Notes on pseudo-marginal methods, variational Bayes and ABC

Christian Andersson Naesseth
October 3, 2016

The Pseudo-Marginal Framework

Assume we are interested in sampling from the posterior distribution defined by

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \tag{1}$$

for some parameters θ and observations y. Furthermore, assume that we can not directly evaluate $p(y|\theta)$ but we have access to a function, $\hat{g}(u; y, \theta)$ (that we can evaluate pointwise), that takes some random numbers ("seed" if you will) $u \sim r(u|\theta)$ and outputs an unbiased (and non-negative) estimate of $p(y|\theta)$. This means

$$\mathbb{E}_{r(u|\theta)}\left[\hat{g}(u; y, \theta)\right] = \int \hat{g}(u; y, \theta) r(u|\theta) du = p(y|\theta),$$

which is the setting of the pseudo-marginal methods [Andrieu and Roberts, 2009]. Now we define a joint target distribution over θ and u as follows

$$\pi(\theta, u) = \frac{\hat{g}(u; y, \theta) r(u|\theta) p(\theta)}{p(y)}, \tag{2}$$

which we can see is proper (integrates to one) and has the posterior as its marginal distribution, i.e.

$$\pi(\theta) = \int \frac{\hat{g}(u; y, \theta) r(u|\theta) p(\theta)}{p(y)} du = \frac{p(y|\theta)p(\theta)}{p(y)} = p(\theta|y).$$

This means that if we design (almost) any approximate inference method and apply it to (2), the marginal of that approximation will also be an approximation to the true target distribution. For example, if we design a Markov chain (θ^i, u^i) that has (2) as its stationary distribution, we get an approximation to the posterior (1) by throwing away u^i and keeping θ^i. Another example would be to fit a variational approximation $q(\theta|\lambda)q(u|\eta, \theta)$ and then maximize the evidence lower bound (ELBO)

$$\int q(\theta|\lambda)q(u|\eta, \theta) \log \frac{\hat{g}(u; y, \theta) r(u|\theta) p(\theta)}{q(\theta|\lambda)q(u|\eta, \theta)} d\theta du,$$

with respect to the variational parameters λ, η. The easiest thing to do, since we do not necessarily know how to evaluate $r(u|\theta)$ point-wise, is to set $q(u|\eta, \theta) = r(u|\theta)$ and the ELBO simplifies

$$\int q(\theta|\lambda) r(u|\theta) \log \frac{\hat{g}(u; y, \theta) p(\theta)}{q(\theta|\lambda)} d\theta du.$$

This is what is called Variational Bayes with Intractable Likelihood (VBIL) [Tran et al., 2015], leveraging pseudo-marginal ideas in variational inference. Note that this is really just standard variational Bayes on the extended space of θ and u with a specific choice of variational approximation. After optimizing with respect to λ we get an approximation to the true posterior (1), i.e. $q(\theta|\lambda^*)$.

As for using Metropolis-Hastings we would need a proposal, e.g. $q(\theta'|\theta) r(u'|\theta')$, and the acceptance step becomes minimum of 1 and

$$\frac{\hat{g}(u'; y, \theta') r(u'|\theta') p(\theta')}{\hat{g}(u; y, \theta) r(u|\theta) p(\theta)} \frac{q(\theta|\theta') r(u|\theta)}{q(\theta'|\theta) r(u'|\theta')} = \frac{\hat{g}(u'; y, \theta') p(\theta')}{\hat{g}(u; y, \theta) p(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}.$$

We could also consider doing EP, INLA, Laplace approximation, sequential Monte Carlo, or any other type of approximate Bayesian inference that you can come up with, using (2) as our target distribution. The marginal distribution over θ would be an approximation to the original posterior.

Approximate Bayesian Computation

In approximate Bayesian computation we (typically) assume we can simulate from the likelihood for a given parameter value, however, we can not evaluate its density point-wise. We solve this by introducing a systematic bias into the model, replacing $p(y|\theta)$ in (1) by

$$p_\epsilon(y|\theta) = \int K_\epsilon(y, x) p(x|\theta) dx,$$

where K_ϵ is a kernel, e.g. $\exp\left\{-\frac{1}{2\epsilon^2}(y - x)^2\right\}$, and $p(x|\theta)$ is our data *simulator*. This means we are now actually interested in approximating the "likelihood-free" (or ABC) posterior

$$p^{\text{LF}}(\theta|y) \propto p_\epsilon(y|\theta) p(\theta) = p(\theta) \int K_\epsilon(y, x) p(x|\theta) dx. \tag{3}$$

This is a systematic error we make that no matter how well we approximate (3) it will never be exactly equal to the true posterior (1) for any non-zero ϵ.

It might seem like we have not really gained anything because to evaluate $p_\epsilon(y|\theta)$ point-wise we evaluate an intractable integral. Well, we can easily generate an unbiased estimate of it by simulating from $x \sim p(x|\theta)$ and returning

$K_\epsilon(y, x)$. This is straightforward to extended to the many-sample case

$$\frac{1}{N}\sum_{i=1}^{N} K_\epsilon(y, x^i), \quad x^i \sim p(x|\theta),$$

which fits very nicely within the pseudo-marginal framework! That is we can identify

$$\hat{g}(u; y, \theta) = \frac{1}{N}\sum_{i=1}^{N} K_\epsilon(y, x^i),$$
$$u := x^{1:N}, \quad r(u|\theta) = \prod_{i=1}^{N} p(x^i|\theta).$$

However, note that now $\hat{g}$ is an unbiased estimate of $p_\epsilon(y|\theta)$ and not $p(y|\theta)$. We can still use the pseudo-marginal framework with the understanding that our approximation will now have the likelihood-free posterior $p^{\text{LF}}(\theta|y)$ as its marginal. For completeness this is the new target distribution

$$\pi(\theta, u) = \pi(\theta, x^{1:N}) \propto p(\theta)\frac{1}{N}\sum_{i=1}^{N} K_\epsilon(y, x^i) \cdot \prod_{i=1}^{N} p(x^i|\theta). \tag{4}$$

Connection with Automatic Variational ABC

We can now show the correctness of the approach taken by Moreno et al. [2016]. It can be interpreted as variational Bayes (i.e. VBIL) with the reparameterization trick on both $q(\theta|\lambda)$ and $r(u|\theta)$ for the target distribution in (4). Note that if we do not want to use $q(u|\theta, \eta) = r(u|\theta)$, Moreno et al. [2016] do not in some of the experiments, we can work under the assumption that $x = f(\omega, \theta),\ \omega \sim m(\omega)$, i.e. reparameterization of $p(x|\theta)$. We now maximize the following ELBO

$$\int q(\theta|\lambda)\prod_{i=1}^{N}\left[q(\omega^i|\theta,\eta)\right] \cdot \log \frac{p(\theta)\frac{1}{N}\sum_{i=1}^{N} K_\epsilon(y, f(\omega^i, \theta))\prod_{i=1}^{N} m(\omega^i)}{q(\theta|\lambda)\prod_{i=1}^{N} q(\omega^i|\theta,\eta)} d\theta d\omega^{1:N},$$

and then apply the reparameterization trick on both $q(\theta|\lambda)$ and $q(\omega^i|\theta, \eta)$. We need to assume that we can evaluate (and differentiate) $m(\omega)$ point-wise. Again, it can be interpreted and motivated as standard variational Bayes on the extended space of θ and $\omega^{1:N}$. The pseudo-marginal framework now says that $q(\theta|\lambda)$ is an approximation for the likelihood-free posterior, see Tran et al. [2015] for details and assumptions for this to hold.

Note that this is not exactly how they motivate it in the paper, however, it is how I would justify what they do.

Connection with Hamiltonian ABC

Assume again that we can further reparameterize the simulator as $x = f(\omega, \theta)$, $\omega \sim m(\omega)$, i.e. this is equivalent with $x \sim p(x|\theta)$. This means we can still use the pseudo-marginal framework but now on the extendend space $(\theta, \omega^{1:N})$ with the target distribution

$$\pi(\theta, \omega^{1:N}) = \frac{p(\theta)\frac{1}{N}\sum_{i=1}^{N} K_\epsilon\left(y, f(\omega^i, \theta)\right)}{p_\epsilon(y)} \prod_{i=1}^{N} m(\omega^i), \tag{5}$$

and it retains the likelihood-free posterior (3) as its marginal over θ. In the Hamiltonian ABC paper [Meeds et al., 2015] they then alternate between sampling $\theta|\omega^{1:N}$ and $w^{1:N}|\theta$ using HMC and another Markov chain method (flipping), respectively. That should mean we need to take gradients of the following with respect to parameters

$$\nabla_\theta \log p(\theta) + \nabla_\theta \log \left(\sum_{i=1}^{N} K_\epsilon\left(y, f(\omega^i, \theta)\right) \right). \tag{6}$$

If we could calculate the above gradient exactly, which they assume in Moreno et al. [2016], we could just use "standard" HMC for updating $\theta|\omega^{1:N}$. In fact conditional on $\omega^{1:N}$ there is no extra randomness arising from the likelihood-simulator! So applying a finite difference approximation to the gradient of the second term in (6) introduces an error for sure, but not really any stochastic error for which we would need the stochastic gradient Monte Carlo framework. The noise I assumed came in through the use of SPSA to approximate the gradient, an *almost* unbiased estimate of the second gradient term above. Together with the stochastic gradient Monte Carlo methods described in the paper it should result in something that more (or less?) samples from the likelihood-free posterior (3).

Another important point here is that (6) might seem to have an unbiased approximation of the ABC likelihood that we have to evaluate through a logarithm. However, because we are using the pseudo-marginal approach this is just part of our target distribution!

References

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2), 04 2009.

E. Meeds, R. Leenders, and M. Welling. Hamiltonian abc. In *31st Conference on Uncertainty in Artificial Intelligence*, 2015.

A. Moreno, T. Adel, E. Meeds, J. M. Rehg, and M. Welling. Automatic Variational ABC. *arXiv:1606.08549*, June 2016.

M.-N. Tran, D. J. Nott, and R. Kohn. Variational Bayes with Intractable Likelihood. *arXiv:1503.08621*, March 2015.