An Introduction to Bayesian Linear Regression

Kaspara Skovli Gåsvær[*]
University of Oslo - Department of Physics

(Dated: June 22, 2021)

The purpose of this paper is to serve as a gentle introduction to Bayesian linear regression. It is a take on the subject from the view of someone not primarily a statistician which we believe can lead to easier understanding and better intuition for most non-statisticians who are first entering the field of Bayesian statistics. We introduce the basics of Bayesian inference before diving a bit deeper into the parts that make up Bayes theorem. We later expand to methods of estimation and how to interpreted them as well as the concepts of model selection and parameter estimation.

I. INTRODUCTION

Anyone pursuing a degree in science is at some point bound to encounter a course in frequentist statistics. At first it might seem alright, maybe even somewhat intuitive and definitely useful for understanding how to interpret data. The problem is that most courses in frequentist statistics is made for statisticians, not for those wishing to use it as a tool without having to be a true expert in the matter. The curriculum often quickly turns in to a soup of instructions and procedure that is hard to get a real grip on and many experience difficulty in combining it all into something useful. This raises the question if there could be another approach more suited to ease the entrance to the subject for physicists, biologists or other scientists who are not primarily statisticians.

This article provides an introduction to bayesian statistics, especially the core concepts of performing bayesian linear regression. Bayesian statistics bases itself on probability, but in a way that focuses mostly on how certain we are that something is true or not. That's to say it uses probability as a way to express our confidence in an experiment or result. It has Bayes' theorem at its core which in itself is quite easy to explain or validate. The theorem can be built on to further ones understanding of statistical methods and we expect that for many a much more tangible approach to statistics than the frequentist one.

We will take you through the main differences in bayesian and frequentist statistics before diving into the specifics of bayesian models. We will cover some ground on the different parts used in Bayes' theorem, probability distributions and methods of estimation. This article is based largely on the books *Data Analysis: A Bayesian Tutorial* (Sivia and Skilling 2006) and *Pattern Recognition and Machine Learning* (Bishop 2006). We strongly recommend these for the eager reader as both have an intuitive take on bayesian statistics and its applications.

II. BAYESIAN VS. FREQUENTIST STATISTICS

The bayesian and frequentist idea of statistics is in many ways two sides of the same coin. Lets say we have two statisticians, Mr.F who is a frequentist and Mr.B who prefers bayesian statistics. They share the same goal of extracting as much useful information about a system as possible, but with different approaches. Given a set of data Mr.F might ask himself "What can my model tell me about this data?" whereas Mr.B might ask himself "What can this data tell me about the model?". When Mr.F is performing an experiment he will make predictions based exclusively on the data from said experiment. In other words he will do something many times, take note of all the outcomes and predict by the outcomes. This is used to calculate the probability of getting the same results if one were to replicate the experiment exactly. On the other hand, Mr.B would stop and think to himself "Do I have any prior knowledge, *inference*, about my system?" before performing any experiment at all. He will use this information to construct a *prior* probability distribution which will be a starting point for his hypothesis about a suitable model. Only then will he start performing his experiment and with each trial he will use the results via a *likelihood* function to update the prior. The likelihood function is a measure of how well the model fits the data and will update the prior accordingly. After performing a number of experiments the updates of the prior will yield a *posterior* distribution, which is what we believe to be the truth of the system based on the data (Sivia and Skilling 2006, Ch.1.3). The posterior is as mentioned a probability distribution, not a point estimate like in the case of Mr.F.

The updating of the prior which helps to produce the posterior is called *bayesian inference* and is done using Bayes' theorem

$$P(H|D,I) = \frac{P(D|H,I)P(H|I)}{P(D|I)}, \tag{1}$$

where H stands for hypothesis/model, D stands for data and I for prior inference about the system. For problems concerning parameter estimation the calculations can most often be made ignoring the term in the denominator, which is called the *evidence*, as it is only a

[*] Code repository: https://github.com/kasparagaasvaer

normalization constant which does not explicitly depend on the hypothesis. When dealing with model selection it can play a rather large role which we will get into later on. We can identify the posterior distribution as $P(H|D,I)$, the likelihood function as $P(D|H,I)$ and the prior distribution as $P(H|I)$ in the above equation.

Whether the approach of Mr.F or Mr.B is the "best" one boils down to a question of preference. One can argue that for some problems one or the other is more suited, but again that really depends on what kind of information you are looking for and the form of that information.

III. PRIORS

Lets return to some of the terms mentioned in the previous section. A prior probability distribution can express our knowledge or ignorance about a system. There are many different priors to choose which can be roughly divided into two main subgroups; *informative* and *uninformative* priors. An informative prior contains definite pre-existing information about a system, for example a prior distribution based on the probability of a coin toss. If we have performed many experiments (tossed the same coin a lot of times) and gotten a roughly 50/50 ratio of heads/tails, we can infer that the coin is fair. For later experimentation on the coin we can then produce a prior probability distribution using that information about the coin. That's to say that the posterior distribution of one set of coin tosses becomes the prior of another set.

Uninformative priors are more objective and are made up of information like an interval we know the variable lies on or something like the fact that the variable is nonzero etc, i.e it often reflects more on the lack of quantifiable knowledge about the system. Using an uninformative prior can be a smart choice when one has little prior knowledge about a system as it reduces the dependence on the prior, while if one has quantifiable inference about the system a better choice would usually be an informative one (Bishop 2006, Ch.1.2.3). A *weakly informative* prior falls somewhere in between the two and is what we call a prior where we either purposefully exclude information in the prior or want to include our ignorance about the system. It leads to a posterior less dependent on the prior and can often be used for regularization (Lemoine 2019).

One special case of prior is a *Conjugate prior*. Conjugate priors ensure that the posterior distribution will take the same form as the prior. If the prior is Gaussian, which is conjugate to it self, and the likelihood function is too then the posterior is guaranteed to be Gaussian as well. Conjugate priors are useful in the way that they reduce the Bayesian inference to a choice of hyperparameters instead of the calculation of often tricky integrals, which is often made use of in Bayesian Linear regression models (Sivia and Skilling 2006, Ch.5.5.2). In this case one needs a way to perform estimation on the hyperparameters which typically calls for a *hyperprior*. This is not a prior distribution for our model itself, but rather a prior over the hyperparameters defining our model. As for the updating of our model prior, the hyperprior is updated according to new evidence as to fit the hyperparameters of the model to the data.

Another thing to note about priors is that the more empirical evidence one gets from doing experiments the less dependent the posterior will be of the prior, that's to say that the likelihood function dominates the posterior. Figure 1 illustrates a scenario where we have n samples drawn randomly from a normal distribution. We define both prior and likelihood function as normal probability density distributions and calculate the posterior using Bayes theorem. One can observe that as the number of samples grows the likelihood function dominates the posterior. This means that for prediction purposes, the choice of prior is important when dealing with small samples but becomes less important with large ones.

IV. THE LIKELIHOOD FUNCTION

The following section is based largely on (Bishop 2006, Ch.3.1).

As previously mentioned the likelihood function is a measure of how well our model fits the data and is used to update the prior distribution when performing Bayesian inference. If we assume a target variable t given by a deterministic function $y(\mathbf{x},\mathbf{w})$ with additive Gaussian noise so that $t = y(\mathbf{x},\mathbf{w}) + \epsilon$, where ϵ is a Gaussian random variable with inverse variance β and

$$y(\mathbf{x},\mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}), \tag{2}$$

that's to say a linear combination of potentially nonlinear basis functions $\phi(\mathbf{x})$ of the input variables where M is the total number of features in the model. The parameter w_0 is called the *bias* and is the intercept of the regression model. It is a constant which allows for fixed offset in the data commonly accompanied by $\phi_0(\mathbf{x}) = 1$ as to separate it from the other weights and make calculations easy.

Using this we can write the likelihood function as

$$\begin{aligned} p(t|\mathbf{x},\mathbf{w},\beta) &= \mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1}) \\ &= \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left\{-\frac{1}{2\beta^{-1}}(t - y(\mathbf{x},\mathbf{w}))^2\right\}. \end{aligned} \tag{3}$$

Now assuming we have a dataset of input variables $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ with corresponding target values $\mathbf{t} = \{t_1, ..., t_N\}$ and that all datapoints are drawn independently so that $P(A \cap B) = P(A)P(B)$ we can rewrite the likelihood function as

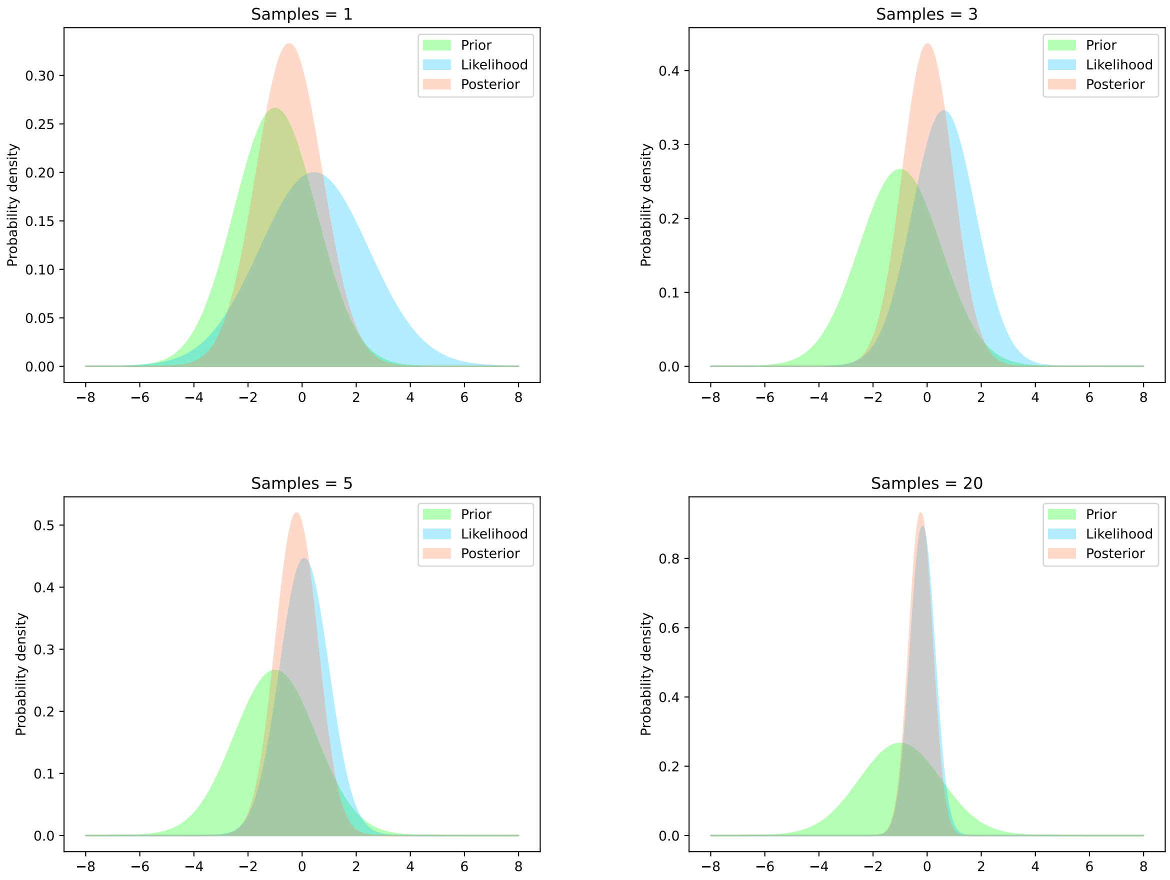

FIG. 1. Prior, likelihood and posterior distributions for samples drawn randomly from a normal distribution. The posterior becomes sharper and dominated by the likelihood function as more samples are introduced.

$$p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n),\beta^{-1}) \tag{4}$$

A. Maximum Likelihood Estimation (MLE) Maximum a Posteriori (MAP)

Maximum likelihood estimation, in short MLE, is a method based on maximizing the likelihood function with regards to the observed data

$$\mathbf{w}_0 = \text{arg max}_{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta), \tag{5}$$

where $p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)$ is the likelihood function. The solution, $\mathbf{w}_0$, is called the *best estimate* and is the solution which maximimizes the probability of the observed data. In other words it maximizes the probability of our model generating the observed data. We can find a MLE from maximizing the expression for the likelihood function expressed in eq.(4) and since

$$\text{arg max}_{\mathbf{w}} p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) = \text{arg max}_{\mathbf{w}} \ln\left[p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)\right],$$

we can maximize the natural logarithm of the likelihood function instead which is much easier. Taking the log yields

$$\begin{aligned}
&\ln\left[(\mathbf{t}|\mathbf{X},\mathbf{w},\beta)\right] \\
&= \ln\left[\prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n),\beta^{-1})\right] \\
&= \sum_{n=1}^{N} \ln\left[\mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n),\beta^{-1})\right] \\
&= \sum_{n=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi\beta^{-1}}}\exp\left\{-\frac{1}{2\beta^{-1}}(t_n-\mathbf{w}^T\phi(\mathbf{x}_n))^2\right\}\right] \\
&= \sum_{n=1}^{N} -\frac{1}{2}\ln\left[2\pi\beta^{-1}\right] - \sum_{n=1}^{N}\frac{1}{2\beta^{-1}}\left[t_n-\mathbf{w}^T\phi(\mathbf{x}_n)\right]^2 \\
&= \frac{N}{2}\ln\left[\beta\right] - \frac{N}{2}\ln\left[2\pi\right] - \beta E_D(\mathbf{w}),
\end{aligned}$$

where

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left[t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\right]^2, \tag{6}$$

is called the *sum-of-squares* error function. Now, maximizing with regards to $\mathbf{w}$ is equivalent to solving $\nabla_{\mathbf{w}} \ln p = 0$

$$\begin{aligned}
&\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) \\
&= -\beta\nabla[E_D(\mathbf{w})] \\
&= -\frac{\beta}{2}\nabla\left[\frac{1}{2}\sum_{n=1}^{N}\left[t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\right]^2\right] \\
&= -\beta\sum_{n=1}^{N}\left[t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\right]\phi^T(\mathbf{x}_n) = 0,
\end{aligned}$$

which rearranged yields

$$\begin{aligned}
\sum_{n=1}^{N} t_n\phi^T(\mathbf{x}_n) &= \mathbf{w}^T\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi^T(\mathbf{x}_n) \\
\implies \Phi^T\mathbf{t} &= \left[\Phi^T\Phi\right]\mathbf{w} \\
\implies \mathbf{w} \equiv \mathbf{w}_{\mathrm{ML}} &= \left[\Phi^T\Phi\right]\Phi^T\mathbf{t},
\end{aligned} \tag{7}$$

where

$$\Phi = \begin{bmatrix} \phi_0(x_1) & \cdots & \phi_{M-1}(x_1) \\ \vdots & \ddots & \vdots \\ \phi_0(x_N) & \cdots & \phi_{M-1}(x_N) \end{bmatrix} \tag{8}$$

is called the *design matrix* which contains all our basis functions evaluated for all input data. This maximum likelihood estimation is equivalent with the ordinary least squares (OLS) estimator (Bishop 2006, Ch.3.1.1). With this we can rewrite eq.6 as

$$E_D(\mathbf{w}) = \frac{1}{2}\|\mathbf{t} - \Phi\mathbf{w}\|^2. \tag{9}$$

This ties in with *maximum a posteriori*, or MAP, which only differs from MLE by making use of a prior. One can think of MLE as a special case of MAP where the prior is uniform or that MAP is equal to MLE multiplied by a prior

$$\mathbf{w}_0 = \arg\max{}_{\mathbf{w}} p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}), \tag{10}$$

where $p(\mathbf{w})$ is the prior probability distribution. Let us introduce a Gaussian prior distribution with mean $\mathbf{m}_0 = \mathbf{0}$ and covariance $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ over the weights

$$\begin{aligned}
p(\mathbf{w}|\alpha) &= \mathcal{N}\big(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}\big) \\
&= \left(\frac{\alpha}{2\pi}\right)^{M+1/2}\exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\},
\end{aligned} \tag{11}$$

where α is the precision of the distribution. From Bayes theorem we know that the posterior can be expressed in terms of the prior as

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha), \tag{12}$$

which using the likelihood function from eq.(4) and the Gaussian prior yields

$$\begin{aligned}
&p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) \\
&= \prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \\
&= \prod_{n=1}^{N}\frac{1}{\sqrt{2\pi\beta^{-1}}}\exp\left\{-\frac{1}{2\beta^{-1}}(t_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2\right\} \\
&\qquad\qquad \times \left(\frac{\alpha}{2\pi}\right)^{M+1/2}\exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\}.
\end{aligned}$$

This can in shorthand notation be written as

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N), \tag{13}$$

where

$$\begin{aligned}
\mathbf{m}_N &= \beta\mathbf{S}_N\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \\
\mathbf{S}_N^{-1} &= \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi},
\end{aligned} \tag{14}$$

are the mean and covariance of the posterior.

Maximizing this with regards to the weights will give us a MAP estimate and as with MLE it is easier done using the natural logarithm of the posterior

$$\begin{aligned}
\ln p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) = &\left[\frac{M+1}{2}\right]\ln\left[\frac{\alpha}{2\pi}\right] - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \\
&+ \frac{N}{2}\ln\left[\beta\right] - \frac{N}{2}\ln\left[2\pi\right] - \beta E_D(\mathbf{w}).
\end{aligned}$$

We know that maximizing the expression is the same as minmizing the negative of the expression which gives us

$$\begin{aligned}
&\nabla_{\mathbf{w}} \ln p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \alpha, \beta) \\
&= \alpha\mathbf{w} + \beta\left[\sum_{n=1}^{N} t_n\phi^T(\mathbf{x}_n) - \mathbf{w}^T\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi^T(\mathbf{x}_n)\right],
\end{aligned}$$

which rearranged yields

$$\begin{aligned} &\alpha\mathbf{w} + \beta\left[\Phi^T\Phi\right]\mathbf{w} = \beta\Phi^T\mathbf{t} \\ \implies &\left[\frac{\alpha}{\beta}\mathbf{I} + \left(\Phi^T\Phi\right)\right]\mathbf{w} = \Phi^T\mathbf{t} \\ &\implies \mathbf{w} \equiv \mathbf{w}_{\mathrm{MAP}} = \left[\lambda\mathbf{I} + \Phi^T\Phi\right]^{-1}\Phi^T\mathbf{t}. \end{aligned} \tag{15}$$

This is equivalent to the *Ridge regression* estimator with regularization term $\lambda = \alpha/\beta$ (Bishop 2006, Ch.3.1.4).

The objective of using such methods is to calculate the parameters of a model so to fit the model best to the observed data points. If the posterior probability distribution is symmetric about a maximum, then the *mean* equals the maximum which in turn is equal to the best estimate. The *mode* of a dataset is the value that appears the most or in other words the value that is most likely to get sampled. The *median* is the middle value of an ordered dataset and for normal distributions we have that the median = mode = mean.

V. THE PREDICTIVE DISTRIBUTION

The following section is based largely on (Bishop 2006, Ch.3.3.2).

The predictive distribution lets us make predictions of t from new input values, which is most often what we are interested in rather than the actual values of $\mathbf{w}$. To find it one needs to evaluate

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)\mathrm{d}\mathbf{w}, \tag{16}$$

in other word the probability of finding new target variables t given the input target variables $\mathbf{t}$ and the precision parameters. Inserting our expressions for the likelihood function in eq. 3 and the Gaussian posterior distribution in eq. 13 we can rewrite the right-hand side as

$$\begin{aligned} p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) &= \int \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)\mathrm{d}\mathbf{w} \\ &= \int \mathcal{N}(t|\mathbf{w}^T\phi(\mathbf{x}), \beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)\mathrm{d}\mathbf{w}. \end{aligned} \tag{17}$$

To solve this we rely on result 2.115 from *Pattern Recognition and Machine Learning* (Bishop 2006) which in short shows how one can find $p(t)$ from the marginal distribution $p(\mathbf{w})$ and the conditional distribution $p(t|\mathbf{w})$ if both have Gaussian shapes. This yields

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}\left(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})\right), \tag{18}$$

where

$$\sigma_N^2(\mathbf{x}) = \beta^{-1} + \phi(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}), \tag{19}$$

is the co-variance matrix of the predictive distribution. One can use the mean and co-variance to produce *Prediction intervals*. They are intervals predicted to be the interval in which future observables lie from a probability based on former observations. One can imagine them as actual intervals in data space containing the areas where future observations will lie. Practically they are defined as the mean of the predictive ± some multiple of standard deviation of the predictive, hence often denoted as $N\sigma$-prediction intervals. It is likely that this is not the most known estimate interval for a reader whom is new to bayesian statistics. A common estimate in frequentist statistics are *Confidence intervals*. These are defined as intervals where we are some percent certain that the true value of our parameter lies in or in other words a range of values related to some parameter of the data. So in a 95% confidence interval we will have 95% certainty that say the mean of some experimentally measured value will lie in that range. The bayesian equivalent is to use *credible intervals*. They are a range of values on the posterior which includes x% of the probability, usually at least 95%, for some parameter of interest. As with confidence intervals they express some percent of certainty concerning what the parameter we are looking at, again for example the mean of the observations, will be. A *highest posterior density interval* is a type of credible interval but more specifically the smallest interval one can use for some confidence level. Another type of credible interval is the *equally tailed* interval which an interval which ensures that the probability of our true parameter being over or under the interval is the same. At first glance it can seem that a prediction interval is a type of credible interval, but there is a difference. Both are made after seeing the data at hand, but a credible interval tells us about the probability of a parameter lying within an interval while a prediction interval tells us about the probability of a future observation lying within an interval.

VI. MODEL SELECTION & PARAMETER ESTIMATION

There are two main ways to ensure better fitting of data when dealing with bayesian linear regression, *model selection* and *parameter estimation*. Model selection is somewhat self explanatory, it is the problem of choosing the best basis functions and prior to serve as our model. Lets imagine a dataset which we want to fit, but that we have little prior knowledge about the data except that it is not generated from a linear function. An approach could be to both try Gaussian basis functions and polynomial basis functions of various degrees and look at the marginal likelihood score. Here is where the evidence enters the picture and why it is so important for model selection, as the marginal likelihood score is found by maximizing the evidence as we will show later on. One can from there compare the marginal likelihood score vs. degree of polynomial/number of Gaussian functions for the basis functions. If one model outperforms or performs just as well as the other with a lower complexity one should select the simpler one. This principle is called *Ockhams razor* and can in bayesian model selection be interpreted as "if all else is equal, pick the simpler model". That is to say if fewer parameters produce as well results as many, use fewer. It is the art of choosing the model that is just complex enough, but not more than it needs. So in summary, the purpose of model selection is to pick out the model (with a corresponding set of parameters) that fit the data best while avoiding it being more complex than need be. Whatever model we chose will consist of a set of parameters which we have to estimate so that the model fits the data as best as it can. This is parameter estimation. Sometimes it is even possible to redeem a model by doing parameter estimation if model selection fails to produce notably better results as more/higher quality data is introduced. Both MLE and MAP are methods for parameter estimation where one tries to find the optimal values for the weights.

A. Basis functions

We see it fit to include a small introduction to the types of basis functions we have previously mentioned which are often met when dealing with bayesian linear regression. *Polynomial* and *Gaussian* basis functions are two very common examples. Polynomial basis functions are defined as

$$\phi_i(x) = x^i, \tag{20}$$

and are the simplest form of basis function. It does not take into account other information than that of the input itself. Gaussian basis functions are defined as

$$\phi_i(x) = \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma^2}\right\}, \tag{21}$$

where μ, the mean, is sampled from a uniform distribution in the same range as the input variable so to define the location of the basis functions in input space while σ, the standard deviation, is a parameter governing the spatial scale of the basis function (Bishop 2006, Ch.3.1).

B. Marginal likelihood/Evidence

To determine values of the hyperparameters of a model one can maximize the *marginal likelihood function*, also called the evidence. The marginal likelihood function is obtained by integrating over the parameters $\mathbf{w}$

$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\mathrm{d}\mathbf{w}, \tag{22}$$

where α, β are the hyperparameters. Determining thses hyperparameters by maximizing the evidence is often denoted *empirical Bayes*. Another approach is to use a *hierarchical Bayesian model* which is defined by introducing a hyperprior as described in the section about *priors*. This is often considered a more "true Bayesian" approach as one performs Bayesian inference on the hyperparameters as well.

Using the likelihood function defined in equation 4 together with the Gaussian conjugate prior defined in equation 11 our evidence function takes the form

$$p(\mathbf{t}|\alpha,\beta) = \\ \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}\left(\frac{\alpha}{2\pi}\right)^{\frac{M}{2}}\int \exp\{-(\beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}))\}\mathrm{d}\mathbf{w}, \tag{23}$$

where N is the number of data points and M are the number of features in the design matrix. $E_D(\mathbf{w})$ is the sum-of-squares error function as defined in eq.9 and $E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$. We wish to rewrite the integral to something with a known analytical solution. Lets begin by expanding the expression

$$\begin{aligned}\beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) &= \\ \frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} &= \\ \frac{\beta}{2}\left[\mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T\Phi\mathbf{w} + \mathbf{w}^T\Phi^T\Phi\mathbf{w}\right] + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} &= \\ \frac{1}{2}\left[\beta\mathbf{t}^T\mathbf{t} - 2\beta\mathbf{t}^T\Phi\mathbf{w} + \mathbf{w}^T\left(\alpha\mathbf{I} + \beta\Phi^T\Phi\right)\mathbf{w}\right],&\end{aligned}$$

and defining $A = \alpha\mathbf{I} + \beta\Phi^T\Phi$ as well as keeping in mind that $A^{-1}A = \mathbf{I}$ so that $\mathbf{w} = \mathbf{w}\mathbf{I} = \mathbf{w}A^{-1}A$ we can rewrite the above expression as

$$\frac{1}{2}\left[\beta\mathbf{t}^T\mathbf{t} - 2\beta\mathbf{t}^T\Phi A^{-1}A\mathbf{w} + \mathbf{w}^TA\mathbf{w}\right].$$

We go on defining $\mathbf{m}_N = \beta A^{-1}\Phi^T\mathbf{t}$. Now we have to do perform some magic to progress, namely adding 0 to the equation as $0 = \mathbf{m}_N^TA\mathbf{m}_N - \mathbf{m}_N^TA\mathbf{m}_N$. This yields

$$\begin{aligned}&\frac{1}{2}\left[\beta\mathbf{t}^T\mathbf{t} - 2\beta\mathbf{t}^T\Phi A^{-1}A\mathbf{w} + \mathbf{w}^TA\mathbf{w} + \mathbf{m}_N^TA\mathbf{m}_N - \mathbf{m}_N^TA\mathbf{m}_N\right] \\ &= \frac{1}{2}\left[\beta\mathbf{t}^T\mathbf{t} - \mathbf{m}_N^TA\mathbf{m}_N + \mathbf{w}^TA\mathbf{w} - 2\mathbf{m}_N^TA\mathbf{w} + \mathbf{m}_N^TA\mathbf{m}_N\right] \\ &= \frac{1}{2}\left[\beta\mathbf{t}^T\mathbf{t} - \mathbf{m}_N^TA\mathbf{m}_N\right] + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^TA(\mathbf{w} - \mathbf{m}_N),\end{aligned}$$

where we in the last line have completed a square. We wish to keep the last term, but further rewrite the first. We begin by yet again sneaking in $0 = \mathbf{m}_N^TA\mathbf{m}_N - \mathbf{m}_N^TA\mathbf{m}_N$

$$\frac{1}{2}\big[\beta\mathbf{t}^T\mathbf{t} - \mathbf{m}_N^T A\mathbf{m}_N\big] = \\ \frac{1}{2}\big[\beta\mathbf{t}^T\mathbf{t} - 2\mathbf{m}_N^T A\mathbf{m}_N + \mathbf{m}_N^T A\mathbf{m}_N\big].$$

Inserting the expressions for $\mathbf{m}_N$ in the second term and A in the last term we get

$$\begin{aligned} &\frac{1}{2}\big[\beta\mathbf{t}^T\mathbf{t} - 2\mathbf{m}_N^T AA^{-1}\Phi^T\mathbf{t}\beta + \mathbf{m}_N^T\big(\alpha\mathbf{I} + \beta\Phi^T\Phi\big)\mathbf{m}_N\big] \\ &= \frac{1}{2}\big[\beta\mathbf{t}^T\mathbf{t} - 2\mathbf{m}_N^T\Phi^T\mathbf{t}\beta + \alpha\mathbf{m}_N^T\mathbf{m}_N + \beta\mathbf{m}_N^T\Phi^T\Phi\mathbf{m}_N\big] \\ &= \frac{1}{2}\big[\beta\big(\mathbf{t}^T\mathbf{t} - 2\mathbf{m}_N^T\Phi^T\mathbf{t} + \mathbf{m}_N^T\Phi^T\Phi\mathbf{m}_N\big) + \alpha\mathbf{m}_N^T\mathbf{m}_N\big] \\ &= \frac{1}{2}\Big[\beta(\mathbf{t} - \Phi\mathbf{m}_N)^T(\mathbf{t} - \Phi\mathbf{m}_N) + \alpha\mathbf{m}_N^T\mathbf{m}_N\Big] \\ &= \frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N. \end{aligned}$$

This gives us a final expression

$$\begin{aligned} &\beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) = \qquad (24) \\ &\frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{m}_N\|^2 + \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T A(\mathbf{w} - \mathbf{m}_N). \end{aligned}$$

Now on to evaluating the integral

$$\begin{aligned} &\int \exp\{-(\beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}))\}\mathrm{d}\mathbf{w} = \\ &\exp\bigg\{-\frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{m}_N\|^2 - \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N\bigg\} \\ &\times \int \exp\bigg\{-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T A(\mathbf{w} - \mathbf{m}_N)\bigg\}\mathrm{d}\mathbf{w}, \end{aligned}$$

we use our knowledge of Gaussian integrals

$$\int \exp\Big\{-\frac{\alpha}{2}\mathbf{x}^2\Big\}\mathrm{d}\mathbf{x} = \sqrt{\frac{2\pi}{\alpha}}^{M},$$

where M is the dimension of the vector $\mathbf{x}$ to write out our integral as

$$\begin{aligned} &\exp\bigg\{-\frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{m}_N\|^2 - \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N\bigg\} \\ &\times \int \exp\bigg\{-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T A(\mathbf{w} - \mathbf{m}_N)\bigg\}\mathrm{d}\mathbf{w} = \\ &\exp\bigg\{-\frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{m}_N\|^2 - \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N\bigg\}(2\pi)^{M/2}|A|^{-1/2}. \end{aligned} \tag{25}$$

We will once again make use of the fact that the maximum of the log of the evidence is the same as the maximum of the evidence itself, but easier to calculate. The log of the evidence can then be written as

$$\begin{aligned} \ln p(\mathbf{t}|\alpha,\beta) &= -\frac{\beta}{2}\|\mathbf{t} - \Phi\mathbf{m}_N\|^2 - \frac{\alpha}{2}\mathbf{m}_N^T\mathbf{m}_N \\ &+ \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{1}{2}\ln|A| - \frac{N}{2}\ln(2\pi),. \end{aligned} \tag{26}$$

which is our final expression for the marginal likelihood function.

VII. SEQUENTIAL UPDATING

In bayesian statistics if we assume that the input values are *exchangeable* (most often equivalent to the frequentist *independent identically distributed* variables) then for a set of parameters λ sequential updating is the same as bulk updating, as long as the posterior of the first update is used as the prior for the second update. More technically we have to be certain that x_1 and x_2 are conditionally independent given the model parameters λ so that

$$p(x_1, x_2|\lambda) = p(x_1|\lambda)p(x_2|\lambda) \tag{27}$$

Sequential updating has many benefits. As one can drop data after updating the posterior it is possible to deal with large datasets without being limited as much by storage issues. One can also update the model in real time giving predictions without having all the data. This is very useful concerning for example search engines which can try to predict what you are looking for right away, but has the chance to become better the more you search.

VIII. FINAL REMARKS

This paper has provided some of the necessary background needed for performing Bayesian Linear Regression. Our hope is that it can be of use for those new to the field when trying out Bayesian Linear Regression on their own datasets. The main take-aways from the paper is the understanding of how priors and likelihood functions work together through Bayes theorem to shed light on the data at hand in the form of a posterior, and how the roles they play differ from one problem to another. Another important piece of knowledge is how model selection and parameter estimation go hand in hand when trying to fit your data optimally and that both can be equally important for success.

[1] D. S. Sivia and J. Skilling, *Data Analysis - A Bayesian Tutorial*, 2nd ed., Oxford Science Publications (Oxford University Press, 2006).

[2] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).

[3] N. P. Lemoine, Oikos **128**, 912 (2019).