Elements Of Parameter Estimation

Shubha Shedthikere

Abstract

In this write up we discuss the concepts involved in the design of optimum procedures for parameter estimation. In particular, given a set of samples $X_1, \cdots, X_N$ from the family of distribution $\mathcal{P} = \{P_\theta; \theta \in \Omega\}$, we understand how to design the optimum procedure to estimate θ. A variety of design principles can be used depending on the available prior information and perfomance criteria chosen. We discuss two basic approaches to parameter estimation, one being the Bayesian, in which the parameter is assumed to be random quantity related statistically to the observation and a second in which the parameter is assumed to be unknown but without being endowed with any probabilistic structure.

1 Introduction

Statistics is concerned with the collection of data and with their analysis and interpretation. We shall not consider the problem of data collection in this book but shall take the data as given and ask what they have to tell us. The answer depends not only on the data, on what is being observed, but also on background knowledge of the situation; the latter is formalized in the assumptions with which the analysis is entered. There have, typically, been three principal lines of approach:

- **Data analysis:** Here, the data are analyzed on their ownterms, essentially without extraneous assumptions. The principal aim is the organization and summarization of the data in ways that bring out their main features and clarify their underlying structure.
- **Classical inference and decision theory:** The observations are now postulated to be the values taken on by random variables which are assumed to follow a joint probability distribution, P, belonging to some known class$\mathcal{P}$. Frequently, the distributions are indexed by a parameter, say θ (not necessarily real-valued), taking values in a set, so that

$$\mathcal{P} = \{P_\theta, \theta \in \Omega\} \tag{1}$$

  The aim of the analysis is then to specify a plausible value for θ (this is the problem of point estimation), or at least to determine a subset of of which we can plausibly assert that it does, or does not, contain θ (estimation by confidence sets or hypothesis testing). Such a statement about θ can be viewed as a summary of the information provided by the data and may be used as a guide to action.

- **Bayesian analysis:** In this approach, it is assumed in addition that θ is itself a random variable (though unobservable) with a known distribution. This prior distribution (specified according to the problem) is modified in light of the data to determine a posterior distribution (the conditional distribution of θ given the data), which summarizes what can be said about θ on the basis of the assumptions made and the data

In terms of the model (1) , suppose that g is a real-valued function defined over Ω and that we would like to know the value of $g(\theta)$ (which may, of course, be θ itself). Unfortunately, θ, and hence $g(\theta)$, is unknown. However, the data can be used to obtain an estimate of $g(\theta)$, a value that one hopes will be close to $g(\theta)$.

An estimation problem involves two basic ingredients:

1. A real-valued function g defined over a parameter space Ω, whose value at θ is to be estimated; we shall call $g(\theta)$ the *estimand*.

2. A *random observable* X (typically vector-valued) taking on values in a sample space $\mathcal{X}$ according to a distribution P , which is known to belong to a family $\mathcal{P}$ as stated in (1).

The problem is the determination of a suitable *estimator*.
Definition (Estimator): An estimator is a real-valued function δ defined over the sample space. It is used to estimate an estimand, $g(\theta)$, a real-valued function of the parameter.

Of course, it is hoped that $\delta(X)$ will tend to be close to the unknown $g(\theta)$, but such a requirement is not part of the formal definition of an estimator. The value $\delta(x)$ taken on by $\delta(X)$ for the observed value x of X is the estimate of $g(\theta)$, which will be our "educated guess" for the unknown value.

The estimator δ is to be close to $g(\theta)$, and since $\delta(X)$ is a random variable, we shall interpret this to mean that it will be close on the average. To make this requirement precise, it is necessary to specify a measure of the average closeness of (or distance from) an estimator to $g(\theta)$. Examples of such measures are

$$P(|\delta(X) - g(\theta)| < c) \qquad \text{for some } c > 0 \tag{2}$$

and

$$E|\delta(X) - g(\theta)|^p \qquad \text{for some } p > 0. \tag{3}$$

Quite generally, suppose that the consequences of estimating $g(\theta)$ by a value d are measured by $L(\theta, d)$. Of the loss function L, we shall assume that

$$L(\theta, d) \geq 0 \qquad \text{for all } \theta, d \tag{4}$$

and

$$L(\theta, g(\theta)) = 0 \qquad \text{for all } \theta, \tag{5}$$

so that the loss is zero when the correct value is estimated. The accuracy, or rather inaccuracy, of an estimator δ is then measured by the **risk function**

$$R(\theta, \delta) = E_\theta\{L[\theta, \delta(X)]\}, \tag{6}$$

the long-term average loss resulting from the use of δ. This is essentially the expected value taken using P_θ. This the average loss of using δ when actually the data is coming from distribution P_θ. One would like to find a δ which minimizes the risk for all values of θ.

As stated, this problem has no solution. For, by (5), it is possible to reduce the risk at any given point θ_0 to zero by making $\delta(x)$ equal to $g(\theta_0)$ for all x. There thus exists no uniformly best estimator, that is, no estimator which simultaneously minimizes the risk for all values of θ, except in the trivial case that $g(\theta)$ is constant.

One way of avoiding this difficulty is to restrict the class of estimators by ruling out estimators that too strongly favor one or more values of θ at the cost of neglecting other possible values. This can be achieved by requiring the estimator to satisfy some condition which enforces a certain degree of impartiality. One such condition requires that the bias $E_\theta[\delta(X)]g(\theta)$, sometimes called the systematic error, of the estimator δ be zero, that is, that

$$E_\theta[\delta(X)] = g(\theta) \qquad \text{for all } \theta \in \Omega \tag{7}$$

This condition of unbiasedness ensures that, in the long run, the amounts by which δ over- and underestimates $g(\theta)$ will balance, so that the estimated value will be correct "on the average." A somewhat similar condition is obtained by considering not the amount but only the frequency of over- and underestimation. This leads to the condition

$$P_\theta[\delta(X) < g(\theta)] = P_\theta[\delta(X) > g(\theta)] \tag{8}$$

or slightly more generally to the requirement that $g(\theta)$ be a median of $\delta(X)$ for all values of θ. To distinguish it from this condition of median-unbiasedness,(7) is called mean-unbiasedness if there is a possibility of confusion.

2 Statistics, Sufficiency and Completeness

3 Information Inequality

4 Average Risk Optimality

The prior distribution is typically selected from a flexible family of prior densities indexed by one or more parameters. We shall denote the density by $\pi(\theta|\gamma)$, where the parameter γ can be real or vector-valued.

We can then write a Bayes model in a general form as:

$$X|\theta \sim f(x|\theta) \tag{9}$$
$$\Theta|\gamma \sim \pi(\theta|\gamma) \tag{10}$$

Thus, conditionall on θ, X has the sampling density $f(x|\theta)$, and conditionally on γ, Θ has prior density $\pi(\theta|\gamma)$. From this model, we calculate the posterior distribution, $\pi(\theta, \gamma)$, from which all Bayesian answers would come. The exact manner in which we deal with the parameter γ or more generally, the prior distribution $\pi(\theta|\gamma)$ will lead us to different types of Bayes analyses.

1. **Single Bayes** In this model, we assume the functional form of the prior and the value of γ is known, so we have one completely specified prior, that is $\gamma = \gamma_0$. Thus the model is,

$$X|\theta \sim f(x|\theta) \tag{11}$$
$$\Theta|\gamma \sim \pi(\theta|\gamma_0) \qquad \text{with } \gamma_0 \text{ known} \tag{12}$$

   Given a loss function $L(\theta, d)$ we then look for the estimatior that minimizes

2. **Hierarchical Bayes**

3. **Empirical Bayes**

5 Single-Bayes

6 Hierarchical Bayes

7 Empirical Bayes

7.1 Estimating parameters through EM algorithm

References

[1] E.L. Lehmann, George Casella, "Theory of Point Estimation"