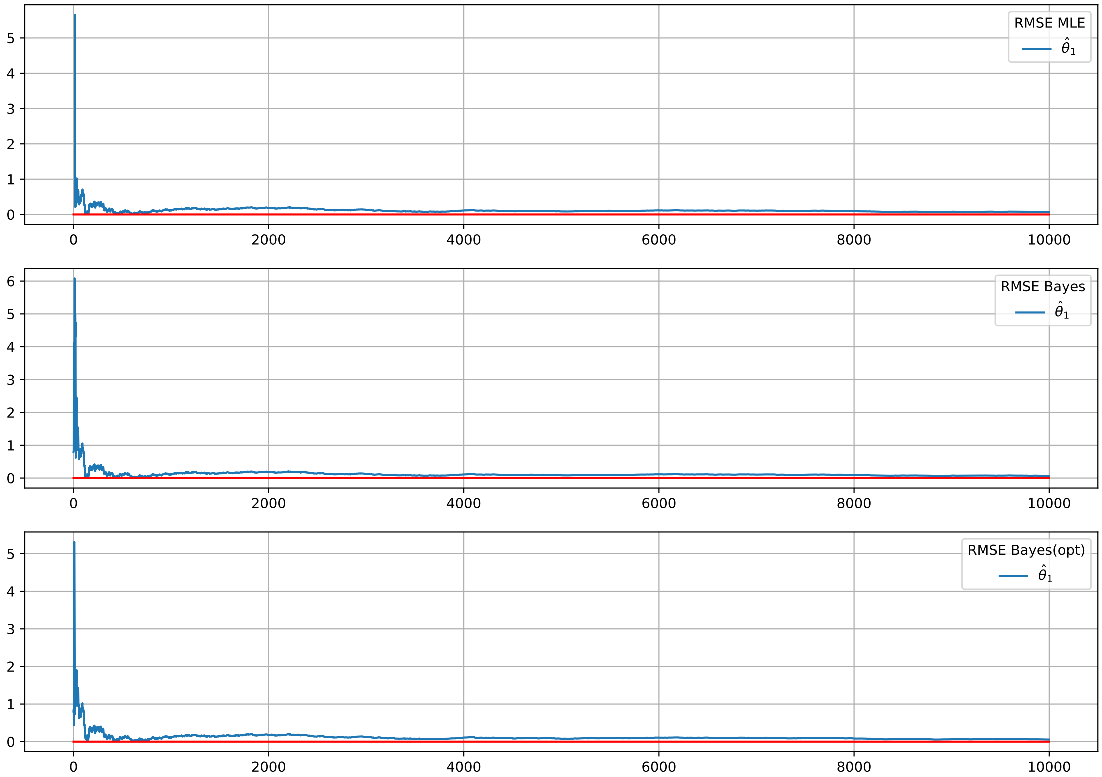


Figure 3.3 Simulated data case 1: RMSE showing perfect convergence

3.2.2 Simulated data case 2: Both Bayes estimators fail to converge in 10000 steps

In this example, the Bayes estimators have not converged to the real θ. This has happened almost exclusively for the data generated using second set of θ parameters where the random component had its values distributed more evenly. In Figure 3.4 both Bayes estimators show slowing convergence as opposed to MLE estimator that converges almost perfectly.

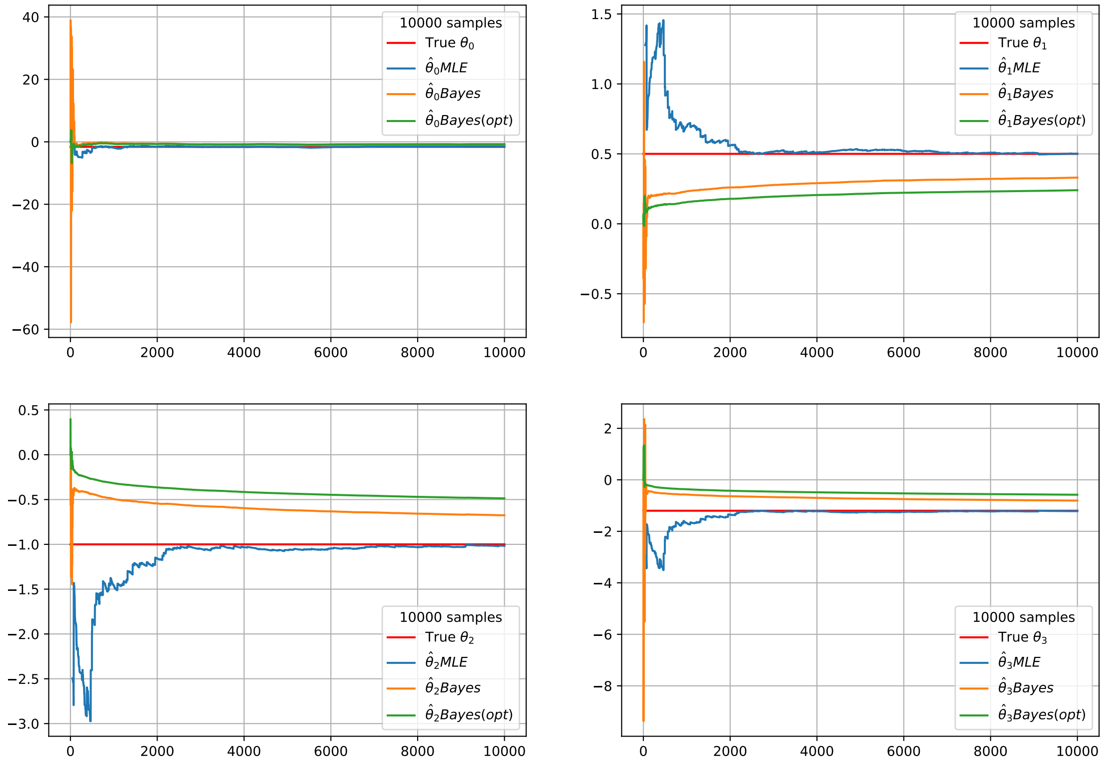


Figure 3.4 Simulated data case 2: None of the estimators converges in 10000 steps

Method	Brier score	RMSE
MLE	0.0206762	0.117725
One-step Bayes	0.02418352	0.598492
Optimized Bayes	0.02164621	0.460211

Table 3.7 Simulated data case 2: Final Brier score and RMSE

RMSE in Table 3.7 confirms better convergence for the MLE. In the Figure 3.5 the third graph (RMSE Bayes(opt)) illustrates the slow convergence very well.

In Table 3.7 we could notice that the Brier score is very low. Compared to the Brier score in the first case (Table 3.3), the values here are almost ten times lower.

Confusion matrices in tables 3.8, 3.9 and 3.10 confirm this, showing us that almost 98% of the samples were classified correctly and that the error is akin for both classes. We may want again to look at how the classification using real θ would look in this case. The Brier score, in this case, was again close to 0.02. This brings us to a thing that the balanced distribution of the values in the random component affects the prediction more than the perfect convergence to parameters used to generate the random component. This may happen due to the element of randomness in generating the data we then classify.

	pred 0	pred 1	sum
0	4166	144	4300
1	141	5549	5700
sum	4307	5693	10000

Table 3.8 Simulated data case 2: Confusion matrix for MLE

	pred 0	pred 1	sum
0	4173	137	4300
1	148	5542	5700
sum	4321	5679	10000

Table 3.9 Simulated data case 2: Confusion matrix for one-step Bayes

	pred 0	pred 1	sum
0	4168	142	4300
1	139	5551	5700
sum	4307	5693	10000

Table 3.10 Simulated data case 2: Confusion matrix for optimized Bayes

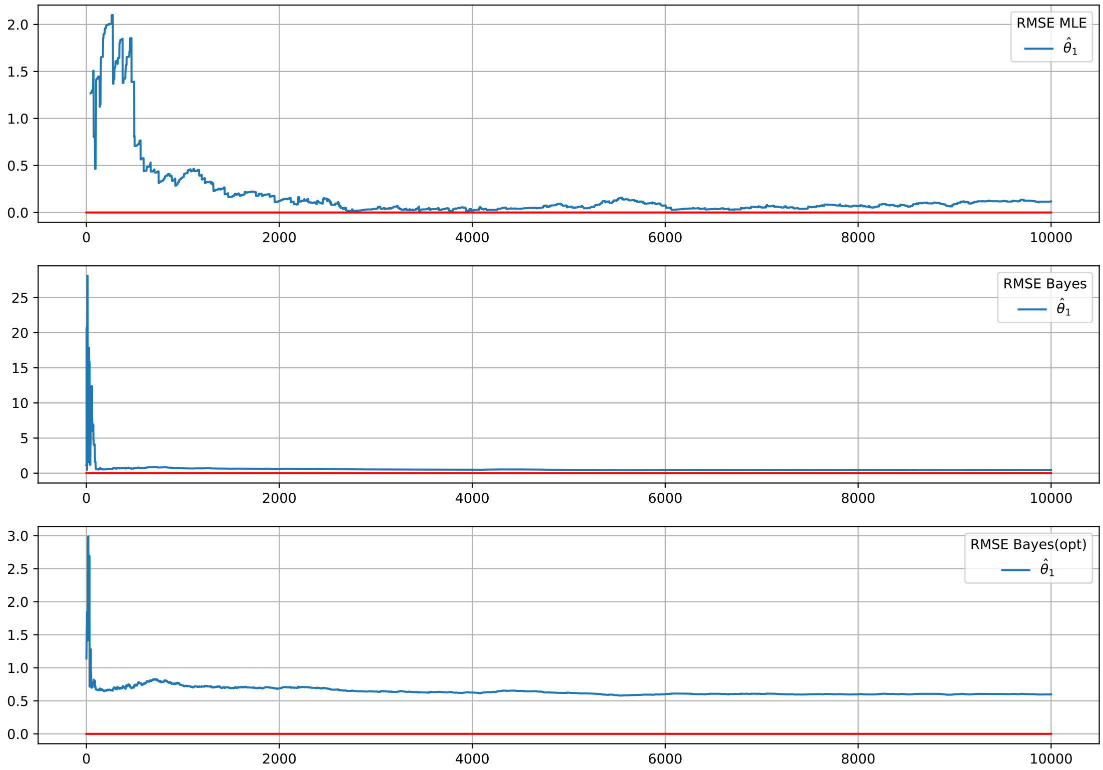


Figure 3.5 Simulated data case 2: RMSE shows that none of the estimators converges

3.2.3 Simulated data case 3: One-step Bayes estimator shows unstandard behaviour

There are cases when the one-step Bayes estimator oscillates wildly around the true values of θ. Although it always converged in 10000 steps for data generated with both coefficient parameters, the convergence did not sometimes occur for smaller datasets. An example of the behavior mentioned can be seen in Figure 3.6. In Figure 3.6 the one-step Bayes estimator deviates extremely for more than 1500 samples. After that, it starts converging to the real values.

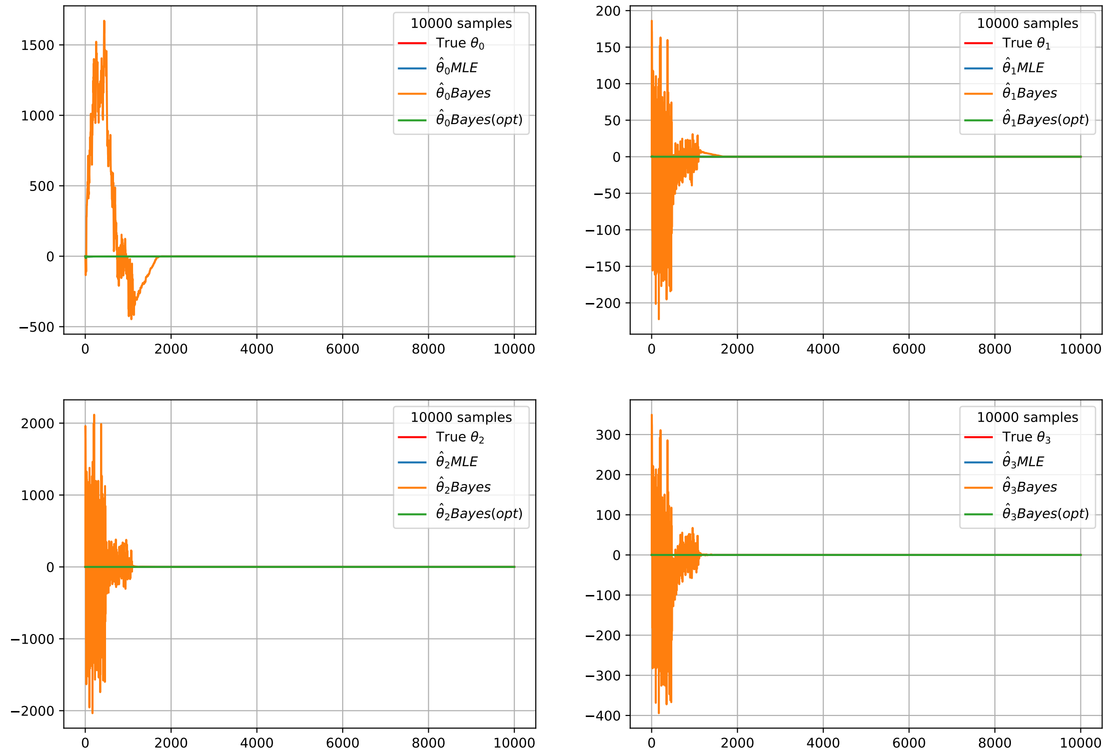


Figure 3.6 Simulated data case 3a: One-step Bayes estimator deviates before converging

We may want to look into the RMSE and Brier score of the final estimates. Due to extreme deviations from the real values, the graph does not show the convergence quality for other models. We can get some idea about it from the graphs depicting the RMSE evolution, shown in Figure 3.7 which shows us that the convergence of other models was similar to other cases with no anomalies. We notice that the MLE estimates were atypically good, but that seems to be a coincidence.

Method	Brier score	RMSE
MLE	0.209024	0.001846
One-step Bayes	0.209030	0.002502
Optimized Bayes	0.209029	0.011509

Table 3.11 Simulated data case 3a: Final Brier score and RMSE

The RMSE in table 3.11 shows better convergence for the one-step Bayes estimator with RMSE lower than 0.005. This occurred every time the one-step Bayes estimator showed similar behavior; the final RMSE was as low or even lower than in the cases when the convergence started early.

The next thing we can look at is the predictions. As the Brier score suggests, confusion matrices will not give us any new information. The prediction quality is similar to other cases when the first set of θ was used.

What may interest us more is how the prediction would be affected if we used the result where the one-step Bayes estimator has not converged at all as in figure 3.8. Here, the estimate for θ_0 (coefficient for the intercept) is receding from its real value. On the contrary, other parameters oscillate around the true θ values. Overall RMSE shown in figure 3.9 does not show any sign of convergence.

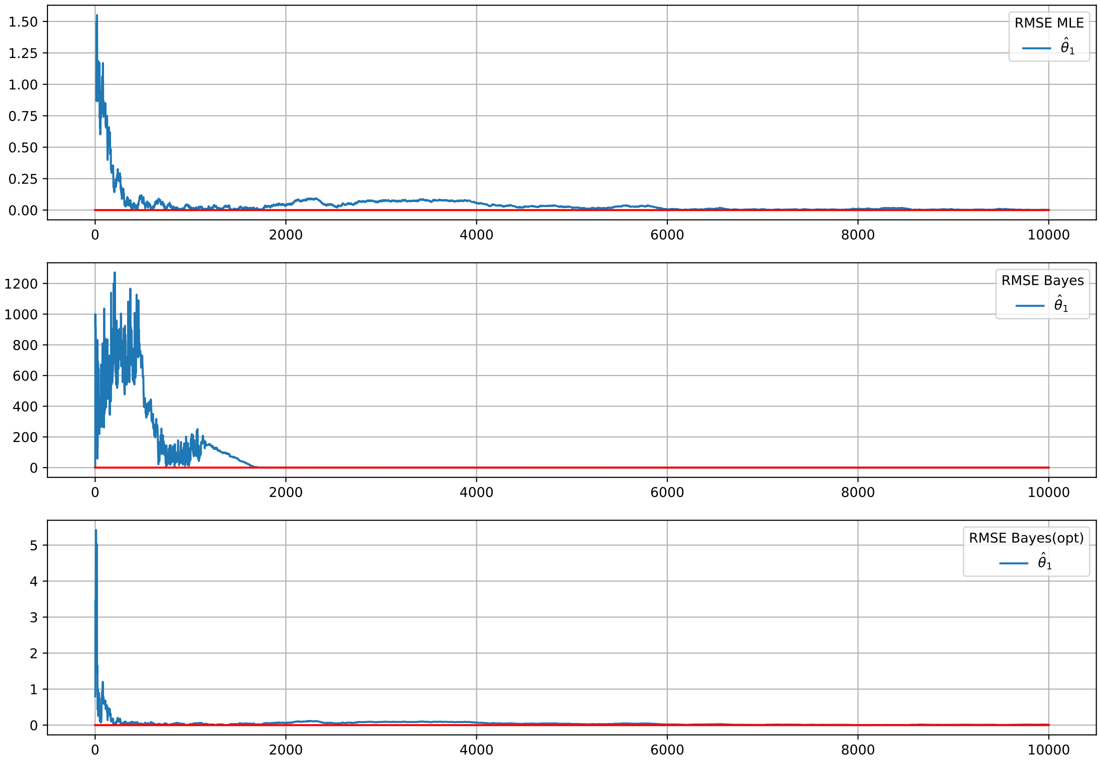


Figure 3.7 Simulated data case 3a: RMSE shows convergence after 2000 samples for one-step Bayes.

RMSE for one-step Bayes estimator shown in Table 3.12 is extremely high compared to other two. The Brier score is still at approximately 0.6. The value may still seem small, but we have to keep in mind that even if the estimator classified all samples into one class, it would still have the Brier score at about 0.7 for the first set of θ coefficients.

When comparing the confusion matrices in Tables 3.13 and 3.14, we see that the one-step Bayes estimator tends to assign most of the samples to the class 1, as opposed to the optimized version. While the behavior of the optimized version makes sense because of the elements of randomness and abundant representation of 0 elements in the random component, the behavior of the one-step version points to the values of the random component of the model being extremely high, thus deviating the estimates towards 1.

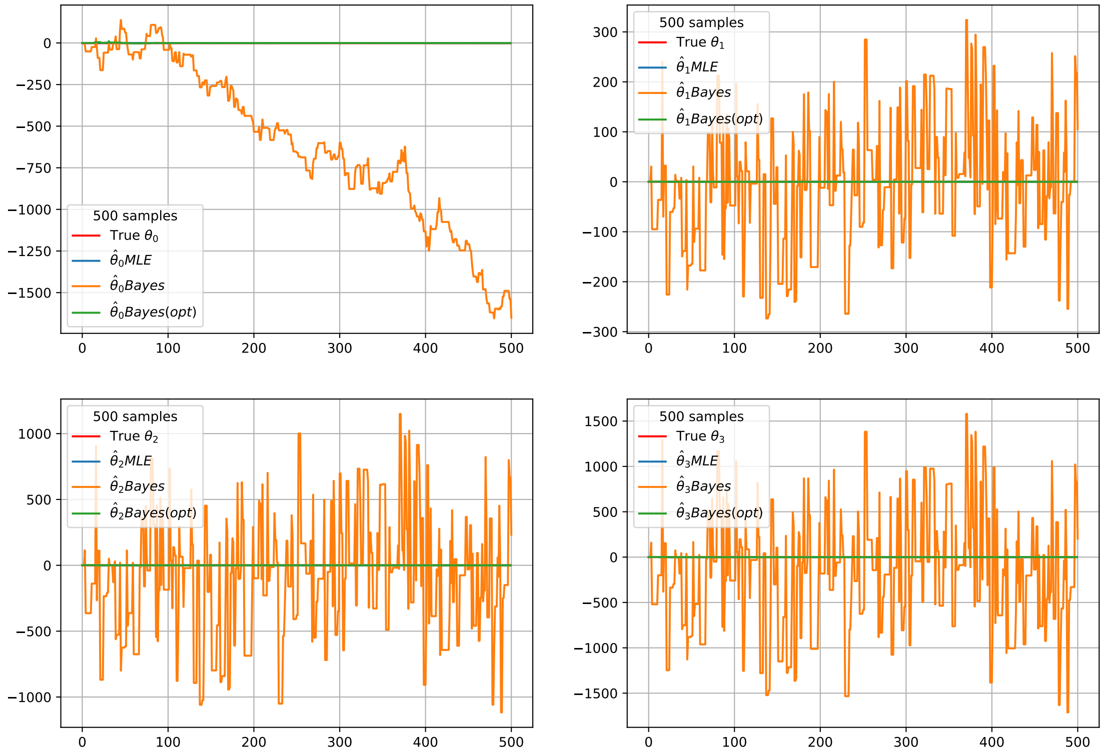


Figure 3.8 Simulated data case 3b: One-step Bayes estimator not converging on smaller datasets

Method	Brier score	RMSE
MLE	0.208173	0.401960
One-step Bayes	0.612693	840.694971
Optimized Bayes	0.208631	0.368250

Table 3.12 Simulated data case 3b: Final Brier score and RMSE

	pred 0	pred 1	sum
0	39	297	336
1	10	154	164
sum	49	451	500

Table 3.13 Simulated data case 3b: Confusion matrix for one-step Bayes

	pred 0	pred 1	sum
0	292	44	336
1	127	37	164
sum	419	81	500

Table 3.14 Simulated data case 3b: Confusion matrix for optimized Bayes

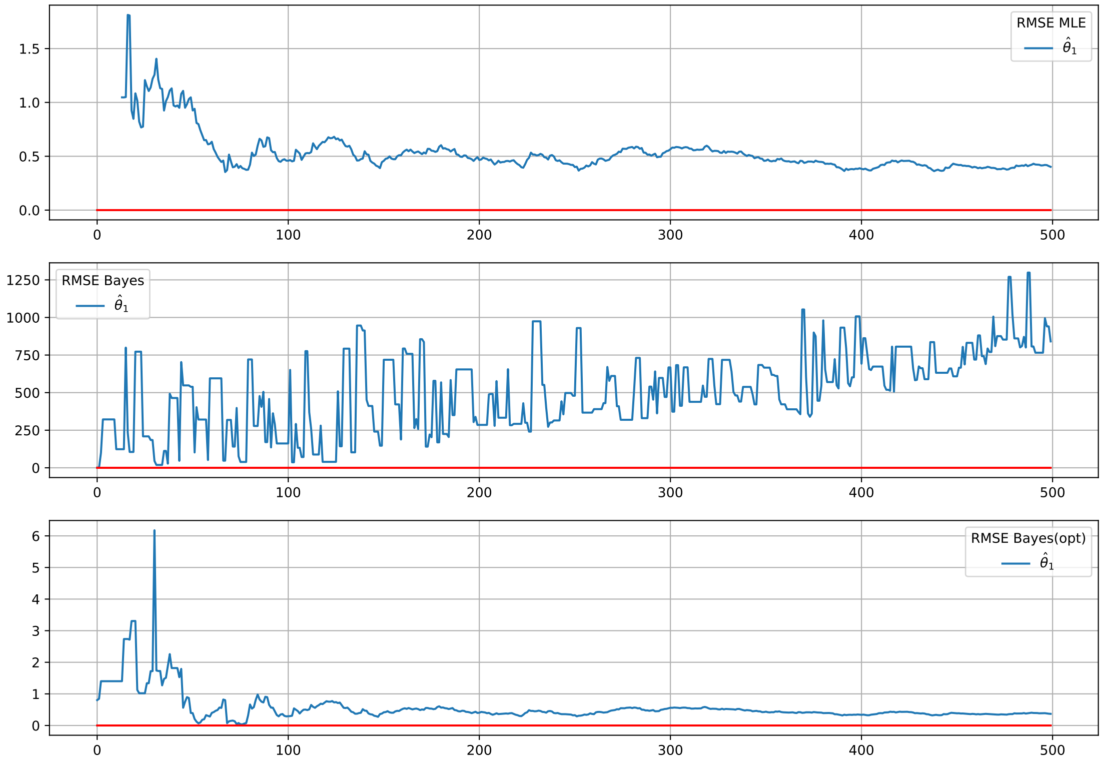


Figure 3.9 Simulated data case 3b: RMSE shows that one-step Bayes estimator diverges from the real θ

In this section, we went through the most typical cases of the behavior of the logistic model that was encountered during our experiments on simulated data samples. We could see that the optimized version of the Bayes estimator had similar results to its one-step counterpart. Its benefits fully manifested in the third case, where it was visible that the optimized Bayes are less prone to extreme deviations for smaller amounts of data (there was no such case in our experiments where the optimized version would behave this way).

What could affect the comparison quality was that the data were generated randomly and had no real correlation. In the next section, we will provide the results from the experiments on a real-life dataset.

3.3 Experiments on real data

The Skin-NonSKin dataset (Bhatt and Dhall [38]) was chosen for experiments on real-life data. The dataset contains 245057 samples (50859 skin and 194198 non-skin samples - the ratio is roughly 20:80) of pixels that either do or do not belong to a human face. The samples consist of BGR values of the pixel, which are our regressor and the classification labels - they will create the random component of our models. To differentiate the results, the dataset was shuffled randomly.

Because for real-life datasets, we usually do not have the real θ values available, thus we are unable to compute RMSE. For this reason, in this section, we will compute the RMSE from the maximum likelihood estimate we receive from using MLE estimator on 100,000 samples.

Predictions using this estimate have a Brier score of about 0.06, and the logistic model has a more than 91% success rate in classifying the samples. For clarification will call it an *overall MLE estimate* in the following text.

During the prediction, we have encountered situations similar to the one described in subsection 3.2.3 more often than with a randomly generated dataset. This can be because the ratio of elements of the random component is even more unbalanced than in the data generated by the first set of θ parameters.

We will now present two cases - one with all estimators converging and one with one-step Bayes deviating.

3.3.1 Real-life data case 1: All estimators converge to the MLE

In this case, all values have converged, as can be seen in figure 3.10.

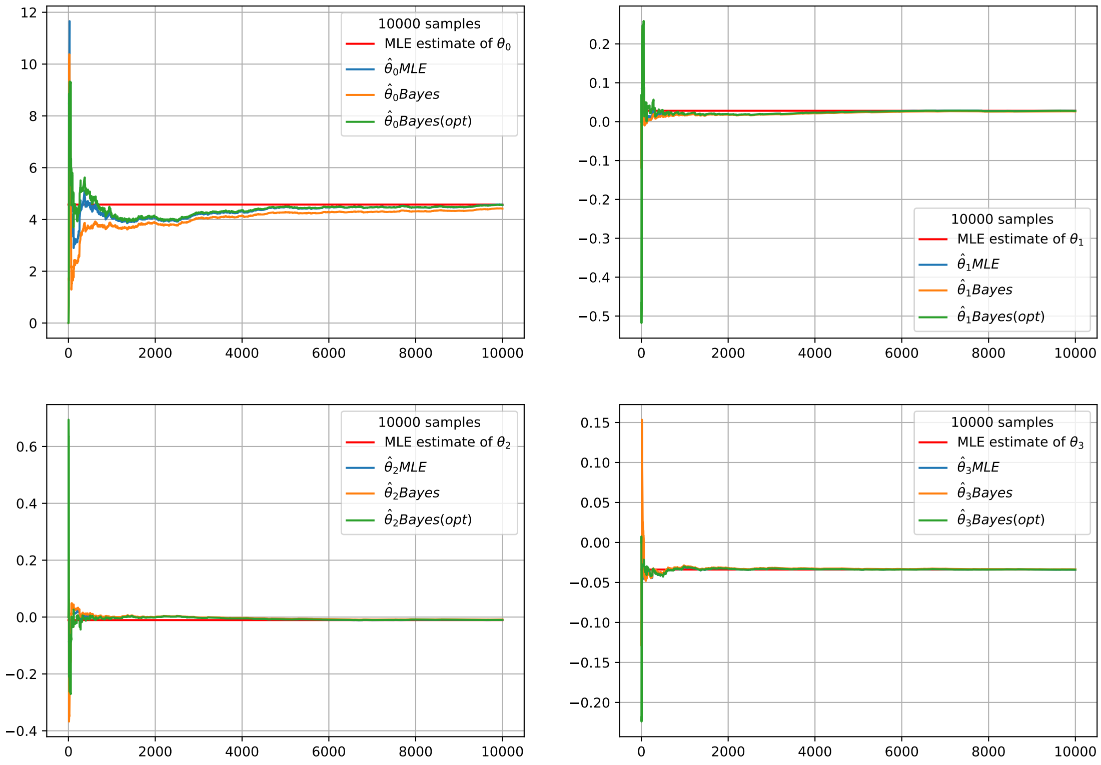


Figure 3.10 Real-life data case 1: All estimators converge to the overall MLE estimate

All estimators have deviated initially, which often happened with this dataset. Then we can see that the convergence was relatively slow for all of them, with the one-step Bayes being slightly slower. RMSE graphs in figure 3.11 also confirm this. The RMSE graphs show that both optimized Bayes and MLE estimator reached the overall MLE estimate very quickly and then deviated a bit before converging. The next thing to notice is that the one-step Bayes estimator does not get as close to the overall MLE estimate as the optimized version does. This is similar in most cases, where the estimators have all converged.

The metrics show in Table 3.15 confirm better convergence for MLE and optimized Bayes estimator. The Brier score also corresponds with the Brier score of the overall MLE estimate.

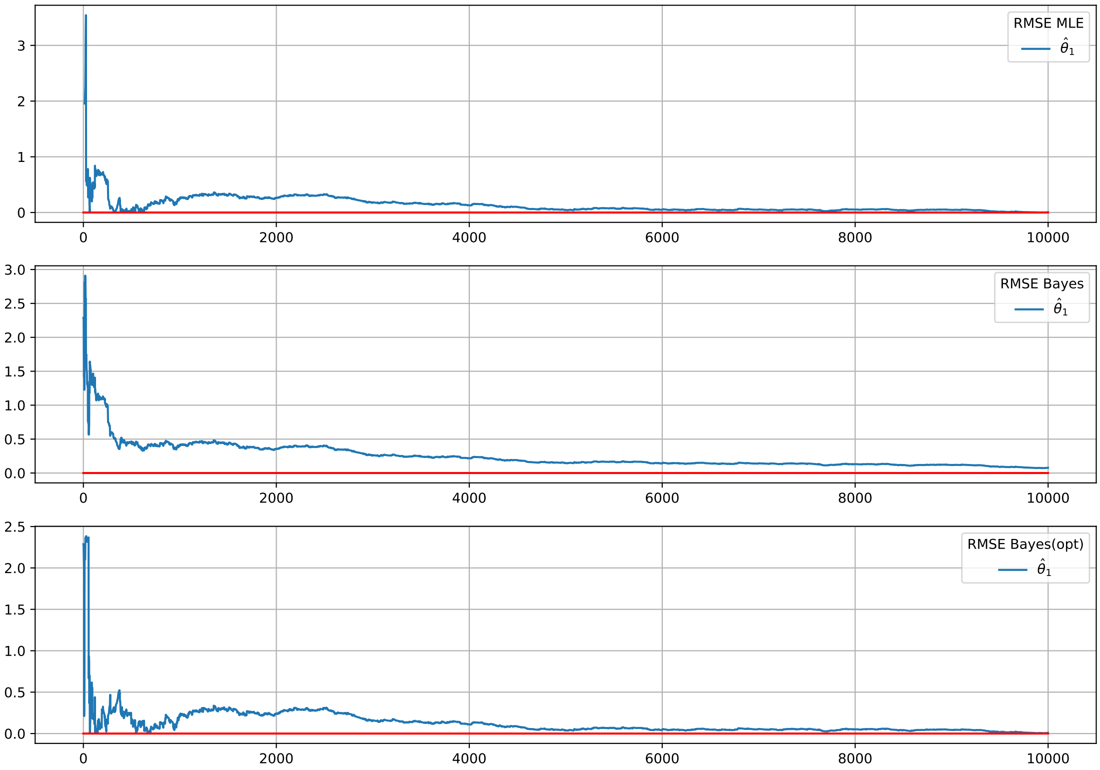


Figure 3.11 Real-life data case 1: RMSE shows good convergence for all estimators, slightly worse for one-step Bayes

The confusion matrices in Tables 3.16 3.17 and 3.18 show that the MLE and optimized Bayes estimators have almost identical prediction with the same number of misclassified samples. The one-step Bayes estimator may seem to return slightly more false positives than the other two, but repeating the experiments has shown that this is a coincidence; the estimators seem to have similar quality of predictions for the 10,000 samples when it has converged

Method	Brier score	RMSE
MLE	0.068342	0.005086
One-step Bayes	0.068641	0.0780082
Optimized Bayes	0.068436	0.009305

Table 3.15 Real-life data case 1: Final Brier score and RMSE

This section has reviewed the case when all estimators converge relatively early. We can see that the final prediction is almost the same quality.

	Pred skin	Pred non-skin	sum
Skin	1642	373	2015
Non-skin	436	7549	7985
sum	2078	7922	10000

Table 3.16 Real-life data case 1: Confusion matrix for MLE

	Pred skin	Pred non-skin	sum
Skin	1641	374	2015
Non-skin	435	7550	7985
sum	2076	7924	10000

Table 3.17 Real-life data case 1: Confusion matrix for one-step Bayes

	Pred skin	Pred non-skin	sum
Skin	1641	374	2015
Non-skin	442	7543	7985
sum	2083	7917	10000

Table 3.18 Real-life data case 1: Confusion matrix for optimized Bayes

3.3.2 Real-life data case 2: One-step Bayes estimator does not converge

In this case, the one-step Bayes deviates significantly from the overall MLE estimate. In Figure 3.12 we can see that the convergence has not begun even after the model processed 10,000 samples. Instead, the intercept value recedes from the overall MLE estimate while the other values oscillate around it. Figure 3.13 shows that the RMSE of the on-step Bayes estimator is instead growing as for the 10,000 samples. On the contrary, the other two graphs show very good convergence. This has not happened with the simulated data - the convergence always begins before reaching 10,000 samples for four regressors.

Method	Brier score	RMSE
MLE	0.064811	0.086263
One-step Bayes	0.1441	0.059961
Optimized Bayes	0.064841	7281.710264

Table 3.19 Real-life data case 2: Final Brier score and RMSE

The metrics in the table 3.19 show us that although the RMSE is extremely high for the one-step Bayes estimator, the Brier score is lower than in most cases from the estimations made on simulated datasets.

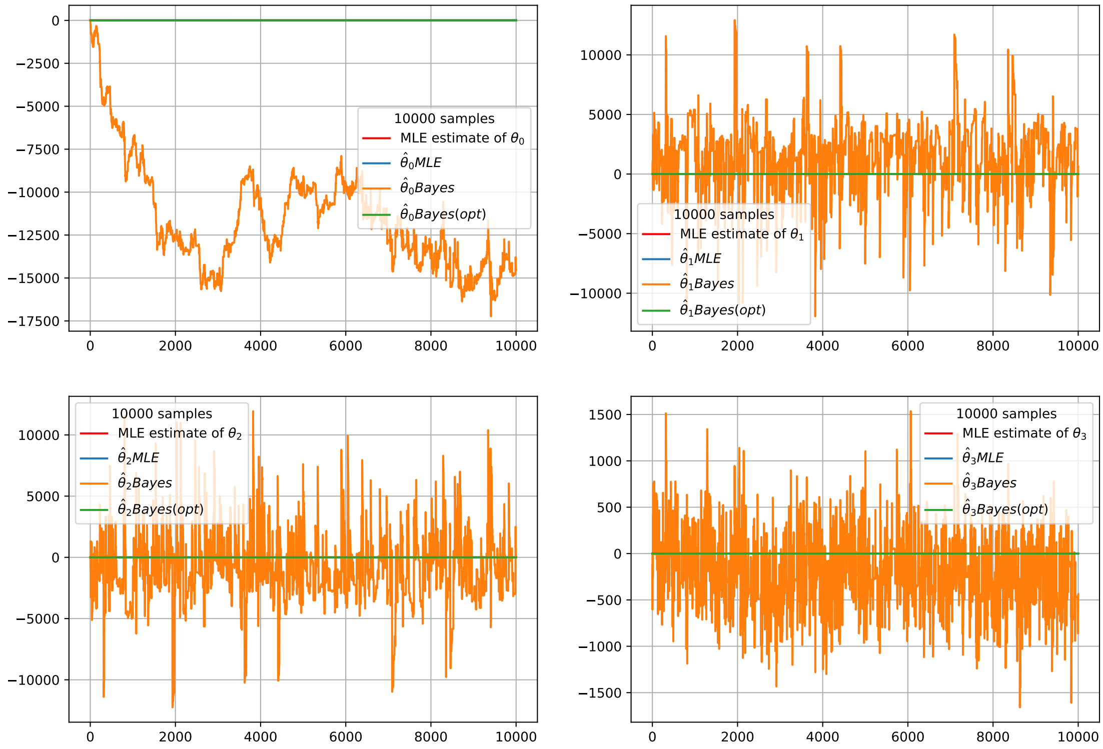


Figure 3.12 Real-life data case 2: Bayes estimator fails

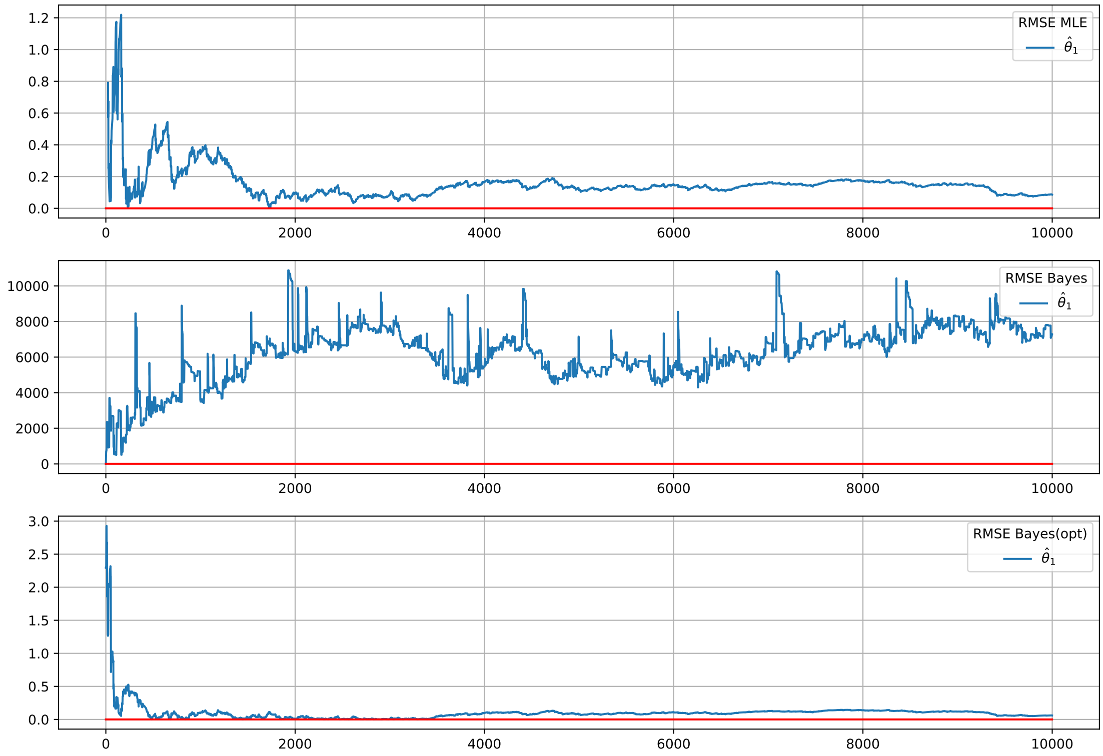


Figure 3.13 Real-life data case 2: RMSE

	Pred skin	Pred non-skin	sum
Skin	1750	355	2105
Non-skin	428	7467	7895
sum	2178	7822	10000

■ **Table 3.20** Real-life data case 2: Confusion matrix for MLE

	Pred skin	Pred non-skin	sum
Skin	1972	133	2105
Non-skin	1308	6587	7895
sum	3280	6720	10000

■ **Table 3.21** Real-life data case 2: Confusion matrix for one-step Bayes

	Pred skin	Pred non-skin	sum
Skin	1750	355	2105
Non-skin	428	7467	7895
sum	2178	7822	10000

■ **Table 3.22** Real-life data case 2: Confusion matrix for optimized Bayes

Confusion matrices in tables 3.20, 3.21 and 3.22 show us that the MLE and optimized Bayes estimator give us identical predictions, comparable to the ones in previous case (tables 3.16 and 3.18). On the other hand, the one-step Bayes tends to give us many false positives, which we suppose not to be the actual property of the one-step Bayes estimator on this dataset - in other cases where one-step Bayes has diverged, the situation was inverse. The model returned a higher number of false negatives.

This section has reviewed the case where the one-step Bayes estimator have not converged. This case was often in our experiments on the Skin non-skin datasets, happening in more than 50% of all cases.

Chapter 4
Conclusion

This thesis aimed to observe the impact of the quality of the approximation of the posterior distribution in Bayesian logistic regression on the estimation of parameters of the logistic model.

Our observations were made on simulated datasets, generated using two different coefficient parameter vectors, and then on a real-life dataset. From the results of the experimental part of this thesis, we can see that the quality of the estimation of the parameters is heavily affected by the distribution of the members of the random component of a used model. With the ratio approaching 1:1, we could see that the convergence was slow compared to the cases where the ratio was closer to 35:65, where the convergence was faster and always happened.

Actual predictions made with estimated parameters showed that even though the convergence was better, for simulated datasets, the number of misclassified samples rose with the inequality in the distribution of members of the random component. This did not happen with the real-life dataset, even though the ratio was close to 1:4. The Brier score was low, and the predictions were good. This leads us to think that the problem is in the random generation of the data, and the missing correlation damages the prediction.

The next thing that could be noticed is that the one-step Bayesian estimator was prone to deviate from the real values of the estimated parameters, as can be seen in 3.6. This did not happen with the optimized version nor with the MLE estimator. All deviations were just for the first few samples, then the convergence began. On the real-life dataset, the deviations were often and happened in more than half of the cases.

In the cases where the convergence occurred, all estimators returned similar results. One could ask why was not the optimized Bayes estimator always better than the one-step version. The problem here could be that even though we have used an algorithm much more likely to find a global maximum, we are still working with approximations. As such, there will always be irregularities and mistakes.

The results of this thesis show advantages of optimizations in Bayes logistic regression, and further work can be done to research its impact combined with other optimization methods, such as the ones described in [35].

Bibliography

1. G, Ambrish; GANESH, Bharathi; GANESH, Anitha; SRINIVAS, Chetana; DHANRAJ; MENSINKAL, Kiran. Logistic Regression Technique for Prediction of Cardiovascular Disease. *Global Transitions Proceedings*. 2022. ISSN 2666-285X. Available from DOI: `https://doi.org/10.1016/j.gltp.2022.04.008`.
2. LEE, Jong Wook; SOHN, So Young. Evaluating borrowers' default risk with a spatial probit model reflecting the distance in their relational network. *PLOS ONE*. 2022, vol. 16, no. 12, pp. 1–11. Available from DOI: `10.1371/journal.pone.0261737`.
3. SAVOLAINEN, Peter T.; SHARMA, Anuj; GATES, Timothy J. Driver decision-making in the dilemma zone – Examining the influences of clearance intervals, enforcement cameras and the provision of advance warning through a panel data random parameters probit model. *Accident Analysis & Prevention*. 2016, vol. 96, pp. 351–360. ISSN 0001-4575. Available from DOI: `https://doi.org/10.1016/j.aap.2015.08.020`.
4. METSIS, Vangelis; ANDROUTSOPOULOS, Ion; PALIOURAS, Georgios. Spam Filtering with Naive Bayes - Which Naive Bayes? In: 2006.
5. FRIEDMAN, Nir; GEIGER, Dan; GOLDSZMIDT, Moises. Bayesian network classifiers. *Machine learning*. 1997, vol. 29, no. 2, pp. 131–163.
6. ROKACH, Lior; MAIMON, Oded. Decision Trees. In: 2005, vol. 6, pp. 165–192. Available from DOI: `10.1007/0-387-25465-X_9`.
7. RAJESWARI, S.; SUTHENDRAN, K. C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. *Computers and Electronics in Agriculture*. 2019, vol. 156, pp. 530–539. ISSN 0168-1699. Available from DOI: `https://doi.org/10.1016/j.compag.2018.12.013`.
8. WINSTER, S Godfrey; KUMAR, M Naveen. Automatic classification of emotions in news articles through ensemble decision tree classification techniques. *Journal of Ambient Intelligence and Humanized Computing*. 2021, vol. 12, no. 5, pp. 5709–5720.
9. SAHU, Garima; KUMAR KHARE, Rakesh. Decision Tree Classification based Decision Support System for Derma Disease. *International journal of computer applications*. 2014, vol. 94, no. 17, pp. 21–26. ISBN 0975-8887.
10. CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. *Machine learning*. 1995, vol. 20, no. 3, pp. 273–297. ISBN 0885-6125.

11. JOACHIMS, Thorsten. Text categorization with Support Vector Machines: Learning with many relevant features. In: NÉDELLEC, Claire; ROUVEIROL, Céline (eds.). *Machine Learning: ECML-98*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 137–142. ISBN 978-3-540-69781-7.

12. COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 1967, vol. 13, no. 1, pp. 21–27. Available from DOI: `10.1109/TIT.1967.1053964`.

13. HAVAEI, Mohammad; JODOIN, Pierre-Marc; LAROCHELLE, Hugo. Efficient Interactive Brain Tumor Segmentation as Within-Brain kNN Classification. In: *2014 22nd International Conference on Pattern Recognition*. 2014, pp. 556–561. Available from DOI: `10.1109/ICPR.2014.106`.

14. KAMENCAY, Patrik; ZACHARIASOVA, Martina; HUDEC, Robert; JARINA, Roman; BENCO, Miroslav; HLUBIK, Jan. A novel approach to face recognition using image segmentation based on spca-knn method. *Radioengineering*. 2013, vol. 22, no. 1, pp. 92–99.

15. EL-DOSUKY, Mohamed A.; SOLIMAN, Mona; HASSANIEN, Aboul Ella. COVID-19 vs influenza viruses: A cockroach optimized deep neural network classification approach. *International Journal of Imaging Systems and Technology*. 2021, vol. 31, no. 2, pp. 472–482. Available from DOI: `https://doi.org/10.1002/ima.22562`.

16. BASU, Saikat; MUKHOPADHYAY, Supratik; KARKI, Manohar; DIBIANO, Robert; GANGULY, Sangram; NEMANI, Ramakrishna; GAYAKA, Shreekant. Deep neural networks for texture classification—A theoretical analysis. *Neural Networks*. 2018, vol. 97, pp. 173–182. ISSN 0893-6080. Available from DOI: `https://doi.org/10.1016/j.neunet.2017.10.001`.

17. JIN, Rize; LU, Liangfu; LEE, Joomin; USMAN, Anwar. Multi-representational convolutional neural networks for text classification. *Computational Intelligence*. 2019, vol. 35, no. 3, pp. 599–609. Available from DOI: `https://doi.org/10.1111/coin.12225`.

18. NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*. 1972, vol. 135, no. 3, pp. 370–384. Available from DOI: `https://doi.org/10.2307/2344614`.

19. CRAMER, Jan Salomon. The origins of logistic regression. 2002.

20. VERHULST, PF. La loi d'accroissement de la population. *Nouveaux Memories de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*. 1845, vol. 18, pp. 14–54.

21. BLISS, C. I. The Method of Probits. *Science*. 1934, vol. 79, no. 2037, pp. 38–39. Available from DOI: `10.1126/science.79.2037.38`.

22. GADDUM, John Henry. *Reports on Biological Standards: III Methods of Biological Assay Depending on a Quantal Response*. HM Stationery Office, 1949.

23. METROPOLIS, Nicholas; ULAM, S. The Monte Carlo Method. *Journal of the American Statistical Association* [online]. 1949, vol. 44, no. 247, pp. 335–341 [visited on 2022-05-11]. ISSN 01621459. Available from: `http://www.jstor.org/stable/2280232`.

24. RUBIN, Donald B. The bayesian bootstrap. *The annals of statistics*. 1981, pp. 130–134.

25. NEWTON, Michael A.; RAFTERY, Adrian E. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1994, vol. 56, no. 1, pp. 3–26.

26. LEWIS, Steven M.; RAFTERY, Adrian E. Estimating Bayes Factors via Posterior Simulation With the Laplace-Metropolis Estimator. *Journal of the American Statistical Association* [online]. 1997, vol. 92, no. 438, pp. 648–655 [visited on 2022-05-11]. ISSN 01621459. Available from: `http://www.jstor.org/stable/2965712`.

27. GEMAN, Stuart; GEMAN, Donald. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*. 1984, no. 6, pp. 721–741.

28. SUBANAR. Data Analysis Comparison Logit and Probit Regression Using Gibbs-Sampler. In: KOR, Liew-Kee; AHMAD, Abd-Razak; IDRUS, Zanariah; MANSOR, Kamarul Ariffin (eds.). *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*. Singapore: Springer Singapore, 2019, pp. 309–315. ISBN 978-981-13-7279-7.

29. BISHOP, Christopher M; NASRABADI, Nasser M. *Pattern recognition and machine learning*. Springer, 2006. No. 4.

30. MENG, Xiao-Li; WONG, Wing Hung. SIMULATING RATIOS OF NORMALIZING CONSTANTS VIA A SIMPLE IDENTITY: A THEORETICAL EXPLORATION. *Statistica Sinica* [online]. 1996, vol. 6, no. 4, pp. 831–860 [visited on 2022-05-11]. ISSN 10170405, ISSN 19968507. Available from: `http://www.jstor.org/stable/24306045`.

31. GELMAN, Andrew; MENG, Xiao-Li. Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling. *Statistical Science* [online]. 1998, vol. 13, no. 2, pp. 163–185 [visited on 2022-05-11]. ISSN 08834237. Available from: `http://www.jstor.org/stable/2676756`.

32. ORMEROD, John T; WAND, Matt P. Explaining variational approximations. *The American Statistician*. 2010, vol. 64, no. 2, pp. 140–153.

33. FLETCHER, Roger. *Practical methods of optimization*. John Wiley & Sons, 2013.

34. VIRTANEN, Pauli; GOMMERS, Ralf; OLIPHANT, Travis E.; HABERLAND, Matt; REDDY, Tyler; COURNAPEAU, David; BUROVSKI, Evgeni; PETERSON, Pearu; WECKESSER, Warren; BRIGHT, Jonathan; VAN DER WALT, Stéfan J.; BRETT, Matthew; WILSON, Joshua; MILLMAN, K. Jarrod; MAYOROV, Nikolay; NELSON, Andrew R. J.; JONES, Eric; KERN, Robert; LARSON, Eric; CAREY, C J; POLAT, İlhan; FENG, Yu; MOORE, Eric W.; VANDERPLAS, Jake; LAXALDE, Denis; PERKTOLD, Josef; CIMRMAN, Robert; HENRIKSEN, Ian; QUINTERO, E. A.; HARRIS, Charles R.; ARCHIBALD, Anne M.; RIBEIRO, Antônio H.; PEDREGOSA, Fabian; VAN MULBREGT, Paul; SCIPY 1.0 CONTRIBUTORS. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*. 2020, vol. 17, pp. 261–272. Available from DOI: `10.1038/s41592-019-0686-2`.

35. MACKAY, David JC. Choice of basis for Laplace approximation. *Machine learning*. 1998, vol. 33, no. 1, pp. 77–86.

36. ALAN, Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley, 2015. Wiley Series in Probability and Statistics. ISBN 9781118730034. Available also from: `https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=941245&lang=cs&site=ehost-live&scope=site`.

37. HAUG, Anton J. *Bayesian estimation and tracking a practical guide*. Wiley, 2012.

38. BHATT, Rajen; DHALL, Abhinav. Skin segmentation dataset. *UCI Machine Learning Repository*. 2010.

Enclosed medium contents

- `readme.txt` ... description of the content
- `src`
  - `impl` ... Source codes of implementation
  - `thesis` ... thesis source code in form of LaTeX
- `text` ... text of the thesis
  - `thesis.pdf` ... text of the thesis in the form of a PDF file