Bayesian Parameter Estimation

Scribe - Algorithms in Computational Biology

Timna Wharton Kleinman

16/03/23

Introduction

Parameter estimation is a fundamental task in statistics that involves estimating the values of unknown parameters in a statistical model based on observed data. There are several methods for parameter estimation including maximum likelihood estimation, Bayesian estimation, method of moments, and more. The accurate estimation of these parameters is critical for making reliable predictions, identifying patterns, and drawing meaningful conclusions from data. In our course, we used parameter estimation, among other things, for sequence alignment algorithms, Markov chains and HMM (particularly CPG islands), and more. This essay aims to provide an overview of parameter estimation and specifically Bayesian parameter estimation in contrast to maximum likelihood estimation and the Frequentist approach. The essay will begin by defining key terms and concepts, followed by introducing a general Bayesian estimator that relies on a selected loss function. It will explore various loss functions and their corresponding Bayesian estimators, before discussing the MLE and frequentist approach. Finally, a comparison between the two approaches will be made.

Definitions and Setting

Before beginning, one needs to define some useful terms - a **parametric model** or **parametric family of distributions** is a set of probability functions P of a random variable X, that depends (and differs from each other) only on a set of parameters θ. Meaning, each value of θ yields a probability function. There are a few common notions for such probability functions - $P_\theta(x)$, $P(x;\theta)$, $P(x \mid \theta)$ which differ depending on the chosen statistical paradigms (frequentist vs. Bayesian, see Comparison section). In this essay we will be using $P(x \mid \theta)$. Each probability function $P(x \mid \theta)$ needs to be a legal distribution (non-negative and sum to one). Since it depends only on θ, one needs a **parameter space** Θ, which corresponds to all values of θ that make P_θ a legal distribution. An example of a parametric family of distributions is the binomial distribution which depends on (1) n - the number of trials and (2) p - the probability to succeed in each Bernoulli trial. Hence, in this case $\theta = (p, n)$ and $\Theta = [0,1] \times \mathbb{N}$.
The setting of the problem is that we have n independent[1] and identically distributed (**iid**) samples or **training set** $D = \{x_1, ..., x_n\}$ drawn from a fixed distribution $P(\cdot \mid \theta)$. P is a parametric family of distributions, that depends on θ which is unknown. We want to estimate θ from the given samples. The

[1]Note that they are independent only for a fixed θ, meaning the tosses are **conditionally independent** given θ

estimator of θ will be marked as $\hat{\theta}$ and is a function of the given observations. Notice that $\hat{\theta}$ is a random variable and thus one can investigate its characteristics (distribution, expectation, etc.).
Another important thing to mention before diving in, is Bayes' theorem -

$$\overbrace{P(\theta \mid D)}^{\text{posterior}} \underset{\text{Rule}}{\overset{\text{Bayes}}{=}} \frac{\overbrace{P(D \mid \theta)}^{\text{likelihood}} \cdot \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(D)}_{\substack{\text{evidence / marginal likelihood}}}}$$

where the reasoning behind the names is that the prior represents the probability distribution of the parameters *before* observing the data, the likelihood represents the probability of observing the data *given* a particular set of parameter values, the posterior represents the probability distribution of the parameters *after* the fact (after observing the data), and the evidence is the probability of the *observations* (the evidence).
Our goal through the next few sections will be to describe different possible estimators - first from the Bayesian approach, and then from the frequentist approach.

Bayesian Approach

After laying all the settings, the question of "how would we choose an estimator?" arises, or "what function of D should we choose?". To answer this we define a **loss function** L (also known as a **cost function** or an **objective function)** which is used to measure the difference between the true and estimated parameters. Given a loss function, we would like to choose an estimator that would be "good". A reasonable possibility to determine what is "good" is to choose an estimator which minimizes the loss while taking into consideration all the possible values of θ, D. Formally,

$$\hat{\theta}^* = \arg\min_{\hat{\theta}} \mathbb{E}_{\theta,D}\left[L\left(\hat{\theta}(D), \theta\right)\right] = \arg\min_{\hat{\theta}} \sum_{\theta}\sum_{D} L\left(\hat{\theta}(D), \theta\right) \cdot P(D, \theta) \tag{1}$$

The outcome of this choice (equation 1) results in the Bayesian estimator[2]. As we will later discuss, the Bayesian approach uses prior knowledge and observed data to infer the probability distribution of the parameters (we can already see it since $P(D, \theta) = P(D \mid \theta) P(\theta)$).
Notice that by choosing $\hat{\theta}^*(D) = \arg\min_{\hat{\theta}} \mathbb{E}_{\theta|D}\left[L\left(\hat{\theta}(D), \theta\right)\right]$ for each given D, we receive the minimum of equation 1. This means that if one has a group of samples, and wants to use the "best estimator", there is no need to go though all the other possible sample groups to find the estimator, but it is enough to look at the current group only.

[2]This is often referred to also as Bayesian decision making and is directly linked to risk analysis.

Proof. Consider any estimator $\hat{\theta}$, then

$$\begin{aligned}
\mathbb{E}_{\theta,D}\left[L\left(\hat{\theta}^*(D),\theta\right)\right] &= \sum_{\theta}\sum_{D} L\left(\hat{\theta}^*(D),\theta\right)\cdot P(D,\theta) = \sum_{\theta}\sum_{D} L\left(\hat{\theta}^*(D),\theta\right)\cdot P(D)\cdot P(\theta\mid D) \\
&= \sum_{D} P(D)\sum_{\theta} L\left(\hat{\theta}^*(D),\theta\right)\cdot P(\theta\mid D) = \sum_{D} P(D)\cdot \mathbb{E}_{\theta|D}\left[L\left(\hat{\theta}^*(D),\theta\right)\right] \\
&\leq \sum_{D} P(D)\cdot \mathbb{E}_{\theta|D}\left[L\left(\hat{\theta}(D),\theta\right)\right] = \sum_{\theta}\sum_{D} L\left(\hat{\theta}(D),\theta\right)\cdot P(D,\theta) \\
&= \mathbb{E}_{\theta,D}\left[L\left(\hat{\theta}(D),\theta\right)\right]
\end{aligned}$$

■

Now, given some sample group we have a general Bayesian estimator $\hat{\theta}^*(D) = \arg\min_{\hat{\theta}} \mathbb{E}_{\theta|D}\left[L\left(\hat{\theta}(D),\theta\right)\right]$ that depends on the chosen loss function. We will attend different possible loss functions and investigate the outcome of choosing the estimator based on them. We will discover that they all rely on the posterior, which is in accordance (due to Bayes rule) with the Bayesian approach that relies on the prior:

L_1 Loss Function

This loss is defined by the L_1 norm, that is $L_1\left(\hat{\theta}(D),\theta\right) = \left\|\hat{\theta}(D)-\theta\right\|_1 = \sum_i \left|\hat{\theta}_i(D)-\theta_i\right|$. When choosing this loss function, we get $\hat{\theta}_{L_1}(D) = \text{median}P(\theta\mid D)$, the posterior median. In other words, if $\Theta=\mathbb{R}$ then $\int_{-\infty}^{\hat{\theta}_{L_1}} P(\theta\mid D)\,d\theta = \int_{\hat{\theta}_{L_1}}^{\infty} P(\theta\mid D)\,d\theta$ (see full proof in the Appendix).

L_2 Loss Function (Squared Error)

Here we will use the L_2 loss function which is defined $L_2\left(\hat{\theta}(D),\theta\right) = \left\|\hat{\theta}(D)-\theta\right\|_2 = \sqrt{\sum_i\left(\hat{\theta}_i(D)-\theta_i\right)^2}$. Notice that this loss amplifies bigger mistakes and is useful , among other things, since it is differentiable. By choosing this loss function, we get the following Bayesian estimator $\hat{\theta}_{L_2}(D) = \mathbb{E}_{\theta|D}[\theta]$ which is the posterior mean. Meaning, the expected value of θ given the samples.

Proof. As shown before, the optimal Bayesian estimator is

$$\begin{aligned}
\hat{\theta}_{L_2}(D) &= \arg\min_{\hat{\theta}} \mathbb{E}_{\theta|D}\left[L_2\left(\hat{\theta}(D),\theta\right)\right] = \arg\min_{\hat{\theta}} \sum_{\theta} P(\theta\mid D)\,L_2\left(\hat{\theta}(D),\theta\right) \\
&= \arg\min_{\hat{\theta}} \sum_{\theta} P(\theta\mid D)\left\|\hat{\theta}(D)-\theta\right\|_2
\end{aligned}$$

to find the minimum we will take the derivative w.r.t. $\hat{\theta}$

$$\sum_{\theta} P(\theta \mid D)\, 2\left(\hat{\theta}-\theta\right) = 2 \cdot \sum_{\theta} P(\theta \mid D)\left(\hat{\theta}-\theta\right) \stackrel{!}{=} 0$$

$$\sum_{\theta} \hat{\theta} \cdot P(\theta \mid D) = \sum_{\theta} \theta \cdot P(\theta \mid D)$$

$$\hat{\theta} \underbrace{\sum_{\theta} \cdot P(\theta \mid D)}_{=1} = \sum_{\theta} \theta \cdot P(\theta \mid D)$$

$$\hat{\theta} = \sum_{\theta} \theta \cdot P(\theta \mid D) = \mathbb{E}_{\theta|D}[\theta]$$

■

Kronecker's Delta Loss Function (MAP Estimator)

In this case, we would use an "everything or nothing" loss. Meaning we would have zero loss if $\hat{\theta} = \theta$, and a loss of one otherwise. Mathematically - $L\left(\hat{\theta}(D), \theta\right) = 1 - \delta_{\hat{\theta},\theta}$. By choosing this loss we get $\hat{\theta}_{\text{MAP}}(D) = \arg\max_{\theta} P(\theta \mid D)$, where MAP stands for Maximum A Posteriori. This estimation makes sense since we are choosing the estimator with the highest probability given the samples.

Proof. Notice that in the case of Kronecker's delta loss function

$$\mathbb{E}_{\theta|D}\left[L\left(\hat{\theta}(D), \theta\right)\right] = \mathbb{E}_{\theta|D}\left[1-\delta_{\hat{\theta},\theta}\right] = \sum_{\theta}\left(1-\delta_{\hat{\theta},\theta}\right) P(\theta \mid D) = \sum_{\theta \neq \hat{\theta}} P(\theta \mid D) = 1 - P\left(\hat{\theta} \mid D\right)$$

And so, the optimal Bayesian estimator is

$$\hat{\theta}_{MAP}(D) = \arg\min_{\hat{\theta}} \mathbb{E}_{\theta|D}\left[L\left(\hat{\theta}(D), \theta\right)\right] = \arg\min_{\hat{\theta}}\left\{1 - P\left(\hat{\theta} \mid D\right)\right\}$$

$$= \arg\min_{\hat{\theta}}\left\{-P\left(\hat{\theta} \mid D\right)\right\} = \arg\max_{\hat{\theta}}\left\{P\left(\hat{\theta} \mid D\right)\right\}$$

■

Summary

Choosing different loss functions resulted in different Bayesian estimators that depend on the posterior -

Loss Function	**Estimator**	**Meaning**
$L_1: \left\|\hat{\theta}(D) - \theta\right\|_1$	median$P(\theta \mid D)$	posterior median
$L_2: \left\|\hat{\theta}(D) - \theta\right\|_2$	$\mathbb{E}_{\theta|D}[\theta]$	posterior mean
$1 - \delta_{\hat{\theta},\theta}$	$\arg\max_{\theta} P(\theta \mid D)$	posterior mode

Relaying on the posterior in our estimator is reasonable, as it is connected to the question - "Given our samples, which θ is most probable?".

Frequentist Approach

MLE

Remember that our training set is defined - $D = \{x_1, ..., x_n\}$. Looking at how the MAP estimator behaves as the number of observations grows, one can notice that the significance of the prior (or log of the prior) goes down -

$$\hat{\theta}_{\text{MAP}}(D) = \arg\max_{\theta} P(\theta \mid D) \overset{\text{Bayes Rule}}{=} \arg\max_{\theta} \frac{P(D \mid \theta) \cdot P(\theta)}{\underbrace{P(D)}_{\text{constant}}} = \arg\max_{\theta} P(D \mid \theta) \cdot P(\theta)$$

$$\overset{\text{log is monotonic}}{=} \arg\max_{\theta} \{\log(P(D \mid \theta) \cdot P(\theta))\} = \arg\max_{\theta} \{\log P(D \mid \theta) + \log P(\theta)\} \tag{2}$$

$$= \arg\max_{\theta} \{\log P(\{x_1, ..., x_n\} \mid \theta) + \log P(\theta)\} \overset{\text{i.i.d}}{=} \arg\max_{\theta} \left\{\log \prod_{i=1}^{n} P(x_i \mid \theta) + \log P(\theta)\right\}$$

$$= \arg\max_{\theta} \left\{\sum_{i=1}^{n} \log P(x_i \mid \theta) + \log P(\theta)\right\}$$

This implies that the prior is somewhat negligible and one can consider maximizing only the first term[3]. So we define the maximum likelihood estimator (or MLE) to be

$$\hat{\theta}_{\text{ML}}(D) = \arg\max_{\theta} P(D \mid \theta)$$

Observe that this is a reasonable estimator since it is connected to the question - "From which θ are the samples most likely to have come?".
Another connection between the MLE and MAP is by assuming a uniform prior. In this case. we get

$$\hat{\theta}_{\text{MAP}}(D) = \arg\max_{\theta} P(\theta \mid D) = \arg\max_{\theta} P(D \mid \theta) \cdot \underbrace{P(\theta)}_{\substack{\text{constant} \\ (\text{same for all } \theta)}}$$

$$= \arg\max_{\theta} P(D \mid \theta) = \hat{\theta}_{\text{ML}}(D) \tag{3}$$

Overall, these two transitions from MLE to MAP (equations 2, 3) emphasize the difference between the Bayesian and the Frequentist approach. While the Bayesian relies on prior knowledge and updates it with

[3]Another reason to use MLE is that it minimizes the cross entropy and the KL-divergence with respect to the true model.

new data, the Frequentists focus on the frequency of observed data and the probability of observing that data.

General

In the Frequentist approach, θ is considered fixed (not a probabilistic entity) and unknown. The goal is to use collected data D to estimate θ. The frequentist perspective can be further divided into two distinct categories: classical and probabilistic, however, we will not provide a detailed explanation of their differences. The MLE method is a key tool in this approach, as it provides a way to estimate the unknown parameters based on the observed data.

Comparison

The philosophical differences between the Bayesian and frequentist approaches to statistics are rooted in their contrasting views of θ and their impact on parameter estimation. In the Bayesian approach, θ is considered a random variable, allowing for the calculation of $P(\theta)$ and $P(D \mid \theta)$. On the other hand, the classical or frequentist approach treats θ as a fixed constant with a single true value. The classical approach distinguishes between parameters, which cannot be repeatedly measured to determine their prevalence, and random variables, which can be investigated to determine their probability distribution. Thus, θ is regarded as a parameter of a distribution (not a random variable like in the Bayesian approach and unlike D), and the appropriate notation is $P(D;\theta)$ or $P_\theta(D)$.

This contrasting view of probability impacts parameter estimation of the different approaches. Bayesianism incorporates prior knowledge into the analysis, allowing for the derivation of posterior probability distributions for the parameters of interest. These distributions provide a range of plausible values for the parameter, and the most likely value can be selected (mean, median, or mode). This approach allows for the quantification of uncertainty in parameter estimates, as well as the incorporation of subjective beliefs and expert opinions. In contrast, frequentism utilizes large-sample theory and asymptotic properties of estimators. This method is often used to derive point estimates, such as the maximum likelihood estimate (MLE) or the method of moments estimate, that represent a single value of the parameter that is most consistent with the data.

The main criticism of Bayesian analysis is that its reliance on prior knowledge can be problematic, as the choice of prior distribution can significantly impact the resulting inferences and may be based on subjective or unreliable information. In addition, Bayesian analysis can be computationally complex, particularly for high-dimensional problems, and may be difficult to interpret and communicate to non-experts.

On the other hand, the frequentist approach relies solely on the data to make inferences and does not allow for the explicit incorporation of prior information or beliefs about the parameters of interest. This can be limiting in situations where prior knowledge or external information is available and can inform the analysis. Additionally, frequentist analysis is often criticized for its reliance on point estimates, which do not provide information about the uncertainty surrounding the parameter estimate. Confidence intervals can be used to estimate the range of plausible values for the parameters, but they do not provide a complete distribution of the parameter estimates or the posterior probability.

Frequentist analysis can also suffer from issues with computational complexity and model selection. Complex models with many parameters can be challenging to fit and require a large amount of data to achieve sufficient power. Model selection procedures can also be prone to overfitting and may not generalize well to new data.

Overall, both approaches have their strengths and weaknesses, and the choice between them depends on the availability of prior knowledge, the context of the problem, the nature of the data, and the researcher's goals and philosophical stance on probability and inference.

References

- Kevin P. Murphy. Machine Learning : A Probabilistic Perspective. MIT Press, 2012.
- Hanns L. Harney. Bayesian Inference Parameter Estimation and Decisions. Springer, 2003.
- Richard Durbin, , Sean R. Eddy, Anders Krogh, and Graeme Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, 1998.
- Daphne Koller, and Nir Friedman. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.

Appendix

L_1 Loss Function is the Posterior Median

Proof. We will prove that for the L_1 loss function - $L_1\left(\hat{\theta}\left(D\right),\theta\right)=\left\|\hat{\theta}\left(D\right)-\theta\right\|_1$ we get that the Bayesian estimator is - $\hat{\theta}_{L_1}\left(D\right)=\text{median}P\left(\theta\mid D\right)$. We will prove this for the case $\theta\in\mathbb{R},\Theta=\mathbb{R}$:

$$\mathbb{E}_{\theta|D}\left[L_1\left(\hat{\theta}\left(D\right),\theta\right)\right]=\int_{-\infty}^{\infty}L_1\left(\hat{\theta}\left(D\right),\theta\right)P\left(\theta\mid D\right)d\theta=\int_{-\infty}^{\infty}\left|\hat{\theta}\left(D\right)-\theta\right|P\left(\theta\mid D\right)d\theta$$

$$=\int_{-\infty}^{\hat{\theta}(D)}\left(\hat{\theta}\left(D\right)-\theta\right)P\left(\theta\mid D\right)d\theta+\int_{\hat{\theta}(D)}^{\infty}\left(\theta-\hat{\theta}\left(D\right)\right)P\left(\theta\mid D\right)d\theta$$

therefore

$$\hat{\theta}_{L_1}\left(D\right)=\arg\min_{\hat{\theta}}\mathbb{E}_{\theta|D}\left[L_1\left(\hat{\theta}\left(D\right),\theta\right)\right]=\arg\min_{\hat{\theta}}\left\{\int_{-\infty}^{\hat{\theta}(D)}\left(\hat{\theta}\left(D\right)-\theta\right)P\left(\theta\mid D\right)d\theta+\int_{\hat{\theta}(D)}^{\infty}\left(\theta-\hat{\theta}\left(D\right)\right)P\left(\theta\mid D\right)d\theta\right\}$$

Now, we will take the derivative w.r.t $\hat{\theta}$, and compare to zero. To do this, we will use the Leibniz integral rule on both parts of the integral and get -
$\frac{\partial}{\partial\hat{\theta}}\left(\int_{-\infty}^{\hat{\theta}(D)}\left(\hat{\theta}\left(D\right)-\theta\right)P\left(\theta\mid D\right)d\theta\right)=\left(\hat{\theta}\left(D\right)-\hat{\theta}\left(D\right)\right)P\left(\hat{\theta}\left(D\right)\mid D\right)d\theta+\int_{-\infty}^{\hat{\theta}(D)}P\left(\theta\mid D\right)d\theta=\int_{-\infty}^{\hat{\theta}(D)}P\left(\theta\mid D\right)d\theta$
$\frac{\partial}{\partial\hat{\theta}}\left(\int_{\hat{\theta}(D)}^{\infty}\left(\theta-\hat{\theta}\left(D\right)\right)P\left(\theta\mid D\right)d\theta\right)=-\left(\hat{\theta}\left(D\right)-\hat{\theta}\left(D\right)\right)P\left(\hat{\theta}\left(D\right)\mid D\right)d\theta-\int_{\hat{\theta}(D)}^{\infty}P\left(\theta\mid D\right)d\theta=\int_{\hat{\theta}(D)}^{\infty}P\left(\theta\mid D\right)d\theta$
And combining them we receive

$$(1)+(2)=\int_{-\infty}^{\hat{\theta}(D)}P\left(\theta\mid D\right)d\theta+\int_{\hat{\theta}(D)}^{\infty}P\left(\theta\mid D\right)d\theta\overset{!}{=}0$$

$$\int_{-\infty}^{\hat{\theta}(D)}P\left(\theta\mid D\right)d\theta=\int_{\hat{\theta}(D)}^{\infty}P\left(\theta\mid D\right)d\theta$$

This result means that $\hat{\theta}_{L_1}$ needs to hold $\int_{-\infty}^{\hat{\theta}_{L_1}}P\left(\theta\mid D\right)d\theta=\int_{\hat{\theta}_{L_1}}^{\infty}P\left(\theta\mid D\right)d\theta$ which is the definition of the posterior median as expected. ■