Statistical Modeling, Causal Inference, and Social Science

Eid ma clack shaw zupoven del ba.

Posted by Dan Simpson on 7 February 2018, 6:36 pm

When I say "I love you", you look accordingly skeptical – Frida Hyvönen

A few years back, Bill Callahan wrote a song about the night he dreamt the perfect song. In a fever, he woke and wrote it down before going back to sleep. The next morning, as he struggled to read his handwriting, he saw that he'd written the nonsense that forms the title of this post.

Variational inference is a lot like that song; dreams of the perfect are ruined in the harsh glow of the morning after.

(For more unnaturally tortured metaphors see my twitter. I think we can all agree setting one up was a fairly bad choice for me.)

But how can we tell if variational inference has written the perfect song or, indeed, if it has laid an egg? Unfortunately, there doesn't seem to be a lot of literature to guide us. We (Yuling, Aki, Me, and Andrew) have a new paper to give you a bit more of an idea.

Palimpsest

The guiding principle of variational inference is that if it's impossible to work with the true posterior $p(\theta \mid y)$, then near enough is surely good enough. (It seldom is.)

In particular, we try to find the member $q^*(\theta)$ of some tractable set of distributions $\mathcal{Q}$ (commonly the family of multivariate Gaussian distributions with diagonal covariance matrices) that minimizes the *Kullback-Leibler* divergence

$$q^*(\theta) = \arg\min_{q \in \mathcal{Q}} KL\left[q(\cdot) \mid p(\cdot \mid y)\right].$$

The Kullback-Leibler divergence in this direction (it's asymmetric, so the order of arguments is important) can be interpreted as the the amount of information lost if we replace the approximate posterior $q(\theta)$ with the true posterior $p(\theta \mid y)$. Now, if this seems like the wrong way around to you [that we should instead worry about what happens if we replace the target posterior $p(\theta \mid y)$ with the approximation $q(\theta)$], you would be very *very* correct. That Kullback-Leibler divergence is *backwards.*

What does this mean? Well it means that we won't penalize approximate distributions that are much less complex than the true one as heavily as we should. *How does this translate into real life?* It means that usually we will end up with approximations $q^*(\theta)$ that are narrower than the true posterior. Usually this manifests as distributions with lighter tails.

(*Quiet side note:* Why are those last few sentences so wishy-washy? Well it turns out that minimizing a Kullback-Leibler divergence in the wrong direction can do all kinds of things to the resulting minimizer and it's hard to really pin down what will happen. But it's almost everyone's experience that the variational posterior $q(\theta)$ is almost always narrower than than the true posterior. So the previous paragraph is *usually* true.)

So variational inference is mostly set up to fail. Really, we should be surprised it works at all.

Cold discovery

There are really two things we need to check when we're doing variational inference. The first is that the optimization procedure that we have used to compute $q^*(\theta)$ has actually converged to a (local) minimum. Naively, this seems fairly straightforward. After all, we don't think of maximum likelihood estimation as being hard computationally, so we should be able to solve this optimization problem easily. But it turns out that if we want our variational inference to be *scalable* in the sense that we can apply it to big problems, we need to be more clever. For example Automatic Differentiation Variational Inference (ADVI) uses a fairly sophisticated stochastic optimization method to find $q^*(\theta)$.

So first we have to make sure the method actually converges. I don't really want to talk about this, but it's probably worth saying that it's not trivial and stochastic methods like ADVI will occasionally terminate too soon. This leads to terrible approximations to the true posterior. It's also well worth saying the if the true posterior is multimodal, there's no guarantee that the minimum that is found will be a (nearly) global one. (And if the approximating family $\mathcal{Q}$ only contains unimodal distributions, we will have some problems!) There are perhaps some ways out of this (Yuling has many good ideas), but the key thing is that if you want to actually know if there is a potential problem, it's important to run multiple optimizations beginning at a diverse set of initial values.

Anyway, let's pretend that this isn't a problem so that we can get onto the main point.

The second thing that we need to check is that the approximate posterior $q^*(\theta)$ is an ok approximation to the true posterior $p(\theta \mid y)$. This is a much less standard task and we haven't found a good method for addressing it in the literature. So we came up with two ideas.

Left only with love

Our first idea was based Aki, Andrew, and Jonah's *Pareto-Smoothed Importance Sampling* (PSIS). The crux of our idea is that if $q(\theta)$ is a good approximation to the true posterior, it can be used as an importance sampling proposal to compute expectations with respect to $p(\theta \mid y)$. So before we can talk about that method, we need to remember what PSIS does.

The idea is that we can approximate any posterior expectation $\int h(\theta)p(\theta \mid y)\, d\theta$ using a self-normalized importance sampling estimator. We do this by drawing S samples $\{\theta_s\}_{s=1}^S$ from the proposal distribution $q(\theta)$ and computing the estimate

$$\int h(\theta)p(\theta \mid y)\, d\theta \approx \frac{\sum_{s=1}^S h(\theta_s)r_s}{\sum_{s=1}^S r_s}.$$

Here we define the *importance weights* as

$$r_s = \frac{p(\theta_s, y)}{q(\theta_s)}.$$

We can get away with using the joint distribution instead of the posterior in the numerator because $p(\theta \mid y) \propto p(\theta, y)$ and we re-normalise the the estimator. This self-normalized importance sampling estimator is consistent with bias that goes asymptotically like $\mathcal{O}(S^{-1})$. (The bias comes from the self-normalization step. Ordinary importance sampling is unbiased.)

The only problem is that if the distribution of r_s has too heavy a tail, the self-normalized importance sampling estimator will have infinite variance. This is not a good thing. Basically, it means that the error in the posterior expectation could be any size.

The problem is that if the distribution of r_s has a heavy tail, the importance sampling estimator will be almost entirely driven by a small number of samples θ_s with very large r_s values. But there is a trick to get around this: somehow tamp down the extreme values of r_s.

With PSIS, Aki, Andrew, and Jonah propose a nifty solution. They argue that you can model the tails of the distribution of the importance ratio with a *generalized Pareto distribution*

$$p(r|\mu,\sigma,k) = \begin{cases} \frac{1}{\sigma}\left(1+k\left(\frac{r-\mu}{\sigma}\right)\right)^{-\frac{1}{k}-1}, & k \neq 0. \\ \frac{1}{\sigma}\exp\left(\frac{r-\mu}{\sigma}\right), & k = 0. \end{cases}.$$

This is a very sensible thing to do: the generalized Pareto is the go-to distribution that you use when you want to model the distribution of all samples from an iid population that are above a certain (high) value. The PSIS approximation argues that you should take the M largest r_s (where M is chosen carefully) and fit a generalized Pareto distribution to them. You then replace those M largest observed importance weights with the corresponding expected order statistics from the fitted generalized Pareto.

There are some more (critical) details in the PSIS paper but the intuition is that we are replacing the "noisy" sample importance weights with their model-based estimates. This reduces the variance of the resulting self-normalized importance sampling estimator and reduces the bias compared to other options.

It turns out that the key parameter in the generalized Pareto distribution is the shape parameter k. The interpretation of this parameter is that if the generalized Pareto distribution has shape parameter k, then the distribution of the sampling weights have $\lfloor k^{-1} \rfloor$ moments.

This is particularly relevant in this context as the condition for the importance sampling estimator to have finite variance (and be asymptotically normal) is that the sampling weights have (slightly more than) two moments. This translates to $k < 1/2$.

VERY technical side note: What I want to say is that the self-normalized importance sampling estimator is asymptotically normal. This was nominally proved in Theorem 2 of Geweke's 1983 paper. The proof there looks wrong. Basically, he applies a standard central limit theorem to get the result, which seems to assume the terms in the sum are iid. The only problem is that the summands

$$\frac{h(\theta_s)r_s}{\sum_{s=1}^S r_s}$$

are not independent. So it looks a lot like Geweke should've used a central limit theorem for weakly-mixing triangular arrays instead. He did not. What he actually did was quite clever. He noticed that the bivariate random variables $(h(\theta_s)r_s, r_s)^T$ are independent and satisfy a central limit theorem with mean $(A, w)^T$. From there you're a second-order Taylor expansion of the function $f(A, w) = A/w$ to show that the sequence

$$f\left(S^{-1}\sum_{s=1}^S h(\theta_s)r_s, S^{-1}\sum_{s=1}^S r_s\right)$$

is also asymptotically normal as long as zero or infinity are never in a neighbourhood of $S^{-1}\sum_{s=1}^S r_s$.

End VERY technical side note!

The news actually gets even better! The smaller k is, the faster the importance sampling estimate will converge. Even better than that, the PSIS estimator seems to be useful even if k is slightly bigger than 0.5. The recommendations in the PSIS paper is that if $\hat{k} < 0.7$, the PSIS estimator is reliable.

But what is $\hat{k}$? It's the sample estimate of the shape parameter k. Once again, some really nice things happen when you use this estimator. For example, even if we know from the structure of the problem that $k < 0.5$, if $\hat{k} > 0.7$ (which can happen), then importance sampling will perform poorly. The value of $\hat{k}$ is strongly linked to the *finite sample* behaviour of the PSIS (and other importance sampling) estimators.

The intuition for why the estimated shape parameter is more useful than the population shape parameter is that it tells you when the sample of r_s that you've drawn *could have come from a heavy tailed distribution.* If this is the case, there isn't enough information in your sample yet to push you into the asymptotic regime and pre-asymptoic behaviour will dominate (usually leading to worse than expected behaviour).

Footprints

Ok, so what does all this have to do with variational inference? Well it turns out that if we draw samples from out variational posterior and use them to compute the importance weights, then we have another interpretation for the shape parameter k:

$$k = \arg\inf_{k'>0} D_{1/k}(p(\theta \mid y) \,||\, q(\theta),$$

where $D_\alpha(p, q)$ is the Rényi divergence of order α. In particular, if $k > 1$, then the Kullback-Leibler divergence in the more natural direction $KL(p(\theta \mid y) \mid q(\theta)) = \infty$ even if $q(\theta)$ minimizes the KL-divergence in the other direction! Once again, we have found that the estimate $\hat{k}$ gives an excellent indication of the performance of the variational posterior.

So why is checking if $\hat{k} < 0.7$ a good heuristic to evaluate the quality of the variational posterior? There are a few reasons. Firstly, because the variational posterior minimizes the KL-divergence in the direction that penalizes approximations with heavier tails than the posterior much harder than approximations with lighter tails, it is very difficult to get a good $\hat{k}$ value by simply "fattening out" the approximation. Secondly, empirical evidence suggests that the smaller the value of $\hat{k}$, the closer the variational posterior is to the true posterior. Finally, if $\hat{k} < 0.7$ we can automatically improve *any* expectation computed against the variational posterior using PSIS. This makes this tool both a diagnostic and a correction for the variational posterior that does not rely too heavily on asymptotic arguments. The value of $\hat{k}$ has also proven useful for selecting the best parameterization of the model for the variational approximation (or equivalently, between different approximation families).

There are some downsides to this heuristic. Firstly, it really does check that the whole variational posterior is like the true posterior. This is a quite stringent requirement that variational inference methods often do not pass. In particular, as the number of dimensions increases, we've found that unless the approximating family is particularlttwell-chosen for the problem, the variational approximation will eventually become bad enough that $\hat{k}$ will exceeded the threshold. Secondly, this diagnostic only considers the full posterior and cannot be modified to work on lower-dimensional subsets of the parameter space. This means that if the model has some "less important" parameters, we still require their posterior be very well captured by the variational approximation.

Let me see the colts

The thing about variational inference is that it's actually often quite bad at estimating a posterior. On the other hand, the centre of the variational posterior is much more frequently a good approximation to the centre of the true posterior. This means that we can get good point estimates from variational inference even if the full posterior isn't very good. So we need a diagnostic to reflect this.

Into the fray steps an old paper of Andrew's (with Samatha Cook and Don Rubin) on verifying statistical software. We (mainly Stan developer Sean) have been playing with various ways of extending and refining this method for the last little while and we're almost done on a big paper about it. (Let me tell you: god may be present in the sweeping gesture, but the devil is definitely in the details.) Thankfully for this work, we don't need any of the new detailed work we've been doing. We can just use the original results as they are (with just a little bit of a twist).

The resulting heuristic, which we call *Variational Simulation-Based Calibration* (VSBC), complements the PSIS diagnostic by assessing the average performance of the implied variational approximation to *univariate* posterior marginals. One of the things that this method can do particularly well is indicate if the centre of the variational posterior will be, on average, biased. If it's not biased, we can apply clever second-order corrections (like the one proposed by Ryan Giordano, Tamara Broderick, and Michael Jordan).

I keep saying "on average", so what do I mean by that? Basically, VSBC looks at how well the variational posterior is calibrated by computing the distribution of $p_j = P(\theta < \tilde{\theta}_j \mid y_j)$ where y_j is simulated from the model with parameter θ_j that is itself drawn from the prior distribution. If the variational inference method is *calibrated*, then Cook *et al.* showed that the histogram of p_j should be uniform.

This observation can be generalized using insight from the forecast validation community: *if the histogram of p_j is asymmetric, then the variational posterior will be (on average over data drawn from the model) biased.* In the paper, we have a specific result, which shows that this insight is exactly correct if the true posterior is symmetric, and approximately true if it's fairly symmetric.

There's also the small problem that if the model is badly mis-specified, then it may fit the observed data much worse or better than the average of data drawn from the model. Again, this contrasts with the PSIS diagnostic that only assesses the fit for the particular data set you've observed.

In light of this, we recommend interpreting both of our heuristics the same way: conservatively. If either heuristic fails, then we can say the variational posterior is poorly behaved in one of two specific ways. If either or both heuristics pass, then we can have some confidence that the variational posterior will be a good approximation to the true distribution (especially after a PSIS or second-order correction), but this is still not guaranteed.

Faith/Void

To close this post out symmetrically (because symmetry indicates a lack of bias), let's go back to a different Bill Callahan song to remind us that even if it's not the perfect song, you can construct something beautiful by leveraging formal structure:

If
If you
If you could
If you could only
If you could only stop
If you could only stop your
If you could only stop your heart
If you could only stop your heart beat
If you could only stop your heart beat for
If you could only stop your heart beat for one heart
If you could only stop your heart beat for one heart beat

Filed under Bayesian Statistics, Stan, Statistical computing, Uncategorized | Permalink

5 Comments

1. *Sameera Daniels* says:
   February 7, 2018 at 6:46 pm

   Andrew, Andrew, Andrew You just gave me a theme for my dance routine.

- *Dan Simpson* says:
  February 7, 2018 at 6:48 pm

  Soooooo not Andrew :p

2. *Max* says:
   February 7, 2018 at 7:14 pm

   The ‘missing’ line “If you could only stop your heart beat for one” makes me nervous for some reason.

   - *Dan Simpson* says:
     February 7, 2018 at 7:23 pm

     It works surprisingly well in the song (which is beautiful and not at al stressful)

     - *Ben Goodrich* says:
       February 7, 2018 at 7:37 pm

       It happens to me from time to time when I sleep. Waking up thinking that you are dead because your heart has stopped beating is stressful because it takes a few seconds for your oxygen-deprived brain to update its posterior probability that you are dead given that you are awake.