6.041SC Probabilistic Systems Analysis and Applied Probability, Fall 2013 Transcript – Lecture 22

PROFESSOR: So we're going to finish today our discussion of Bayesian Inference, which we started last time. As you probably saw there's not a huge lot of concepts that we're introducing at this point in terms of specific skills of calculating probabilities. But, rather, it's more of an interpretation and setting up the framework.

So the framework in Bayesian estimation is that there is some parameter which is not known, but we have a prior distribution on it. These are beliefs about what this variable might be, and then we'll obtain some measurements. And the measurements are affected by the value of that parameter that we don't know. And this effect, the fact that X is affected by Theta, is captured by introducing a conditional probability distribution-- the distribution of X depends on Theta. It's a conditional probability distribution.

So we have formulas for these two densities, the prior density and the conditional density. And given that we have these, if we multiply them we can also get the joint density of X and Theta. So we have everything that's there is to know in this second.

And now we observe the random variable X. Given this random variable what can we say about Theta? Well, what we can do is we can always calculate the conditional distribution of theta given X. And now that we have the specific value of X we can plot this as a function of Theta.

OK. And this is the complete answer to a Bayesian Inference problem. This posterior distribution captures everything there is to say about Theta, that's what we know about Theta. Given the X that we have observed Theta is still random, it's still unknown. And it might be here, there, or there with several probabilities.

On the other hand, if you want to report a single value for Theta then you do some extra work. You continue from here, and you do some data processing on X. Doing data processing means that you apply a certain function on the data, and this function is something that you design. It's the so-called estimator. And once that function is applied it outputs an estimate of Theta, which we call Theta hat.

So this is sort of the big picture of what's happening. Now one thing to keep in mind is that even though I'm writing single letters here, in general Theta or X could be vector random variables. So think of this-- it could be a collection Theta1, Theta2, Theta3. And maybe we obtained several measurements, so this X is really a vector X1, X2, up to Xn.

All right, so now how do we choose a Theta to report? There are various ways of doing it. One is to look at the posterior distribution and report the value of Theta, at which the density or the PMF is highest. This is called the maximum a posteriori estimate. So we pick a value of theta for which the posteriori is maximum, and we report it. An alternative way is to try to be optimal with respects to a mean squared error. So what is this?

If we have a specific estimator, g, this is the estimate it's going to produce. This is the true value of Theta, so this is our estimation error. We look at the square of the estimation error, and look at the average value. We would like this squared estimation error to be as small as possible. How can we design our estimator g to make that error as small as possible?

It turns out that the answer is to produce, as an estimate, the conditional expectation of Theta given X. So the conditional expectation is the best estimate that you could produce if your objective is to keep the mean squared error as small as possible. So this statement here is a statement of what happens on the average over all Theta's and all X's that may happen in our experiment.

The conditional expectation as an estimator has an even stronger property. Not only it's optimal on the average, but it's also optimal given that you have made a specific observation, no matter what you observe. Let's say you observe the specific value for the random variable X. After that point if you're asked to produce a best estimate Theta hat that minimizes this mean squared error, your best estimate would be the conditional expectation given the specific value that you have observed.

These two statements say almost the same thing, but this one is a bit stronger. This one tells you no matter what specific X happens the conditional expectation is the best estimate. This one tells you on the average, over all X's may happen, the conditional expectation is the best estimator.

Now this is really a consequence of this. If the conditional expectation is best for any specific X, then it's the best one even when X is left random and you are averaging your error over all possible X's.

OK so now that we know what is the optimal way of producing an estimate let's do a simple example to see how things work out. So we have started with an unknown random variable, Theta, which is uniformly distributed between 4 and 10. And then we have an observation model that tells us that given the value of Theta, X is going to be a random variable that ranges between Theta - 1, and Theta + 1. So think of X as a noisy measurement of Theta, plus some noise, which is between -1, and +1.

So really the model that we are using here is that X is equal to Theta plus U -- where U is uniform on -1, and +1. one, and plus one. So we have the true value of Theta, but X could be Theta - 1, or it could be all the way up to Theta + 1. And the X is uniformly distributed on that interval. That's the same as saying that U is uniformly distributed over this interval.

So now we have all the information that we need, we can construct the joint density. And the joint density is, of course, the prior density times the conditional density. We go both of these.

Both of these are constants, so the joint density is also going to be a constant. 1/6 times 1/2, this is one over 12. But it is a constant, not everywhere. Only on the range of possible x's and thetas. So theta can take any value between four and ten, so these are the values of theta. And for any given value of theta x can take values from theta minus one, up to theta plus one.

So here, if you can imagine, a line that goes with slope one, and then x can take that value of theta plus or minus one. So this object here, this is the set of possible x and theta pairs. So the density is equal to one over 12 over this set, and it's zero everywhere else. So outside here the density is zero, the density only applies at that point.

All right, so now we're asked to estimate theta in terms of x. So we want to build an estimator which is going to be a function from the x's to the thetas. That's why I chose the axis this way-- x to be on this axis, theta on that axis-- Because the estimator we're building is a function of x. Based on the observation that we obtained, we want to estimate theta.

So we know that the optimal estimator is the conditional expectation, given the value of x. So what is the conditional expectation? If you fix a particular value of x, let's say in this range. So this is our x, then what do we know about theta? We know that theta lies in this range. Theta can only be sampled between those two values. And what kind of distribution does theta have? What is the conditional distribution of theta given x?

Well, remember how we built conditional distributions from joint distributions? The conditional distribution is just a section of the joint distribution applied to the place where we're conditioning. So the joint is constant. So the conditional is also going to be a constant density over this interval. So the posterior distribution of theta is uniform over this interval.

So if the posterior of theta is uniform over that interval, the expected value of theta is going to be the meet point of that interval. So the estimate which you report-- if you observe that theta-- is going to be this particular point here, it's the midpoint.

The same argument goes through even if you obtain an x somewhere here. Given this x, theta can take a value between these two values. Theta is going to have a uniform distribution over this interval, and the conditional expectation of theta given x is going to be the midpoint of that interval.

So now if we plot our estimator by tracing midpoints in this diagram what you're going to obtain is a curve that starts like this, then it changes slope. So that it keeps track of the midpoint, and then it goes like that again. So this blue curve here is our g of x, which is the conditional expectation of theta given that x is equal to little x.

So it's a curve, in our example it consists of three straight segments. But overall it's non-linear. It's not a single line through this diagram. And that's how things are in general. g of x, our optimal estimate has no reason to be a linear function of x. In general it's going to be some complicated curve.

So how good is our estimate? I mean you reported your x, your estimate of theta based on x, and your boss asks you what kind of error do you expect to get? Having observed the particular value of x, what you can report to your boss is what you think is the mean squared error is going to be. We observe the particular value of x. So we're conditioning, and we're living in this universe.

Given that we have made this observation, this is the true value of theta, this is the estimate that we have produced, this is the expected squared error, given that we have made the particular observation. Now in this conditional universe this is the expected value of theta given x. So this is the expected value of this random variable inside the conditional universe.

So when you take the mean squared of a random variable minus the expected value, this is the same thing as the variance of that random variable. Except that it's the variance inside the conditional universe. Having observed x, theta is still a random variable. It's distributed according to the posterior distribution. Since it's a random variable, it has a variance. And that variance is our mean squared error.

So this is the variance of the posterior distribution of Theta given the observation that we have made. OK, so what is the variance in our example? If X happens to be here, then Theta is uniform over this interval, and this interval has length 2. Theta is uniformly distributed over an interval of length 2. This is the posterior distribution of Theta. What is the variance? Then you remember the formula for the variance of a uniform random variable, it is the length of the interval squared divided by 12, so this is 1/3.

So the variance of Theta -- the mean squared error-- is going to be 1/3 whenever this kind of picture applies. This picture applies when X is between 5 and 9. If X is less than 5, then the picture is a little different, and Theta is going to be uniform over a smaller interval. And so the variance of theta is going to be smaller as well.

So let's start plotting our mean squared error. Between 5 and 9 the variance of Theta -- the posterior variance-- is 1/3. Now when the X falls in here Theta is uniformly distributed over a smaller interval. The size of this interval changes linearly over that range. And so when we take the square size of that interval we get a quadratic function of how much we have moved from that corner.

So at that corner what is the variance of Theta? Well if I observe an X that's equal to 3 then I know with certainty that Theta is equal to 4. Then I'm in very good shape, I know exactly what Theta is going to be. So the variance, in this case, is going to be 0.

If I observe an X that's a little larger than Theta is now random, takes values on a little interval, and the variance of Theta is going to be proportional to the square of the length of that little interval. So we get a curve that starts rising quadratically from here. It goes up forward 1/3. At the other end of the picture the same is true. If you observe an X which is 11 then Theta can only be equal to 10.

And so the error in Theta is equal to 0, there's 0 error variance. But as we obtain X's that are slightly less than 11 then the mean squared error again rises quadratically. So we end up with a

plot like this. What this plot tells us is that certain measurements are better than others. If you're lucky, and you see X equal to 3 then you're lucky, because you know Theta exactly what it is.

If you see an X which is equal to 6 then you're sort of unlikely, because it doesn't tell you Theta with great precision. Theta could be anywhere on that interval. And so the variance of Theta -- even after you have observed X -- is a certain number, 1/3 in our case.

So the moral to keep out of that story is that the error variance-- or the mean squared error-- depends on what particular observation you happen to obtain. Some observations may be very informative, and once you see a specific number than you know exactly what Theta is. Some observations might be less informative. You observe your X, but it could still leave a lot of uncertainty about Theta.

So conditional expectations are really the cornerstone of Bayesian estimation. They're particularly popular, especially in engineering contexts. There used a lot in signal processing, communications, control theory, so on. So that makes it worth playing a little bit with their theoretical properties, and get some appreciation of a few subtleties involved here.

No new math in reality, in what we're going to do here. But it's going to be a good opportunity to practice manipulation of conditional expectations. So let's look at the expected value of the estimation error that we obtained. So Theta hat is our estimator, is the conditional expectation. Theta hat minus Theta is what kind of error do we have? If Theta hat, is bigger than Theta then we have made the positive error.

If not, if it's on the other side, we have made the negative error. Then it turns out that on the average the errors cancel each other out, on the average. So let's do this calculation. Let's calculate the expected value of the error given X. Now by definition the error is expected value of Theta hat minus Theta given X.

We use linearity of expectations to break it up as expected value of Theta hat given X minus expected value of Theta given X. And now what? Our estimate is made on the basis of the data of the X's.

If I tell you X then you know what Theta hat is. Remember that the conditional expectation is a random variable which is a function of the random variable, on which you're conditioning on. If you know X then you know the conditional expectation given X, you know what Theta hat is going to be.

So Theta hat is a function of X. If it's a function of X then once I tell you X you know what Theta hat is going to be. So this conditional expectation is going to be Theta hat itself. Here this is-- just by definition-- Theta hat, and so we get equality to 0. So what we have proved is that no matter what I have observed, and given that I have observed something on the average my error is going to be 0.

This is a statement involving equality of random variables. Remember that conditional expectations are random variables because they depend on the thing you're conditioning on. 0 is

sort of a trivial random variable. This tells you that this random variable is identically equal to the 0 random variable.

More specifically it tells you that no matter what value for X you observe, the conditional expectation of the error is going to be 0. And this takes us to this statement here, which is inequality between numbers. No matter what specific value for capital X you have observed, your error, on the average, is going to be equal to 0.

So this is a less abstract version of these statements. This is inequality between two numbers. It's true for every value of X, so it's true in terms of these random variables being equal to that random variable. Because remember according to our definition this random variable is the random variable that takes this specific value when capital X happens to be equal to little x.

Now this doesn't mean that your error is 0, it only means that your error is as likely, in some sense, to fall on the positive side, as to fall on the negative side. So sometimes your error will be positive, sometimes negative. And on the average these things cancel out and give you a 0 --. on the average.

So this is a property that's sometimes giving the name we say that Theta hat is unbiased. So Theta hat, our estimate, does not have a tendency to be on the high side. It does not have a tendency to be on the low side. On the average it's just right.

So let's do a little more playing here. Let's see how our error is related to an arbitrary function of the data. Let's do this in a conditional universe and look at this quantity.

In a conditional universe where X is known then h of X is known. And so you can pull it outside the expectation. In the conditional universe where the value of X is given this quantity becomes just a constant. There's nothing random about it. So you can pull it out, the expectation, and write things this way. And we have just calculated that this quantity is 0. So this number turns out to be 0 as well.

Now having done this, we can take expectations of both sides. And now let's use the law of iterated expectations. Expectation of a conditional expectation gives us the unconditional expectation, and this is also going to be 0. So here we use the law of iterated expectations. OK.

OK, why are we doing this? We're doing this because I would like to calculate the covariance between Theta tilde and Theta hat. Theta hat is, ask the question -- is there a systematic relation between the error and the estimate?

So to calculate the covariance we use the property that we can calculate the covariances by calculating the expected value of the product minus the product of the expected values.

And what do we get? This is 0, because of what we just proved. And this is 0, because of what we proved earlier. That the expected value of the error is equal to 0.

So the covariance between the error and any function of X is equal to 0. Let's use that to the case where the function of X we're considering is Theta hat itself.

Theta hat is our estimate, it's a function of X. So this 0 result would still apply, and we get that this covariance is equal to 0.

OK, so that's what we proved. Let's see, what are the morals to take out of all this? First is you should be very comfortable with this type of calculation involving conditional expectations. The main two things that we're using are that when you condition on a random variable any function of that random variable becomes a constant, and can be pulled out the conditional expectation.

The other thing that we are using is the law of iterated expectations, so these are the skills involved. Now on the substance, why is this result interesting? This tells us that the error is uncorrelated with the estimate. What's a hypothetical situation where these would not happen? Whenever Theta hat is positive my error tends to be negative.

Suppose that whenever Theta hat is big then you say oh my estimate is too big, maybe the true Theta is on the lower side, so I expect my error to be negative. That would be a situation that would violate this condition. This condition tells you that no matter what Theta hat is, you don't expect your error to be on the positive side or on the negative side. Your error will still be 0 on the average.

So if you obtain a very high estimate this is no reason for you to suspect that the true Theta is lower than your estimate. If you suspected that the true Theta was lower than your estimate you should have changed your Theta hat.

If you make an estimate and after obtaining that estimate you say I think my estimate is too big, and so the error is negative. If you thought that way then that means that your estimate is not the optimal one, that your estimate should have been corrected to be smaller. And that would mean that there's a better estimate than the one you used, but the estimate that we are using here is the optimal one in terms of mean squared error, there's no way of improving it. And this is really captured in that statement. That is knowing Theta hat doesn't give you a lot of information about the error, and gives you, therefore, no reason to adjust your estimate from what it was.

Finally, a consequence of all this. This is the definition of the error. Send Theta to this side, send Theta tilde to that side, you get this relation. The true parameter is composed of two quantities. The estimate, and the error that they got with a minus sign. These two quantities are uncorrelated with each other. Their covariance is 0, and therefore, the variance of this is the sum of the variances of these two quantities.

So what's an interpretation of this equality? There is some inherent randomness in the random variable theta that we're trying to estimate. Theta hat tries to estimate it, tries to get close to it. And if Theta hat always stays close to Theta, since Theta is random Theta hat must also be quite random, so it has uncertainty in it.

And the more uncertain Theta hat is the more it moves together with Theta. So the more uncertainty it removes from Theta. And this is the remaining uncertainty in Theta. The uncertainty that's left after we've done our estimation. So ideally, to have a small error we want this quantity to be small. Which is the same as saying that this quantity should be big.

In the ideal case Theta hat is the same as Theta. That's the best we could hope for. That corresponds to 0 error, and all the uncertainly in Theta is absorbed by the uncertainty in Theta hat.

Interestingly, this relation here is just another variation of the law of total variance that we have seen at some point in the past. I will skip that derivation, but it's an interesting fact, and it can give you an alternative interpretation of the law of total variance.

OK, so now let's return to our example. In our example we obtained the optimal estimator, and we saw that it was a nonlinear curve, something like this. I'm exaggerating the corner of a little bit to show that it's nonlinear.

This is the optimal estimator. It's a nonlinear function of X -- nonlinear generally means complicated.

Sometimes the conditional expectation is really hard to compute, because whenever you have to compute expectations you need to do some integrals. And if you have many random variables involved it might correspond to a multi-dimensional integration. We don't like this. Can we come up, maybe, with a simpler way of estimating Theta? Of coming up with a point estimate which still has some nice properties, it has some good motivation, but is simpler. What does simpler mean? Perhaps linear.

Let's put ourselves in a straitjacket and restrict ourselves to estimators that's are of these forms. My estimate is constrained to be a linear function of the X's. So my estimator is going to be a curve, a linear curve. It could be this, it could be that, maybe it would want to be something like this. I want to choose the best possible linear function.

What does that mean? It means that I write my Theta hat in this form. If I fix a certain a and b I have fixed the functional form of my estimator, and this is the corresponding mean squared error. That's the error between the true parameter and the estimate of that parameter, we take the square of this.

And now the optimal linear estimator is defined as one for which these mean squared error is smallest possible over all choices of a and b. So we want to minimize this expression over all a's and b's. How do we do this minimization?

Well this is a square, you can expand it. Write down all the terms in the expansion of the square. So you're going to get the term expected value of Theta squared. You're going to get another term-- a squared expected value of X squared, another term which is b squared, and then you're going to get to various cross terms. What you have here is really a quadratic function of a and b.

So think of this quantity that we're minimizing as some function h of a and b, and it happens to be quadratic.

How do we minimize a quadratic function? We set the derivative of this function with respect to a and b to 0, and then do the algebra. After you do the algebra you find that the best choice for a is this 1, so this is the coefficient next to X. This is the optimal a.

And the optimal b corresponds of the constant terms. So this term and this times that together are the optimal choices of b. So the algebra itself is not very interesting. What is really interesting is the nature of the result that we get here.

If we were to plot the result on this particular example you would get the curve that's something like this. It goes through the middle of this diagram and is a little slanted. In this example, X and Theta are positively correlated. Bigger values of X generally correspond to bigger values of Theta.

So in this example the covariance between X and Theta is positive, and so our estimate can be interpreted in the following way: The expected value of Theta is the estimate that you would come up with if you didn't have any information about Theta. If you don't make any observations this is the best way of estimating Theta.

But I have made an observation, X, and I need to take it into account. I look at this difference, which is the piece of news contained in X? That's what X should be on the average. If I observe an X which is bigger than what I expected it to be, and since X and Theta are positively correlated, this tells me that Theta should also be bigger than its average value.

Whenever I see an X that's larger than its average value this gives me an indication that theta should also probably be larger than its average value. And so I'm taking that difference and multiplying it by a positive coefficient. And that's what gives me a curve here that has a positive slope.

So this increment-- the new information contained in X as compared to the average value we expected apriori, that increment allows us to make a correction to our prior estimate of Theta, and the amount of that correction is guided by the covariance of X with Theta. If the covariance of X with Theta were 0, that would mean there's no systematic relation between the two, and in that case obtaining some information from X doesn't give us a guide as to how to change the estimates of Theta.

If that were 0, we would just stay with this particular estimate. We're not able to make a correction. But when there's a non zero covariance between X and Theta that covariance works as a guide for us to obtain a better estimate of Theta.

How about the resulting mean squared error? In this context turns out that there's a very nice formula for the mean squared error obtained from the best linear estimate. What's the story here?

The mean squared error that we have has something to do with the variance of the original random variable. The more uncertain our original random variable is, the more error we're going to make. On the other hand, when the two variables are correlated we explored that correlation to improve our estimate.

This row here is the correlation coefficient between the two random variables. When this correlation coefficient is larger this factor here becomes smaller. And our mean squared error become smaller. So think of the two extreme cases. One extreme case is when rho equal to 1 -- so X and Theta are perfectly correlated.

When they're perfectly correlated once I know X then I also know Theta. And the two random variables are linearly related. In that case, my estimate is right on the target, and the mean squared error is going to be 0.

The other extreme case is if rho is equal to 0. The two random variables are uncorrelated. In that case the measurement does not help me estimate Theta, and the uncertainty that's left-- the mean squared error-- is just the original variance of Theta. So the uncertainty in Theta does not get reduced.

So moral-- the estimation error is a reduced version of the original amount of uncertainty in the random variable Theta, and the larger the correlation between those two random variables, the better we can remove uncertainty from the original random variable.

I didn't derive this formula, but it's just a matter of algebraic manipulations. We have a formula for Theta hat, subtract Theta from that formula. Take square, take expectations, and do a few lines of algebra that you can read in the text, and you end up with this really neat and clean formula.

Now I mentioned in the beginning of the lecture that we can do inference with Theta's and X's not just being single numbers, but they could be vector random variables. So for example we might have multiple data that gives us information about X.

There are no vectors here, so this discussion was for the case where Theta and X were just scalar, one-dimensional quantities. What do we do if we have multiple data? Suppose that Theta is still a scalar, it's one dimensional, but we make several observations. And on the basis of these observations we want to estimate Theta.

The optimal least mean squares estimator would be again the conditional expectation of Theta given X. That's the optimal one. And in this case X is a vector, so the general estimator we would use would be this one.

But if we want to keep things simple and we want our estimator to have a simple functional form we might restrict to estimator that are linear functions of the data. And then the story is exactly the same as we discussed before. I constrained myself to estimating Theta using a linear function of the data, so my signal processing box just applies a linear function.

And I'm looking for the best coefficients, the coefficients that are going to result in the least possible squared error. This is my squared error, this is (my estimate minus the thing I'm trying to estimate) squared, and then taking the average. How do we do this? Same story as before.

The X's and the Theta's get averaged out because we have an expectation. Whatever is left is just a function of the coefficients of the a's and of b's. As before it turns out to be a quadratic function. Then we set the derivatives of this function of a's and b's with respect to the coefficients, we set it to 0.

And this gives us a system of linear equations. It's a system of linear equations that's satisfied by those coefficients. It's a linear system because this is a quadratic function of those coefficients. So to get closed-form formulas in this particular case one would need to introduce vectors, and matrices, and metrics inverses and so on.

The particular formulas are not so much what interests us here, rather, the interesting thing is that this is simply done just using straightforward solvers of linear equations. The only thing you need to do is to write down the correct coefficients of those non-linear equations. And the typical coefficient that you would get would be what? Let say a typical quick equations would be -- let's take a typical term of this quadratic one you expanded.

You're going to get the terms such as a1x1 times a2x2. When you take expectations you're left with a1a2 times expected value of x1x2. So this would involve terms such as a1 squared expected value of x1 squared. You would get terms such as a1a2, expected value of x1x2, and a lot of other terms here should have a too.

So you get something that's quadratic in your coefficients. And the constants that show up in your system of equations are things that have to do with the expected values of squares of your random variables, or products of your random variables. To write down the numerical values for these the only thing you need to know are the means and variances of your random variables. If you know the mean and variance then you know what this thing is. And if you know the covariances as well then you know what this thing is.

So in order to find the optimal linear estimator in the case of multiple data you do not need to know the entire probability distribution of the random variables that are involved. You only need to know your means and covariances. These are the only quantities that affect the construction of your optimal estimator.

We could see this already in this formula. The form of my optimal estimator is completely determined once I know the means, variance, and covariance of the random variables in my model. I do not need to know how the details distribution of the random variables that are involved here.

So as I said in general, you find the form of the optimal estimator by using a linear equation solver. There are special examples in which you can get closed-form solutions. The nicest simplest estimation problem one can think of is the following-- you have some uncertain parameter, and you make multiple measurements of that parameter in the presence of noise.

So the Wi's are noises. I corresponds to your i-th experiment. So this is the most common situation that you encounter in the lab. If you are dealing with some process, you're trying to measure something you measure it over and over. Each time your measurement has some random error. And then you need to take all your measurements together and come up with a single estimate.

So the noises are assumed to be independent of each other, and also to be independent from the value of the true parameter. Without loss of generality we can assume that the noises have 0 mean and they have some variances that we assume to be known. Theta itself has a prior distribution with a certain mean and the certain variance.

So the form of the optimal linear estimator is really nice. Well maybe you cannot see it right away because this looks messy, but what is it really? It's a linear combination of the X's and the prior mean. And it's actually a weighted average of the X's and the prior mean. Here we collect all of the coefficients that we have at the top.

So the whole thing is basically a weighted average. 1/(sigma_i-squared) is the weight that we give to Xi, and in the denominator we have the sum of all of the weights. So in the end we're dealing with a weighted average. If mu was equal to 1, and all the Xi's were equal to 1 then our estimate would also be equal to 1.

Now the form of the weights that we have is interesting. Any given data point is weighted inversely proportional to the variance. What does that say? If my i-th data point has a lot of variance, if Wi is very noisy then Xi is not very useful, is not very reliable. So I'm giving it a small weight. Large variance, a lot of error in my Xi means that I should give it a smaller weight.

If two data points have the same variance, they're of comparable quality, then I'm going to give them equal weight. The other interesting thing is that the prior mean is treated the same way as the X's. So it's treated as an additional observation. So we're taking a weighted average of the prior mean and of the measurements that we are making. The formula looks as if the prior mean was just another data point. So that's the way of thinking about Bayesian estimation.

You have your real data points, the X's that you observe, you also had some prior information. This plays a role similar to a data point. Interesting note that if all random variables are normal in this model these optimal linear estimator happens to be also the conditional expectation. That's the nice thing about normal random variables that conditional expectations turn out to be linear.

So the optimal estimate and the optimal linear estimate turn out to be the same. And that gives us another interpretation of linear estimation. Linear estimation is essentially the same as pretending that all random variables are normal. So that's a side point. Now I'd like to close with a comment.

You do your measurements and you estimate Theta on the basis of X. Suppose that instead you have a measuring device that's measures X-cubed instead of measuring X, and you want to estimate Theta. Are you going to get to different a estimate? Well X and X-cubed contained the same information. Telling you X is the same as telling you the value of X-cubed.

So the posterior distribution of Theta given X is the same as the posterior distribution of Theta given X-cubed. And so the means of these posterior distributions are going to be the same. So doing transformations through your data does not matter if you're doing optimal least squares estimation. On the other hand, if you restrict yourself to doing linear estimation then using a linear function of X is not the same as using a linear function of X-cubed. So this is a linear estimator, but where the data are the X-cube's, and we have a linear function of the data.

So this means that when you're using linear estimation you have some choices to make linear on what? Sometimes you want to plot your data on a not ordinary scale and try to plot a line through them. Sometimes you plot your data on a logarithmic scale, and try to plot a line through them. Which scale is the appropriate one? Here it would be a cubic scale. And you have to think about your particular model to decide which version would be a more appropriate one.

Finally when we have multiple data sometimes these multiple data might contain the same information. So X is one data point, X-squared is another data point, X-cubed is another data point. The three of them contain the same information, but you can try to form a linear function of them. And then you obtain a linear estimator that has a more general form as a function of X.

So if you want to estimate your Theta as a cubic function of X, for example, you can set up a linear estimation model of this particular form and find the optimal coefficients, the a's and the b's.

All right, so the last slide just gives you the big picture of what's happening in Bayesian Inference, it's for you to ponder. Basically we talked about three possible estimation methods. Maximum posteriori, mean squared error estimation, and linear mean squared error estimation, or least squares estimation. And there's a number of standard examples that you will be seeing over and over in the recitations, tutorial, homework, and so on, perhaps on exams even. Where we take some nice priors on some unknown parameter, we take some nice models for the noise or the observations, and then you need to work out posterior distributions in the various estimates and compare them.

6.041SC Probabilistic Systems Analysis and Applied Probability
Fall 2013