

Applied Time-Series Analysis
Prof. Arun. K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 108
Lecture 47A - MLE and Bayesian Estimation 3

Very good morning, so let us get a going on bayesian estimation, I explain to you yesterday the essential principle of bayesian estimation you start with the prior uncertainty on the parameter and again to remind you in the bayesian frame work, we assume that the truth is not fixed as in the classical methods, but then of course, the truth is assume to be random, but the way to perceive it is that it is not the truth per say that is random, it is our knowledge of the truth that is kind of a uncertain and we essentially mask we are looking at the truth plus our uncertainty together and the hope is that the data will refined that uncertainty and get us a more precise estimate.

So, therefore, we have a concept of prior and posterior and the main or the central task in bayesian estimation is in the computation of this posterior given this prior.

(Refer Slide Time: 01:19)

ML and Bayesian estimation - Introduction

The posterior p.d.f.

Given an observation set y_N (evidence), the denominator is a fixed quantity, allowing us to write

$$f(\theta|y_N) = C \frac{f(y_N|\theta)f(\theta)}{\quad} \quad (21)$$

(Posterior) (Likelihood)(Prior)

The constant C is usually adjusted such that $f(\theta|y)$ is a legitimate p.d.f.

1 2 3 4 5 6 7 8 9 10 11 12

Arun K. Tangirala Applied TSA November 8, 2018

And of course, the likelihood you can say or the conditional observations of the measurements or parameterized p.d.f conditional p.d.f of the observations given the parameters, which you have to assume. So, the user has to supply both f of θ and f of

y_N given θ and C is a constant stemming from the fact that, we had this f of Y_N here and we said that f of Y_N is independent of the θ .

(Refer Slide Time: 01:45)

ML and Bayesian estimation - Introduction

Basic Idea

Prior to the experiment we have a (large) uncertainty in parameters, and post experiment (and data analysis), this uncertainty should (hopefully) shrink.

The starting point is therefore, the conditional p.d.f., written using Bayes' rule:

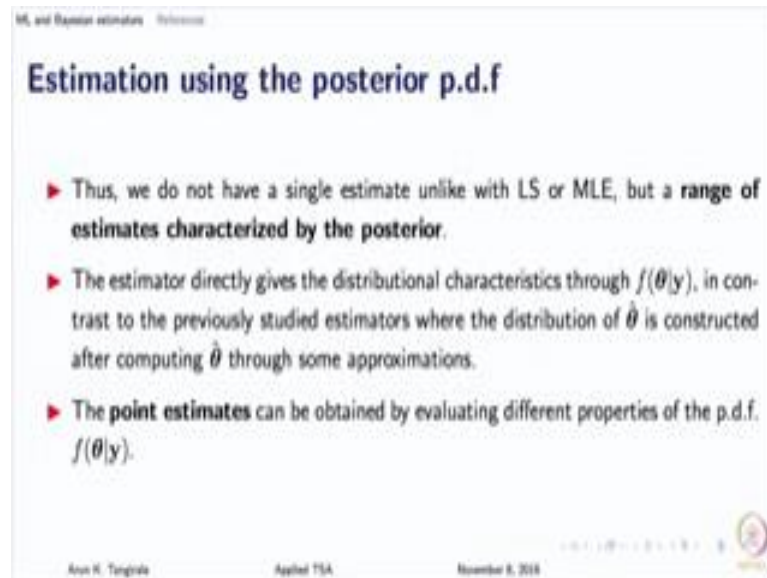
$$f(\theta|y_N)f(y_N) = f(y_N|\theta)f(\theta)$$
$$\implies f(\theta|y_N) = \frac{f(y_N|\theta)f(\theta)}{f(y_N)} \quad (20)$$

Anil K. Tongolo Applied TSA November 6, 2018 20

Nevertheless it is a constant that has to be estimated if you cannot ignore it, but the nice thing is you do not have to evaluate f of Y_N , you treat it as a constant and evaluate the constant in such a way that this left hand side is the legitimate p.d.f, what we mean by legitimate p.d.f is a area under it should be one.

So, that is how you estimate the constant that is it. So, beyond this as I said it is the users call as to what the user would like to do with posterior, but the main point again apart from the fact that, there is a start difference between the philosophies of the classical methods and the bayesian methods, the second significant difference is we obtain interval estimates, we do not have obtain point estimate; however, we can derive point estimate out of this distribution of θ in a few different ways, I will go through to that.

(Refer Slide Time: 02:53)



ML and Bayesian estimation

Estimation using the posterior p.d.f

- ▶ Thus, we do not have a single estimate unlike with LS or MLE, but a **range of estimates characterized by the posterior**.
- ▶ The estimator directly gives the distributional characteristics through $f(\theta|y)$, in contrast to the previously studied estimators where the distribution of $\hat{\theta}$ is constructed after computing $\hat{\theta}$ through some approximations.
- ▶ The **point estimates** can be obtained by evaluating different properties of the p.d.f. $f(\theta|y)$.

Amr H. Tawfik Applied TSA November 6, 2018

For example, I could say that I would like to pick the average of the estimates that I have obtain; f of theta given by is the p.d.f of conditional p.d.f of theta given by after the experiment has been perform, I can ask for the average which turns out to be the conditional expectation.

(Refer Slide Time: 03:11)



Mean (cond) $E(\theta|y)$ MMSE
Median (cond)
Mode (-do-)

So, the first point estimate that I construct is the mean. In fact, the conditional mean specifically of theta, in other words expectation of theta given Y which is nothing, but the average of f of theta given y . Now that we have decided to work with some number

an average of the interval of estimates that we have obtain, we can turn to other averages as well; what do you mean by other average? This is we can look at median of that is condition median again or we can look at mode once again conditional.

So, we can look at any of this we can work with the any of this estimates and derive a point estimate, which is completely the flipped way of what we have been doing in the classical methods. So, whatever point estimates that you pick fortunately now, you can show that picking any of this point estimates amounts to solving an optimization problem. So, here we are not solving any optimization problem, you must be you must have observed by now, we are not maximizing likelihood we are not minimizing approximation errors nothing. We are just saying I will compute the posterior and from the posterior I am going to derived this point one of this estimates and the nice result is regardless of which of this point estimates you work with, you would essentially be solving an optimization problem at behind a since, where you would be minimizing what is known as risk function or a cost function.

(Refer Slide Time: 05:00)

Point estimates and their optimality

► Further, these point estimates are optimal in the sense that they minimize a risk function $\mathcal{R}(\epsilon_\theta) = E(C(\epsilon_\theta))$, which is the averaged user-defined cost function.

$$\hat{\theta} = \arg \min_{\theta} \mathcal{R}(\epsilon) = \arg \min_{\theta} E(C(\epsilon)_{\theta}) \quad (22)$$

Anon N. Tangirala Applied TSA November 6, 2018

So, let us quickly go through this 3 estimates different point estimates and ask what is the equivalent optimization problem I would be solving, I would still not be solving that optimization problem, but it is heartening to know that these estimates respective estimates are optimal in some sense.

(Refer Slide Time: 05:12)

ML and Bayesian estimation

Bayesian estimate

Three popular choice of cost functions associated with three related properties of the p.d.f. are discussed below:

1. **Bayesian estimate, $E(\theta|y)$** : This is the estimate that minimizes the cost function

$$C(\epsilon_\theta) = \|\theta - \theta_0\|_2^2 \quad (23)$$

In other words, it is the MMSE of θ and also the mean of $f(\theta|y)$.

$$\hat{\theta}_{\text{BA}} = E(\theta|y) = \arg \min_{\theta} E((\theta - \theta_0)^2) \quad (24)$$

Amr H. Tawfik Applied TSA November 6, 2018

So, the first in order is this conditional expectation, which is known as the bayesian estimate. Although the general bayesian estimation when we talk about it, it always refers to computation of the posterior, specifically when you say I am working with the bayesian estimate, you are working with this conditional expectation which also happens to be the minimum square mean square error estimate; that is this minimizes the sum square error between the estimate and the truth.

If we are estimate is theta hat then you would be essentially this is a solution minimizing, find theta hats as the this is minimized and that is the condition expectation you know that already right and that is exactly what your equation is telling you, equation 23 telling you the cost function that you are minimizing with the slight change of notation. But you should understand that by working with this conditional expectation, you are essentially solving an optimization problem of this tipper, where you are minimizing the mean square error.

So, this is very nice because early on in estimation we said I would not know the truth, so therefore, I can never solve this problem, but here is a solution to that problem without knowing theta naught, we are actually able to obtain the minimum mean square error estimate and that is why the bayesian estimation is or the bayesian estimate is very good.

The celebrated Kalman filter is actually a Bayesian estimator, you can show that it is a minimum mean square error estimator and of course, we have already known we have already know what is this why is this result optimal as we have been saying the conditional expectation is the minimum mean square error estimator and how would you compute this from the posterior using this expression? Of course in practice you may not have an analytical expression for f or you may have that depends on the situation, it completely depends on the situation. In early days of Bayesian estimation may be or the last two decades or so, the prior and posterior sorry the prior and the likelihood, but chosen in such a way that you would get an analytical expression for the posterior.

So, let me give you an example, suppose I assume that the measurements have Gaussian distributed errors condition on for the fixed value of θ , f of y given θ is Gaussian like we went through the example yesterday; then what prior should I choose so that I can obtain a closed form expression for the posterior? The likelihood will have a closed form expression, I can write an analytical expression for the joint Gaussian p.d.f, but there is no guaranty that when I multiply 2 p.d.f.s. I can show that the resulting p.d.f has some closed form expression, simply because as multiply 2 p.d.f.s it does not mean that I will always be able to express that the resulting product in the form of a known p.d.f. When the priors are chosen such that the posterior p.d.f could be expressed in a known p.d.f form, those priors are called conjugate priors.

So, that was the basis for choosing the prior, it seems a bit strange because we said prior is suppose to represent my knowledge prior knowledge of θ and now you are saying prior is driven by computation consideration, yes that is true. When you are following short of computation power, you will have to make some compromises and that compromise was made here then that question that arises in mind is what if the prior that is governed by computation convenience is not the same as the prior that I would like that I have a belief about θ , what if they 2 or different, will the estimate go (Refer Time: 09:36) will estimate be in error? No, actually in general in Bayesian estimation even if you were to choose a prior that is different from the so called true prior.

You would still end up obtaining decent estimates, but the number of observations that you have to collect to obtain a certain precision is going to change, what is the premise behind Bayesian estimation? You start with some uncertainty before the experiment,

perform the experiment analyze your data so that you obtain a more precise estimate, that is lesser uncertainty.

Suppose I were to fix the uncertainty, I were to say that I would like to have an estimate with the certain precision and I start with prior 1 which is the truth and prior 2 which is different from the true prior, then would it make a difference to the precision that I want? No. I can still achieve the precision that I want, but where it makes the difference number of observation that will require, that would be require to achieve the desire precision. When I show the example part of this can be understood, the other part may be you have to simulate and figure out, but that is a point because this is a question that is of an asked when by beginners in bayesian estimation, what if I choose a wrong prior a nothing like a wrong prior, we will take more observations to achieve a certain precision.

So, the to go back to the choice of prior essentially you would choose the prior in early days, even to a certain extent today such that the product of this prior and the likelihood can be expressed in a closed form in some known form of p.d.f. Today you do not have the restriction; because of the computational power. Essentially what you now have is very high simulation power to over the last few years I would say probably less an a decade, we have now what are known as a Markov chain Monte Carlo simulators MCMC simulations, which allow us to work with p.d.fs that do not have a closed form expression, after all why do I need to know the posterior I would like to be able to compute this conditional expectation or the median or the mode, I should have the p.d.f either in a mathematical form or some values.

So, what this Markov chain Monte Carlo simulators would do is, they would actually run for different realizations of y , they would actually figure out what is the resulting theta the possibilities for theta or for a given y you can say for different priors you can figure out what the posterior is. So, it is would sample the space and that sampling the outcomes space has to be done in a computationally efficient manner and that is the key in the Markov chain Monte Carlo simulators. So, in other words today this p.d.f this posterior is need not be available in a close form expression, it can be available numerical form. It essentially gives you the probabilities on small intervals in the theta space and using those numerical p.d.fs, you can compute the conditional expectation. So, you replace the integral with the numerical integration.

(Refer Slide Time: 13:11)

Bayesian estimate ... contd.

The support comes from the classical result in prediction theory, which states that given a random variable Y , the best prediction of an unknown X is its conditional expectation in the minimum MSE sense.

The estimate is computed using the definition of conditional expectation (21),

$$E(\theta|y) = \int \theta f(\theta|y) d\theta \quad (25)$$

Arav K. Tongala Applied TSA November 6, 2018

And go ahead and compute of course, there going to be more intensive then you are regular mle or even least squares, but that is today we have we can offered large computational power and you must have a observed in many mathematical feels today people are now slowly resorting to proving certain theorems by way of enlisting all enumerating all possibilities. So, it is like a brute force method which was considered bad earlier, now is being considered good bit it was the brute force method that is a doing an a exhaustive search for all possibilities if I am proving some theorem, was considered bad the pen and paper version was considered anlagen it is still anlagen, but it was considered bad because primarily because we did not have computational power to do an exhaustive search.

So, there was a reason formats conjectures, I do not know if you must have heard that was actually not it does not possible to prove at least until such a time, analytically or mathematically whether that conjecture is correct or not; where as they did an exhaustive search. But using clever computational clever search methods and showed that the conjecture is wrong, but it required quite a bit of intelligence to do that exhaustive search, one small inefficiency in that intellectual in that search would really blow up the time takes or the time taken to figure out whether that conjecture is good or not, hope eventually approved that conjecture does not hold.

So, here in Markov chain Monte Carlo simulations some intelligent kind of sampling is used and for those of your familiar with Kalman filters you would also heard of particle filters and so on. So, this particle filters also rely on this sampling in the outcomes space approach at clever sampling of the outcomes space. So, to come back to the discussion to compute this conditional like expectation, either you need f in analytical form closed form and still may be you will not be able to evaluate this integral, you would perform a numerical integration, but today you have the luxury of working out this integral again in numerical form even though you may not have f in a analytical form.

(Refer Slide Time: 15:33)

MAP estimate

2. MAP estimate: The associated cost function is

$$C(\epsilon_{\theta}) = \begin{cases} 0, & |\epsilon_{\theta}| < \delta \\ 1, & |\epsilon_{\theta}| > \delta \end{cases} \quad (26)$$

This leads to the maximum a posteriori estimate, which is essentially the mode of $f(\theta|y)$:

$$\hat{\theta}_{MAP} = \text{arg max}_{\theta} f(\theta|y) = \text{arg max}_{\theta} f(y|\theta)f(\theta) \quad (27)$$

$$\text{or } \hat{\theta}_{MAP} = \text{arg max}_{\theta} (\ln f(y|\theta) + \ln f(\theta)) \quad (28)$$

The estimate is also known as the hit-or-miss estimate.

Anu K. Tingle Applied TSA November 8, 2018 41

The second estimate has got to do with the mode although we written here median in the order, the second popular estimate the first one is called the bayesian estimate which is the conditional expectation. The second one is called the MAP. MAP stands for maximum aposterior estimate; what we are saying is we will if you know from the theory of from the probability theory.

(Refer Slide Time: 16:00)

an (cond) $E(\theta|y)$ MLE $E - \min(\theta - \theta)$
dian (cond)
de (-do-) $\max_{\theta} f(\theta|y)$
 $f(\theta|y) = f(y|\theta)f(\theta)$
 $\ln(\text{Post}) = \ln(\text{Likelihood}) + \ln(\text{Prior})$

Mode is nothing, but the maximum of the p.d.f. So, in this case it is a maximum of the posterior, we know that mode is essentially the value of the random variable at which the p.d.f peaks and that is why it is called the maximum a posterior estimate. And you can show that once again this map minimizes some kind of a cost function and that cost function is given here, that is the cost function is again in terms of the error, error between theta and theta naught. So, if you look at the cost function it is a kind of a discontinues kind of cost function, it basically sorry it is minimizing this cost function and it is called the hit or miss estimate; that means, either you are in the correct region or you are not with this kind of an estimate, but the beauty with this map is it is very close similarities to MLE.

So, let me go back to this expression, the original expressions where a where you have posterior as the function of likelihood times a prior, if you look at this expression carefully and I write this in logarithmic form like we have done there. So, log of posterior is logarithm of likelihood plus logarithm of prior right.

So, if you look at this carefully, the first point to observe is suppose I assume f of theta to be constant, where when is the p.d.f constant for what kind of distribution is the p.d.f constant uniform? Which means I am saying any value of theta is likely, I have no specific knowledge of which are the more likely values and less likely values I do not know average temperature anything between 0 Kelvin and may be 6000 Kelvin or

whatever is the temperature of the sun that is possible for this average room temperature. It ridicules, but you can in principle say that and still get away with that. So, let us say we say that any value is equally likely then basically this boils down to a constant yeah there is a constant here. So, let me write that constant as well.

So, this is observed into this constant. So, what you are left with this simply the posterior being the likelihood scaled likelihood and when you are evaluating the mode of this, all your doing is picking up the MLE. So, the MLE therefore, as much as to Fisher dislike is the Bayesian estimate with the uniform prior that is all, it is a Bayesian estimator with uniform prior. If I assume any values equally likely and perform a Bayesian estimate Bayesian I carry out this and pick the map. So, that is very important. So, the complete statement is MLE is the MAP estimate with the uniform prior, not simply Bayesian estimation, Bayesian estimation would be conditional expectation, it is a maximum a posteriori estimate with the uniform prior.

Now, we know we have a different perspective of MLE, we have been thinking all along, we have been viewing MLE all along as an estimate that maximizes the likelihood, but now we can take a different stamp and we can say well now it is actually an estimate that is the maximum of the posterior, after I have conducted the experiment among all possibilities I have picked up the maximum, that maximizes the f of θ given y , if I begin with the uniform prior. This is the perspective that Fisher obviously, vehemently argued against because the framework in which MLE is set up is θ is deterministic, the truth θ is deterministic whereas in the Bayesian framework the truth is random. So, the philosophy is themselves are different, how can you by sheer mathematical equivalence observe my approach into the Bayesian.

So, as I said fine people said that fine how long will you stay to argue with all due respect to Fisher is it fine, no problem a few papers from you that is it right we will wait for some time and then start teaching that MLE is a special case of Bayesian.

So, I hope wherever Fisher is looking at from the sky's is not upset with me and I keep saying this, but anyway you should still nevertheless keep that settled, but very important difference in mind while mathematically accepting the equivalence, alright.