Variational Bayes Analysis

Meeting Transcript

David Ewing

Transcript

[0:00:00] Speaker 1: Question as well. That's why it's variational data say that again. So if you look so for example, if you're looking at the Wikipedia page on Laplace approximation, it will tell you that

[0:00:20] Speaker 1: it's yours. You're replacing the posterior with a normal and you are right that in many settings and variational phase, that's what you do. The difference is how you work out the mean and variance of that normal distribution, because they are not going to be the same. Okay? So in Laplace approximation, the mean is simply the value of theta, where the first derivative of the log of the joint distribution is equal to zero. In variational phase, the positive the the posterior means are the values where the expected value of the first derivative is equal to zero, okay, and under the approximating distribution and in Laplace approximation, The variance covariance matrix is equal to the inverse of the negative of the second derivative of the log joint distribution evaluated at theta hat and variational base, that variance covariance matrix will be equal to the expected value of it. That's why. When you look at some of your examples, you will notice Laplace is the most spiked

[0:01:50] Speaker 2: the most. So if you look

[0:01:53] Speaker 1: here on page eight and debug two, you'll notice it's the Laplace. I believe that's gone, got the highest density at the peak, okay, which is telling you it's the one that's most heavily underestimating the true uncertainty in the problem. That makes sense, because you're putting in a point is essentially from how the point estimate is constructed. Okay, the variational base does beta, because your those point the parameters of the approximating distributions are determined by the expectations of things like derivatives, not by just saying the derivative is equal to something. They will not be the same thing. In this particular example, you will notice that the Laplace approximation, the modes are the same, is just showing up on the variance. If I'd given you a GLM, you would actually find that both the modes and variances would change, but we're not going to cover that. Okay. Well, the thing, and if you want, you don't even need to include a Laplace approximation in here, particularly, is, I see you haven't done the Laplace approximation for the u's and the tau's. So I suspect the Laplace approximation here is not on the same model, because you've only plotted the Laplace on the betas.

[0:03:38] Speaker unknown: Okay, I've been struggling to get every all this done. That's all,

[0:03:42] Speaker 1: yeah, I haven't done so it's over you. I'd skip the Laplace.

[0:03:46] Speaker 2: Okay, well, this was what I ended up taking in coming up with for the comparison of the estimated value of tau u. So that's good. The third one is basically the last two are basically with a stronger

[0:04:10] Speaker 1: prior, and a stronger prior will be draining information away. So it's not just a question of stronger prior, it's where that stronger prior is taking you to so is that, because if you put a very strong prior to center it at 0.5 it would have been drained very close to 0.5 you'll put a very strong prior on that's clearly draining it away from 0.5 so I need to look at what that strong prior is to work out what value it's trying to emphasize. And the other thing I would say is, with the towels, the emphasis because, as you can see with these they're quite close. What you're more interested in is showing the changes in the posterior variance compared to the Gibbs sampling. Yes, yes, yes. So that's, that's the whole idea. So in terms of what you're trying to present, yes, you probably want to include and show you will, more or less always the unreasonably unbiased estimates of the model parameters. What we're really interested in here because, I mean, that's fairly well known variational based what you're really interested in is showing how accurately you can act, how accurately you can reproduce the distribution, ie the posterior distribution, which means the focus is much more on the variance, because the variance is telling you how spread the distribution is, and also the shape. But because you don't mean to your variational base in this setting, you know, you've got the shape probably right. The question is whether you've got the spread right. So in other words, for example, what I'm trying to say is, you know, tau u and tau e should look gamma distributed, and the variational Bayes in the mean field case does produce gammas. So you know that that's fine. You know, it's quite good on the modes. The question is, How bad is it on the variant,

[0:06:32] Speaker 2: and how do we end up taking it? Isn't that what the that's what

[0:06:36] Speaker 1: that's what we do, and that's why you're having to run these Gibbs samples alongside it, because the Gibbs sampling. Yes, it's sampling base. But you know, if you take enough samples, you will get a sample that looks the sample you have will look like draws from the posterior at the two posterior distribution. And in an example like this, where you can't work out analytically what the posterior distribution is. This is the best you can do. And so the but you obviously know this is slow so very so the comparison for variational base is going to be with what you estimate from a Gibbs sample, and see how close you are. And that's why I wanted you to change these group sizes, because what you will find is you will do better with six groups than you will do with 15 than you will do with 15.

[0:07:32] Speaker 2: Is that because there's less samples per group? Yes.

[0:07:37] Speaker 1: So the big thing with this model, and this is why I wanted to choose this particular model, because it relatively like it can be coded up. I should say it can be code either. Yeah, you've got the Gibbs sample, you've got the variational base. It can be coded up. And it's

quite an easy one to understand to in terms of showing some your depend, your dependence on sample size, so the betas. So as you can see, the thing that's hardest to do in this model is the variance components. It's not the betas and the years. It's the variance components or the precisions, so taus or sigma squared, depending on how you want the parameters. Those are, those are the hardest things to do. They are particularly hard to do when you have a small sample size. Okay, that makes sense, and that's why I wanted you. That's why, I think, for the port, it's do this model with multiple groups, with changing the group sizes, but leaving the total number of observations unchanged. Because, if you because, then someone, no one, can say to you and say, Well, of course you're going to do better when you have more groups, because you've got more you've got more observations altogether. The idea is the total number of observations should be the same, but we're changing how many times each random effect level is seen, and what you will find is, the more times you see the same random effect level, the better you will do it, estimating the variance components. And then it needs to do with shrinkage. So if you don't see a random effect level that many times the base uni submit of that random effect that will be heavily shrunk, and variational base doesn't take if you look at the up, look at what it implies about the cost. Can the approximate posterior distribution for these precisions. It doesn't take shrinkage into account, okay? And that's really what we're trying to get across here and in your presentation and in your report.

[0:09:57] Speaker 2: No, I think I've asked you this before, but it still just doesn't make sense. When we end up taking doing an elbow, it's based on, on what we're looking for, for the mean, is it not, or the other, the expected value. We're not. We don't have an elbow based on the

[0:10:15] Speaker 1: since it's not the mean. I'm sorry, yes. I mean,

[0:10:23] Speaker unknown: the elbow is not it's not the mean of beta,

[0:10:29] Speaker 1: it's the elbow is measuring. We have minimized a distance between two distributions. So you can think of it is a mean. But as a mean, I want to say is a mean between two log distributions, not the betas and the years. Yeah, so it's a log distribution. You're trying to minimize difference between so. But of course, that log distribution has information about both the mean and variance. Okay, because the distribution itself is obviously a function of the parameters of the model, and parameters of the model control both mean and variance and all the other moments. So you are dear. So when you're optimizing the Ewing, you are actually optimizing something that is information about, yes, okay, but in practice, the choice of optimization, but how you work out something that optimizes them so variational Bayes and mean field cases, what you would do is essentially where, in a Gibbs sample, you would work out conditional posterior distributions and variational base these are your independent posterior distributions. And what you're trying to do is find the parameters that define these approximate posteriors. So the question and so obviously you pick something that, in theory, optimizes the difference between that, those approximations you've chosen, and the true block distribution, but because you can think of it as well. You're sort of saying the conditional posterior from it gives in many ways of being turned into unconditional

posteriors in the variational phase. The question is, where does the conditional really differ from the unconditional? And that in this sort of model is examples where it's a variance components. That's where it really, really differs, because it doesn't, because the the conditional posterior doesn't take into account the shrinkage. So if I because, obviously, I can code a whole set of conditional posterior distributions, but it's cycling through them that deals with, deals with the dependencies between subsets that obviously is getting ignored in variational ways.

[0:13:30] Speaker unknown: I'm not sure what to do with that. I don't think I understand that piece.

[0:13:34] Speaker 1: So that's why I think focus on a focus only on like this, do the different groups, and emphasize the sample size samples.

[0:13:44] Speaker 2: So the emphasis is on sample size. That ends up taking a mean that the so it

[0:13:50] Speaker 1: seems, it seems, yes, do it on sample size, and emphasize that this sort of model, the variational based approximations, will do a lot better with larger sample sizes. That's what you and if you wanted to explain why this is the case, what you would then do is look at the use so not not to understand why you're doing better with the posterior distributions of the variant of the tau's. When the sample size is bigger, you actually have to go back one step and look at the use, okay.

[0:14:28] Speaker 2: Answer. Here's a cross correlation between the between the the I mean the user, the cross correlation between the betas.

[0:14:36] Speaker 1: No. What is no, no. The your random effects. Okay, so your model is x, beta plus z, U, yes. So these u's are assumed random, yes. And so the U is being drawn, in this case, from divided tau u. That's why you've got the second variance component. There's an epsilon, and the epsilon is drawn also, and that's where the tau e comes in. Okay, yes. And so what you want? And so to understand why tell you will end up looking much better when the number of times you see tremor feet level, the you the individual use increases. What You Should you can then do is plot

[0:15:35] Speaker 1: the variance, the posterior variance, of the use

[0:15:41] Speaker unknown: over the prior variances,

[0:15:47] Speaker unknown: the posterior variance over the

[0:15:49] Speaker 1: prior variance. So in the examples, so in the examples, where you do quite badly with tell you the sorry, I just need to talking about

[0:16:20] Speaker 1: E, so yes, so when you do badly with tau u, this value will look quite high. When you do well, this value will be low.

[0:16:40] Speaker unknown: So if you want, are you

[0:16:42] Speaker unknown: talking about over one or bigger than one or no?

[0:16:45] Speaker 1: It's, by construction, going to be a number between zero and one. Okay. But what you should find is is you do well with approximating tau u, this ratio will go to zero, and if the posterior variance over the prior variance will go to zero, as your variational approximation for the posterior of tau u

[0:17:08] Speaker 2: in prose, does that mean that this is either going to go the infinite

[0:17:13] Speaker 1: or this? No, this is finite. Okay. How does it go to zero? Because the posterior variance is going to zero. Okay, the posterior variance, the posterior variance, goes to zero, and as the posterior variance team becomes smaller and smaller, you will do better with your variational approximation for tau u, okay,

[0:17:35] Speaker 2: so that means that more like this, not on this one, but I mean for tau u, yes, that's basically, it's going to be

[0:17:42] Speaker 1: a lot more. Yeah. So essentially, the narrower the posterior distributions for the use are, the better you will do it across, getting the posterior for tau u, okay, the same thing will although you're not looking at it, because we always use the same total number of observations. But if we started varying the total number of observations, you'd see the same thing occur with tau e. Not that it matters, because the sample sizes shows

[0:18:14] Speaker unknown: high enough, although you might get some

[0:18:19] Speaker 1: you might get some effect from the U's coming in as well. So as q, then you'll find Q goes to the true posterior. And this, of course, in practice, it was just Q, the number of levels divided by tau u, okay, because your prior is that they're all independent, and you've got, well, that's the same region, yes, yeah, you've got Q levels.

[0:18:52] Speaker unknown: Okay, all right, we have to digest that.

[0:18:58] Speaker 1: So then that's essentially the big thing I've been trying to get you to get towards is that realization that there are just some things you can't do very well with variational phase. But generally speaking, most of them are simply functions of sample size. And that if you've got a large enough sample and enough observations for each subset of the main effect in the first level of the hierarchy, it will be probably

[0:19:39] Speaker 2: fine the intention of going variational basis. I understand it is the main things go fast, yes. So if you end up taking better with a bigger sample size, there's a there's a trade off there,

[0:19:50] Speaker 1: because that's going to be the Gibbs Sampler. Mtmc is going to be slower to increase the sample size as well. Okay, so you've got to think of it like that. It's not that you'll give sample it gets any easier to do or any quicker to do as your sample size increases. Yes, obviously variational base takes longer to do, but bit more down, yeah, but it's still going to always be faster than the EMC.

[0:20:17] Speaker unknown: So, but the threat,

[0:20:21] Speaker 2: the just where, how close do you want to be is effectively

[0:20:25] Speaker unknown: real, the big question,

[0:20:28] Speaker 1: okay, and I'm and then on top of it, there are also ways you can deliberately use the title,

[0:20:36] Speaker unknown: okay, So,

[0:20:39] Speaker 1: and I do have code for something that you can deliberately Miss other items, but

[0:20:44] Speaker 2: that's a bit right now, that's outside the scope of what I'm looking at

[0:20:49] Speaker 1: because so we could just leave it at this. Because the other thing to really be set up is so the code I send to you update, assumes the code note, the VB you're Working with

[0:21:17] Speaker 1: is assumes the parameterization yes of the posterior as

[0:21:28] Speaker unknown: now, if I wanted to really mess us up,

[0:21:32] Speaker 1: I would change beta Q of beta u to every individual beta and every individual u as a okay, and that will look a lot worse. No, is

[0:21:46] Speaker unknown: it working because they're because

[0:21:50] Speaker unknown: the tenants on one to the other Exactly?

[0:21:52] Speaker 1: Essentially, there are two ways to mess up variational phase. One is to ignore interdependencies that you know exist. And so, for example, if we had factorizes, so every independent, every grant, every single scalar parameter was given its own approximate posterior, that would completely screw it up. But we screwed it up by our choices, whereas this is based on your our level of analytical ability is the

[0:22:26] Speaker unknown: you know that that's as blocked as we can.

[0:22:30] Speaker 1: And that means, okay, that should imply, because we've done all the blocking, that should imply we should get the best approximation we can manage. The question is, how close is this? And then we've got to ask questions of, okay, what features of variational Bayes just are not does a variational based algorithm is unable to take into account? And that's very much to do with the idea. Well, it's very much like give sampling. You're replacing a conditional. You're saying conditional posterior. In a way, it's becoming more unconditional posterior, where, in which situations does that not matter? Which situations doesn't matter? It probably doesn't matter very much for beating you in these sort of models, because we already know that once you've got enough of sample size, if you think of linear regression when you got taught, you use t distributions for confidence intervals for regression coefficients. Well, once the sample size is about 30 or 50, depending on perspective, you can't tell the difference between a T and a normal drink a t distribution. Oh, I see, yes. So you've got beta and you you're probably going to do fine. You probably don't care. So that leads the precisions. And the issue with the conditional posteriors is

so conditional posteriors don't take into consideration the loss of degrees of freedom in estimating some parameters when you're updating the precisions, which effectively is a bit like it's essentially, it assumes the sample size and doesn't knock out. Now, obviously that's much that loss of degree of freedom is much more pronounced if you have relatively few observations for each random effect level. Okay, so that's why, for the example, I was going to say, change the number of times each random effect level is observed. And what you will see is the approximation of the tau u will get better as you increase the number of observations, because that's telling you we haven't taken to account effective degrees of freedom. We don't actually know what the effective degrees of freedom are, but we know the loss of degree of freedom compared to the number of times the number of levels we have will decline as we increase the number of times we see intramurphic level. And that's what we see

[0:25:12] Speaker unknown: that and so that was the real rain.

[0:25:14] Speaker 2: So what I need to do is this again, but end up taking and doing it with just the changing of the level.

[0:25:24] Speaker 1: So it seems so, as I said, based on stuff you've been sending through to me, you've got the theory. This is going to be the example that goes into the report. We want to illustrate some features. If we have time, we might,

[0:25:40] Speaker unknown: because we could do that.

[0:25:43] Speaker 1: We could do that at the very beginning or the very end, and saying, This is user defined faults with variational base. And then we need to finish off with this example. Okay, we want the total number of observations to stay unchanged, but we're going to vary the number of random effect levels, which means varying the number of times each random effect level is seen, the number of times number of observations associated with each random effect level, the more times a random effect levels, the more observations associated with each random effect level, the better you will do it. Tell you, and that makes sense. If you look back to our proximate posterior for tau u, which has

[0:26:28] Speaker unknown: Q over two in it,

[0:26:31] Speaker unknown: which is very much like,

[0:26:36] Speaker 1: is very much like what you would get, which is what you get in the conditional posterior, but it doesn't, but that means it's not taking them out the loss of degree of

[0:26:48] Speaker unknown: freedom. And so that's how we're going to end

[0:26:52] Speaker 2: them, right? I will go away and do this, take this injury down. I have. Here's saw my original report, and I haven't given that back to you, because I didn't end up taking an understanding some of the fine. But I would like to, you know, get your eyes across it again.

[0:27:13] Speaker unknown: That's fine. Okay.

[0:27:17] Speaker 1: Yes. So, as I said, and obviously a lot of this will go into the appendix, but at times you obviously take out the exact yes and the Laplace, because it's not needed either. Okay, it's just variational based versus GIFs. Oh, and make sure you run multiple chains of the GIF sampler, because you need to show that it converges okay. And then so in terms of your report, you'll probably find you put things like this in to the actual report, and probably you won't need to put in all the betas in the US, because we're much more interested in particularly tau u. So you might find this is the posterior of tau u when we did this number of observations associated with each this each random effect level. So let's say we started with five, and then we went to team, being 20, being 50, and this is what happened to the posterior distribution of telling you compared to what

[0:28:27] Speaker 2: we got. Okay, and so should I end up taking, having those chart you could, you could see, I thought it was gonna get nasty. That's why I ended up trying to get one characteristic.

[0:28:36] Speaker 1: And no, it's, it's because you're, it's the distributions. You're really and don't, because there's no need. Don't, don't change the priors, because that's not that's not trying, that doesn't help you. Illustrate what we what you want to illustrate, which is the idea of variational base tends to struggle with shrinkage.

[0:29:03] Speaker 2: I'll tell you, I started out with basically that the where I where I thought that the smaller samples were going to be worse and they were better. Or this isn't this particular one I ended up taking and going through. But it was just struggling with what was going on with that. I made an error effectively, but it was just that that's where I spent a lot of time getting that that worked out so

[0:29:34] Speaker unknown: right, based on what anything else,

[0:29:36] Speaker 2: not at the moment, although I probably would ask to see you later this week, once I get this done,

[0:29:41] Speaker 1: email, mail by their arrangement time. Thank you very much. Applause.