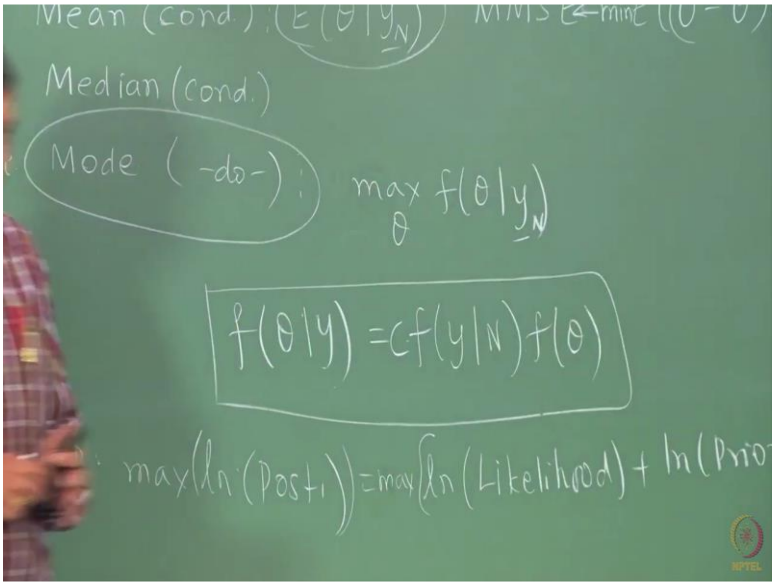
Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture – 109
Lecture 47B - MLE and Bayesian Estimation 4

Now, what happens if the prior is not uniform? So, you have to ignore this, this is not going to anyway play any part in the optimization; what I am going to pick in map is I am going to pick the maximum of this, whether I pick the maximum of p.d.f or maximum of logarithm of p.d.f it does not make a difference right, we have talked about in MLE as well.

(Refer Slide Time: 00:26)



So, I am going to pick here the maximum, when I am going to when I work with map essentially I am picking maximum of this objective function or minimum of negative log likelihood, plus negative of that log prior.

Now, it turns out if you turn to optimization, I do not know how many of you are familiar with the a parameter estimation literature, when you are estimating parameters we have talked about over parameterization and one of the symptom of over parameterization is large errors, if you take least squares for example, which is more or less MLE under some conditions, in the least square subjective function or even in the MLE there is no explicit expression that is telling the optimizer, if the to minimize the number of parameters that is being used; there is nothing, both MLE and least squares are essentially telling the optimizer find theta such that the approximation errors are minimized that is all.

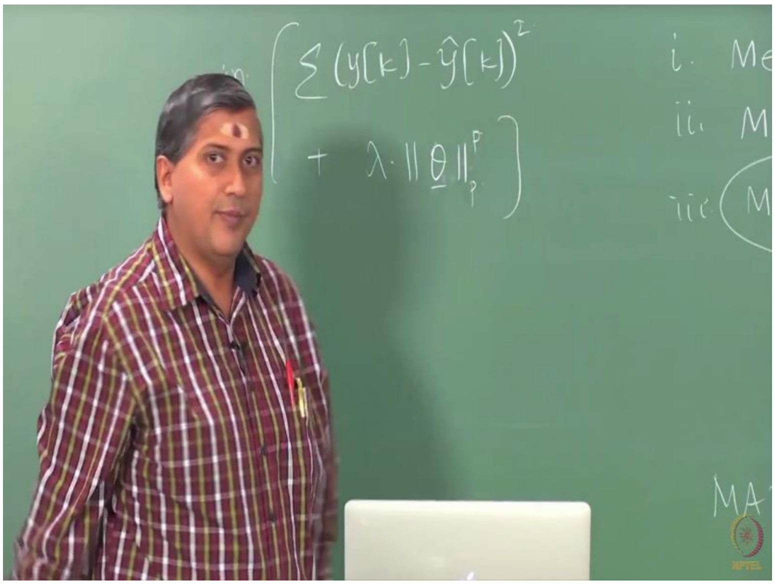
So, the focus is always on minimizing the prediction error or minimizing the approximation error. So, it is the more theta you give me I will use them, because you want you want me to give a minimum right it is a users responsibility therefore, post estimation to figure out whether more thetas have been used, that is more than necessary has been used and the way we figure that out? Is by looking at the errors in the estimates and akaike information criteria like measures try to achieve a tradeoff, they give you a tradeoff between how much good fit you have obtained versus how much error you have incurred in a parametric estimates and we know that at some point there is a good healthy tradeoff.

Gradually people said why not I tell the let the optimizer figure out, why should I as a user be poking my nose into this right? I would like to have minimal user intervention, never think of full automated, fully automated would mean you will become jobless. So, never ever propose to your boss of a fully automated one and it is not possible, it is not a great idea either there should be some user intervention, but you want to minimize it. In other words you want the optimizer to automatically tell you that although you have include you have given me 100 parameter model, I will use only 30 of that or 20 of that or even 3 of that and that is when the idea of regularization came along.

So, as I always say your parameter is your manpower, there are there to do some job for you that job is fitting the data. The more man people you hire yes you will get more and more work done, but then the amount of budget whether it is cost or food, money or food that you have is limited, that is the information content in the data. The more people you invite the more work you will get done, but the less you will pay them and they will go dissatisfied that is what we mean by large errors in parameters estimates.

So, if you regularize things then you will although you 100 people have been sent to you by some human resources organization, you will be very careful if you are you know sensible person you will be clever in not using all of them, not employing all of them you will say I need only 20 the rest can sleep, in IT field they say they are on bench, but these parameters will be on bench sleeping also that is what you want. So, this idea of penalizing including a penalty function in the objective function, so that the optimizer knows there is a cost associated with every parameter is called regularization.

(Refer Slide Time: 04:34)


$\sum (y[k] - \hat{y}[k])^2$
$+ \lambda \cdot \|\underline{\theta}\|_p^p$
i. Me
ii. M
iii. M
MA

And typically so if you take least squares, you have here y k minus y hat of k whole square, this is your vanilla least squares without regularization. Now you would have regularization and this regularization is in some form of it is a function that is typically a norm of the parameter vector.

Now, the goal of the optimizer is to find theta such that this is minimized. So, you can see that these two terms are conflicting and the early one of the early penalty function that was suggested was the two norm, that is not only the optimizer will not only minimize this some square approximation errors, but also minimize the square two norms of theta; that means, it is reducing the variance energy of theta not the variance the energy. So, that as many thetas are possible are low value, ideally you want them to be zero value, but initially the low value problem was being solved and that is why the squared two norm was used and that is called tikhonov regularization, but then gradually people realized that this is still not doing the job, because just because you are minimizing the squared two norm, it does not mean that the unnecessary thetas will be driven to 0, they will be low value all of them will be awake you do not want them to be active, which thetas are unnecessary they should be 0 valued.

So, then came in mid 90s originally was this theta 2 square, then in mid nineties came the one norm minimization one norm regularization, which was due to Tibshirani who is at Stanford, and he called this optimizer as lasso least absolute shrinkage selection operator that is the name for it.

Basically he showed that if you were to minimize this kind of an objective function then if the truth is such that the thetas some thetas are zero or many thetas are zero as many as possible, all those thetas will be driven to zero. So, that you as a user can relax, you can spend more time on your devices and let the optimizer do the job for you, it will come out and say you supplied 100. It just fall short of scolding, you it will say well only 3 were required. So, that is the called that that is now the latest development in parameter estimation and you will see this trend these kind of objective functions being used everywhere, people no longer solve just me least squares the vanilla least squares, they would always solve a regularized least squares.

Now, coming back to the connection here although we have used least squares here you could replace this with a likelihood as well log likelihood, because we know these forms a part of the log likelihood; the vanilla likelihood did not have these regularization, but now you can think of regularized MLE and that is exactly what is happening here, this logarithm of prior can be thought of as a penalty function, because it is purely a function of theta, it turns out that it becomes MLE with tikhonov regularization if the if you use a Gaussian prior. Because if you use a Gaussian prior for theta, f of lon logarithm of f of theta would become square two norm of course, with the mean shifted theta, but that is.

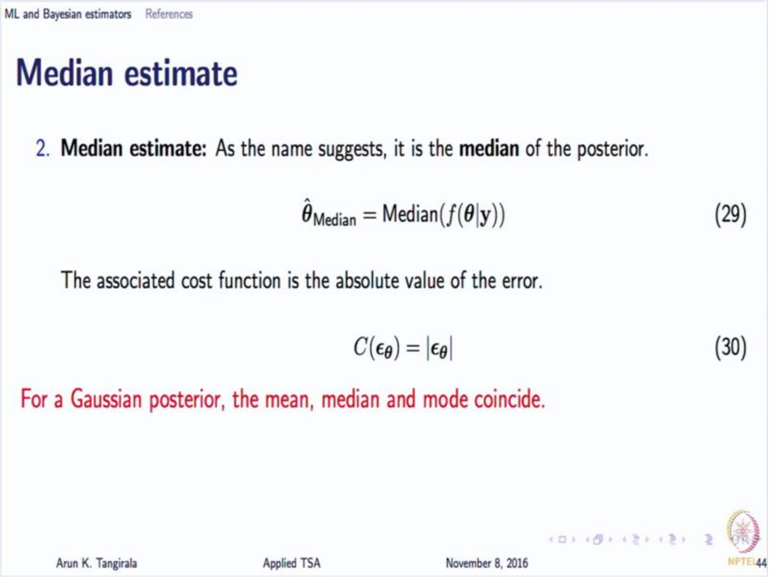
And you can show that if you choose a laplacian prior f of theta here laplacian, what is the difference between laplacian p.d.f and Gaussian p.d.f, do you know what is a laplacian p.d.f?

Student: (Refer Time: 08:32).

So, instead of square you have mod. So, laplacian prior is different is no different from Gaussian, but except that instead of a square you have a e to the minus modulus. If you choose such a prior then you can show that working with map is equivalent to working with a regularized MLE with one norm penalty function, again why? If f of theta is e to the minus mod something theta minus something, then when you take the logarithm it will essentially amount to working with one norm. So, these equivalences are excellent really very good, because you then know what problem you are essentially working with.

Therefore this map is a very popular estimator that people work with, but the computationally yeah it is certainly the; it is not as superior as the conditional expectation, because conditional expectation is like super dada you know comes to minimum mean square error. Map is maybe a slightly sub optimal estimate or in a optimal in some other sense, it is not the minimum mean square error estimator nevertheless it is your least squares with a penalty function or MLE with a penalty function and so on, but it has now the feature of some regular min minimization of approximation errors plus a penalty function, you can show that essentially working now with map amounts to working with regularized has estimation problems so, that is the connection.

(Refer Slide Time: 10:15)



ML and Bayesian estimators References

Median estimate

2. **Median estimate:** As the name suggests, it is the **median** of the posterior.

$$\hat{\theta}_{\text{Median}} = \text{Median}(f(\boldsymbol{\theta}|\mathbf{y})) \qquad (29)$$

The associated cost function is the absolute value of the error.

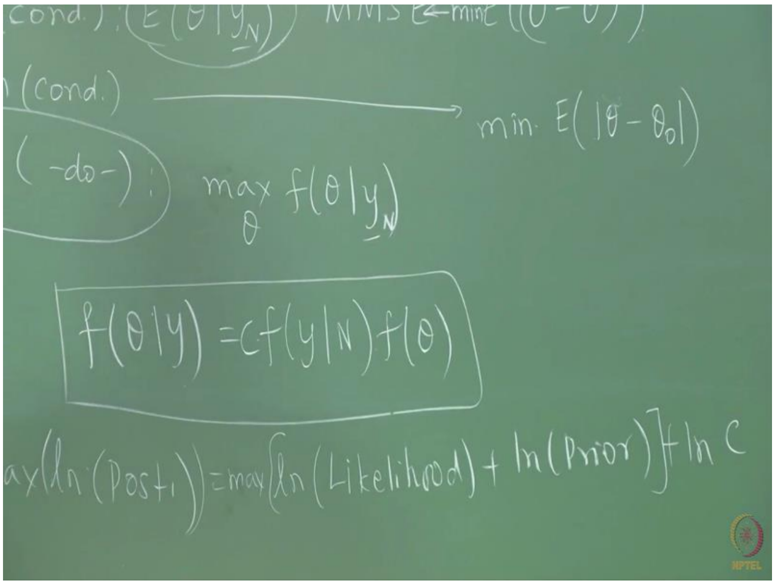
$$C(\epsilon_{\boldsymbol{\theta}}) = |\epsilon_{\boldsymbol{\theta}}| \qquad (30)$$

For a Gaussian posterior, the mean, median and mode coincide.

Arun K. Tangirala Applied TSA November 8, 2016 44

And finally, will move on to median. So, the median we know is what is the median of a p.d.f? It is the quantile which divided the cumulative into 2 halves; you can show that working with the median of the posterior amounts to minimizing this kind of a cost function in equation 30.

(Refer Slide Time: 10:42)



cond.) : $E(\theta|\underline{y}_N)$ MMSE ⟵ min E($(\theta - \theta)$)

(cond.) $\longrightarrow$ min. $E(|\theta - \theta_0|)$

(-do-) : $\max_{\theta} f(\theta|\underline{y}_N)$

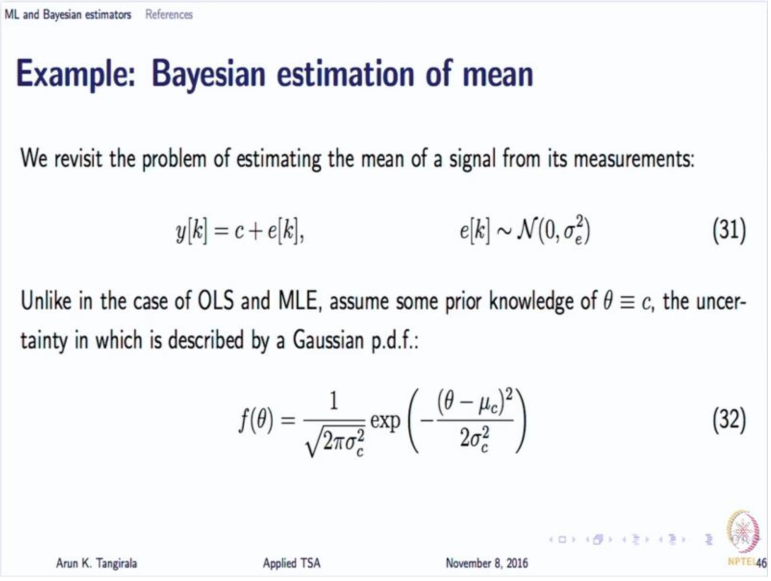
$$\boxed{f(\theta|y) = c f(y|N) f(\theta)}$$

ax(ln (Post.)) = max{ln (Likelihood) + ln (Prior)} + ln C

Which is basically minimizing expectation of median amounts to minimizing expectation of theta hat or you can say theta minus theta naught; that is different from your conditional expectation, we know that already that that is the property of a median for a Gaussian posterior that is if you somehow end up with a posterior that is Gaussian, the mean median and mode are the same.

In general MMSE is a most preferred, but; obviously, nature is not going to be so kind, it is computationally the most intrinsic, that is the only drawback, but it is computational power is not a big deal today and maybe not even a deal in future. But with a Gaussian posterior all these different estimators are simple. So, to summarize we obtain the posterior and derive point estimate from the posterior and you can derive any number of point estimators, 3 are popular the mean, that is the conditional expectation, the map which is the mode and then the median and if you have a Gaussian posterior it does not matter what you are working with.

(Refer Slide Time: 11:56)



ML and Bayesian estimators References

Example: Bayesian estimation of mean

We revisit the problem of estimating the mean of a signal from its measurements:

$$y[k] = c + e[k], \qquad e[k] \sim \mathcal{N}(0, \sigma_e^2) \qquad (31)$$

Unlike in the case of OLS and MLE, assume some prior knowledge of $\theta \equiv c$, the uncertainty in which is described by a Gaussian p.d.f.:

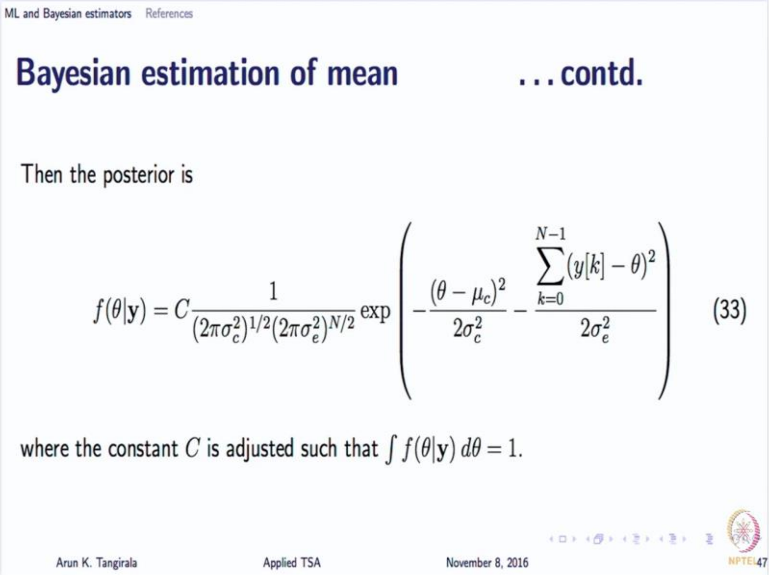
$$f(\theta) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(\theta - \mu_c)^2}{2\sigma_c^2}\right) \qquad (32)$$

Arun K. Tangirala Applied TSA November 8, 2016 46

So, let us very quickly go through an example again we come back to our familiar estimation of mean, until now we have been you have solved this using method of moments, we have solved this using least squares and MLE in all those 3 we have not assumed any prior knowledge of zee, but suppose I am given the prior knowledge of z which is theta here in this form. Suppose I say that I know that imagine this to be the average temperature, I know that the average temperature has a Gaussian prior, some values are more likely than others; which means now I have to specify mu c this is the mean of the prior and sigma square c as well, which is the variance of the prior it is a measure of the uncertainty that you have.

The idea is the data that I am going to use n observations will refine this prior; that means, it will sigma square c is the prior uncertainty; post estimation the uncertainty should reduce.

(Refer Slide Time: 13:00)



ML and Bayesian estimators References

Bayesian estimation of mean ...contd.

Then the posterior is

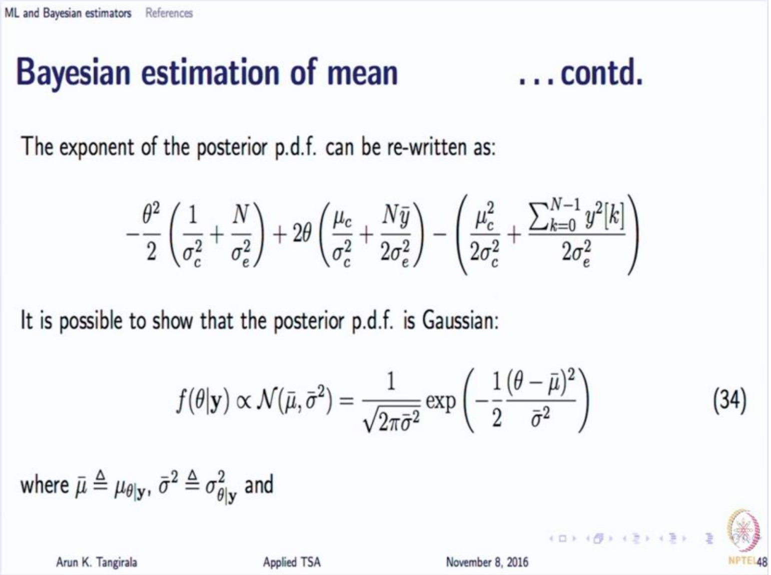
$$f(\theta|\mathbf{y}) = C\frac{1}{(2\pi\sigma_c^2)^{1/2}(2\pi\sigma_e^2)^{N/2}}\exp\left(-\frac{(\theta-\mu_c)^2}{2\sigma_c^2} - \frac{\sum\limits_{k=0}^{N-1}(y[k]-\theta)^2}{2\sigma_e^2}\right) \qquad (33)$$

where the constant C is adjusted such that $\int f(\theta|\mathbf{y})\,d\theta = 1$.

Arun K. Tangirala Applied TSA November 8, 2016 47

So, what happens when we do that? If you assume since you have already given by the way, the measurements fall out of a Gaussian white noise process, we club together the prior and the likelihood everything that is what we have done here and when we do this we will of course, run into this constant we will adjust this constant, such that the posterior is a legitimate p.d.f. So, this is your big posterior, is it not yet in a palatable form, it is just in the form of a product of p.d.f.

(Refer Slide Time: 13:31)



ML and Bayesian estimators References

Bayesian estimation of mean ...contd.

The exponent of the posterior p.d.f. can be re-written as:

$$-\frac{\theta^2}{2}\left(\frac{1}{\sigma_c^2}+\frac{N}{\sigma_e^2}\right)+2\theta\left(\frac{\mu_c}{\sigma_c^2}+\frac{N\bar{y}}{2\sigma_e^2}\right)-\left(\frac{\mu_c^2}{2\sigma_c^2}+\frac{\sum_{k=0}^{N-1}y^2[k]}{2\sigma_e^2}\right)$$

It is possible to show that the posterior p.d.f. is Gaussian:

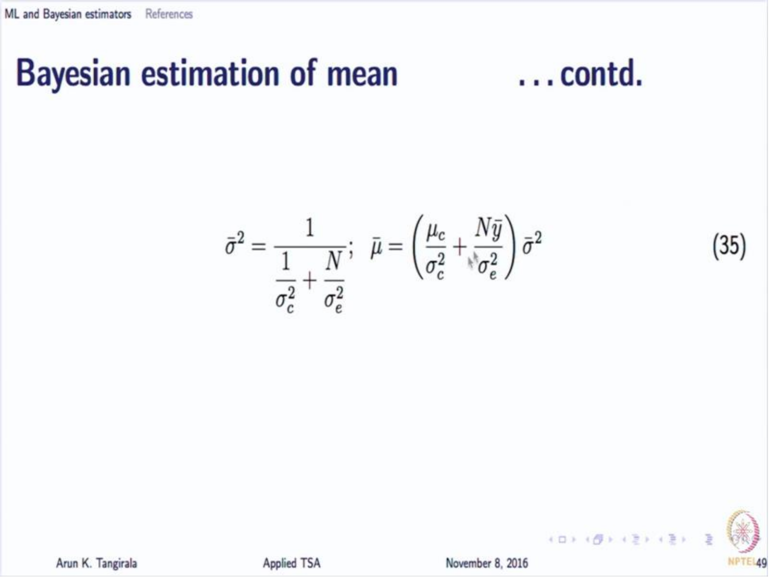
$$f(\theta|\mathbf{y}) \propto \mathcal{N}(\bar{\mu},\bar{\sigma}^2) = \frac{1}{\sqrt{2\pi\bar{\sigma}^2}}\exp\left(-\frac{1}{2}\frac{(\theta-\bar{\mu})^2}{\bar{\sigma}^2}\right) \qquad (34)$$

where $\bar{\mu} \triangleq \mu_{\theta|\mathbf{y}}$, $\bar{\sigma}^2 \triangleq \sigma^2_{\theta|\mathbf{y}}$ and

Arun K. Tangirala Applied TSA November 8, 2016 48

But you can show that you can express the posterior in the form of a Gaussian, and in fact, this is true whenever your prior is Gaussian and the measurements are Gaussian, you can show that the product can be expressed as a Gaussian.

(Refer Slide Time: 13:54)



ML and Bayesian estimators References

Bayesian estimation of mean ...contd.

$$\bar{\sigma}^2 = \frac{1}{\frac{1}{\sigma_c^2}+\frac{N}{\sigma_e^2}}; \quad \bar{\mu} = \left(\frac{\mu_c}{\sigma_c^2}+\frac{N\bar{y}}{\sigma_e^2}\right)\bar{\sigma}^2 \qquad (35)$$

Arun K. Tangirala Applied TSA November 8, 2016 49

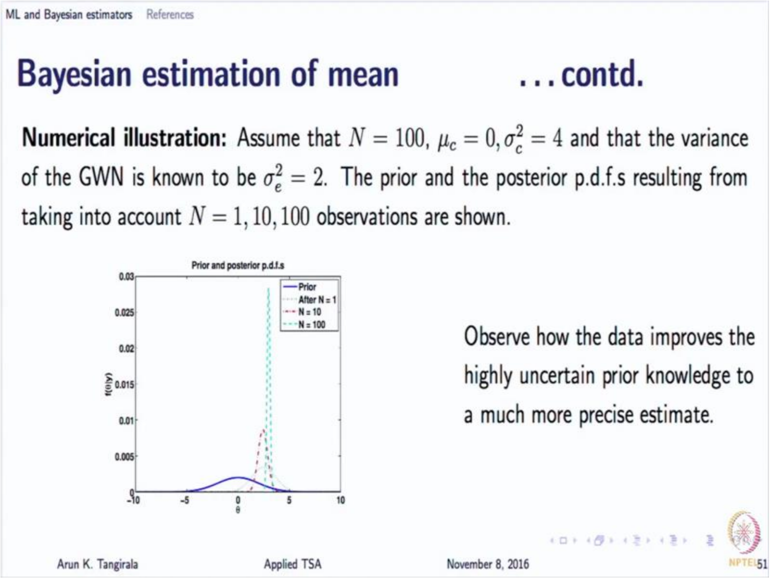
Now, the mean and variance of the resulting posterior are given here I will show you. So, the mean of the posterior is this and the variance of the posterior is given by this. So, you should quickly indentify these terms here mu c is the average of the prior and sigma square c is the variance associated with that, y bar is your sample mean it is coming from data and sigma square e over N, although I have written N by sigma square e, sigma square e over N what is it? It is the variance of y bar. So, what are we doing essentially here? We are constructing a weighted average of the prior and that comes from the data, what are we weighing it with? The inverse of the uncertainties; you see that mu c is the mean of the prior and there is a variance associated with the prior which is sigma square c and y bar is purely if you have to compute from data, that has a variance of sigma square e by N.

Now, you are fusing them to produce a new average, this is not the estimate remember what bayesian philosophy says, this is the posterior now you pick the maximum or the mean or median it is your choice. If I were to pick the conditional expectation of this, what would be the estimate? If I were to work with conditional expectation of the theta given by, the mean itself this is this mean here this it is the conditional mean.

The mean of the posterior is the conditional mean given y and that is the expression here. So, it says that the conditional expectation is a fusion of 2 estimates: 1 your prior guess, 2 you're the guess the estimate that you derive from data and what is it that we are hoping that the variance of this posterior estimate is going to be lesser, than what I begin with? And that you should actually check here right and you can see here again sigma square c comes from the prior uncertainty and sigma square e comes from the sigma square e over N comes from data. Suppose sigma square e is extremely high that is the case, then what is happening in this expression here in terms of the waiting that we had given?

You have giving it very low weightage; that means, that the data is almost garbage, it has what we mean by sigma square e by [noise]; it is actually sigma square e by N, it could be that sigma square e is very high or N is very small whatever may be the case, the sigma square e over N if that is very high then; that means, the data is not so reliable, you will stick to your prior. If the prior is very poor and the data is reliable more weightage is given to the data that is the understanding. Let us look at the simulation here very quickly and then close the discussion.

(Refer Slide Time: 17:05)



ML and Bayesian estimators References

Bayesian estimation of mean ...contd.

Numerical illustration: Assume that $N = 100$, $\mu_c = 0, \sigma_c^2 = 4$ and that the variance of the GWN is known to be $\sigma_e^2 = 2$. The prior and the posterior p.d.f.s resulting from taking into account $N = 1, 10, 100$ observations are shown.

Observe how the data improves the highly uncertain prior knowledge to a much more precise estimate.

Arun K. Tangirala Applied TSA November 8, 2016 51

So, what we begin with here I do not know how well you can see this figure, I have assume that I have 100 I mean forget about number of observations, let us assume that I begin with a 0 prior; that means, the prior as a 0 average and variance 4 and the true sorry the data has let us say variance of 2 then the what I show you here is a posterior p.d.fs resulting from different observations. So, let me actually zoom in here, is this visible now?

So, if you look at it there are basically you have this blue line which is a prior right, prior is without before you begin and then you have this dotted line after N equals 1 if you are able to see. I can zoom in further. So, this is the dotted line, what has happened when I began with the prior? We said the average is 0 and huge uncertainty, all right when I obtain 1 observation what happen to the uncertainty, did it shrink? Yes a bit and the average also starts to shift; which means the conditional expectation is giving me an improvement over the initial. As I increase the data size to 10 and to100 you can see the uncertainty is shrinking.

So, this is very nice it is telling us very clearly, the role of the prior and the number of observations also required to shrink that how quickly that uncertainty can be shrunk, quickly in the sense in terms of sample size. If your uncertainty was less, then the number of observations required to go from prior to a certain precision will be more that is all. So, this a very nice and simple example to understand what bayesian estimation is doing for you; any questions on this example, any questions on bayesian estimation? So, is from the book by Tangirala, in chapter 15, you can of course, the codes are given in math lab, you can import you can rewrite in r if you wish anyway.

So, this brings us to the conclusion of bayesian estimation. So, with this we are come to the close closure of all estimation methods that I wanted to discuss, what remains to be learnt is how to apply these estimation methods to two things estimating signal properties namely: mean, variance, covariance, correlation, autocorrelation, cross correlation and power spectral density and time series modeling, but time series modeling part we have been discussing in bits and pieces for example, we have already talked about Yule walker equations, that is nothing but applying the method of moments to estimating the parameters of a time series model. When we spoke of least squares I showed you how to set up the regressors and so on. Nevertheless we will go through a round roundup of how to take we will omit the bayesian part, bayesian was only is only a preliminary basis, we will not learn how to use bayesian for time series model, the philosophy has been put forth and hopefully you have as grasped the concepts.

We will learn how to actually apply the MLE for example, in estimating auto regressive models because the challenge there is setting up the likelihood function, because we are going to have a correlated series. So, there is a small trick, once you know the trick then you are on your way that is all and then we will talk about estimation of ARMA models, we know ARMA models lead to non-linear least squares problems, there is not much to learn there, but some bits and pieces and then will close the discussion with how to forecast using those models and hopefully we will have time for a short case study.