Particle Filtering with Optimal Proposal Distribution Estimated via Variational Inference

Gibson, Reich, and Ray in some order

December 3, 2017

Introduction

All of the text below is borderline plagiarism of Wikipedia; revise heavily.

Particle filtering is an approach to estimating the posterior distribution of the state variables given the observed variables in a hidden Markov model. In turn, this estimate of the distribution of the state variables can be used to obtain Monte Carlo approximations to the likelihood function of a statistical model, and can therefore be used as a step in parameter estimation in either a frequentist or Bayesian framework.

There are many variations on particle filtering algorithms, but most of these follow a broad outline in which two steps are iterated:

1. **Predict:** Predict the value of the state at time $t + 1$, conditional on the observed variables up through time t. The predictive distribution for the value of the state at time $t + 1$ is represented via a sample from that distribution.
2. **Update:** Update the predictive distribution from step 1 to obtain an estimate of the distribution of the state at time $t + 1$, conditional on the observed variables up through time $t + 1$.

Although particle filtering works well in many settings, it can suffer from limitations including:

- particle depletion; the effective size of the sample used to represent the predictive or filtered state distribution can be much smaller than the nominal sample size.
- high Monte Carlo variability; as a consequence (?) of particle depletion, there can be high Monte Carlo variability in the estimates of the unobserved states, and as a result, in the estimates of the likelihood function if particle filtering is being used in a larger parameter estimation procedure.

A variety of strategies have been proposed to address these problems. For example, if the effective sample size is too small, it is common to augment the sample by using resampling to obtain additional particles near those particles in the sample that have large weight (cite Kitagawa 1996, clarify).

The proposal distribution used in the predict step also has important consequences for the variance of the filtered distribution estimate. It has also been shown (Del Moral, 1996, 1998) that the optimal proposal distribution (in terms of minimizing the variance of the filtered distribution estimate) is the target distribution. Although this target distribution is typically unknown, common schemes involve approximating it by an appropriately weighted sample from the transition density of the hidden Markov model.

In this work, we suggest another solution to this problem, in which the optimal proposal distribution for a particle filter is estimated via variational inference. This imposes some additional computational costs on the particle filter, but we show (hopefully. . .) that this approach mitigates the problem of particle depletion and reduces the Monte Carlo variability of the filtered state distribution estimates.

Method Derivation

Let X_t, $t = 1, \dots, T$ denote an unobserved state vector at each time t. For now, this state has to be continuous but we might be able to discretize later?

Let Y_t, $t = 1, \ldots, T$ denote an observed value at each time t.

The details about when we start observing the Y_t's relative to the first X_t are unimportant.

For now, we're just going to write down the method to evaluate the likelihood for a fixed set of parameters. This is not explicitly Bayesian or frequentist.

But we could likely differentiate the approximation to the likelihood derived here with respect to parameters θ and use that to do maximum likelihood or Bayesian inference.

Model Structure

States:

- $X_1 \sim g_1(x_1; \xi)$
- $X_t | X_{t-1} \sim g(x_t | x_{t-1}; \xi)$ for all $t = 2, \ldots, T$

Observations:

- $Y_t | X_t \sim h(y_t | x_t; \zeta)$

Here, $g_1(\cdot)$ and $g(\cdot)$ are appropriately defined probability density functions depending on parameters ξ and $h(\cdot)$ is an appropriately defined probability density function or probability mass function depending on parameters ζ.

Define $\theta = (\xi, \zeta)$ to be the full set of model parameters.

Evaluating the Likelihood via Filtering

Our goal (for now) is to evaluate the likelihood function. We develop a variant on sequential importance resampling for evaluating the likelihood here.

$$\begin{aligned}
L(\theta | y_{1:T}) &= f(y_{1:T}; \theta) \\
&= f(y_1; \theta) \prod_{t=2}^{T} f(y_t | y_{1:t-1}; \theta) \\
&= \int_{x_1} f(y_1, x_1; \theta) dx_1 \prod_{t=2}^{T} \int_{x_t} f(y_t, x_t | y_{1:t-1}; \zeta) dx_t \\
&= \int_{x_1} f(y_1 | x_1; \zeta) f(x_1; \xi) dx_1 \prod_{t=2}^{T} \int_{x_t} f(y_t | x_t, y_{1:t-1}; \zeta) f(x_t | y_{1:t-1}; \xi) dx_t \\
&= \int_{x_1} f(y_1 | x_1; \zeta) f(x_1; \xi) \frac{\pi(x_1)}{\pi(x_1)} dx_1 \prod_{t=2}^{T} \int_{x_t} f(y_t | x_t; \zeta) f(x_t | y_{1:t-1}; \xi) \frac{\pi(x_t | y_{1:t-1})}{\pi(x_t | y_{1:t-1})} dx_t \\
&\approx \sum_{x_1^{(k)}} \left\{ w_1^{(k)} f(y_1 | x_1^{(k)}; \zeta) \frac{f(x_1^{(k)}; \xi)}{\pi(x_1^{(k)})} \right\} \prod_{t=2}^{T} \sum_{x_{t|t-1}^{(k)}} \left\{ w_t^{(k)} f(y_t | x_t^{(k)}; \zeta) \frac{f(x_t^{(k)} | y_{1:t-1}; \xi)}{\pi(x_t^{(k)} | y_{1:t-1})} \right\}, \text{ where}
\end{aligned}$$

$(w_1^{(k)}, x_1^{(k)})$, $k = 1, \ldots, K$ are a weighted sample from the distribution $\pi(x_1; \xi)$ and $(w_t^{(k)}, x_t^{(k)})$ are a weighted sample from $\pi(x_t | y_{1:t-1}; \xi)$. Here, $\pi(x_1; \xi)$ and $\pi(x_t | y_{1:t-1}; \xi)$ are the proposal distributions (also referred to as importance distributions) for importance sampling estimates of

$$\int_{x_t} f(y_t|x_t;\zeta) f(x_t|y_{1:t-1};\xi) dx_t = E_{X_t|y_{1:t-1}}[f(y_t|X_t;\zeta)] \tag{1}$$

Going forward, for notational simplicity we will focus on the proposal at step t.

It is known that the optimal proposal distribution (in terms of minimizing the variance of the estimated expected value in Equation (1)) has the form $\pi(x_t|y_{1:t-1};\xi) \propto f(y_t|x_t;\zeta) f(x_t|y_{1:t-1};\xi)$, where the proportionality is as a function of x_t. Note that this optimal proposal distribution is actually the conditional distribution $f(x_t|y_{1:t};\xi)$.

We propose the following filtering algorithm:

Algorithm:

In step t, suppose we have a weighted sample $(w_{t-1}^{(k)}, x_{t-1}^{(k)}) \sim f(x_{t-1}|y_{1:t-1};\xi)$

1. **Estimate Proposal Distribution:**
   a. Use variational inference methods such as SVGD or BBVI to obtain a distributional estimate of the optimal proposal distribution

$$\begin{aligned}\pi(x_t|y_{1:t-1};\xi) &\propto f(y_t|x_t;\zeta) f(x_t|y_{1:t-1};\xi) \\ &\approx f(y_t|x_t;\zeta) \sum_{i=1}^{K} w_{t-1}^{(k)} f(x_t|x_{t-1}^{(k)})\end{aligned}$$

   b. Possibly, add a defensive mixture component to the estimated optimal proposal that's an estimate of $f(x_t|x_{t-1}) \approx \sum_{i=1}^{K} w_{t-1}^{(k)} f(x_t|x_{t-1}^{(k)})$, itself obtained as a mixture based on the sample from the previous step.
2. **Sample from the Proposal:** Draw a sample $x_t^{(k)} \sim \pi(x_t|y_{1:t-1};\xi)$, $k = 1, \ldots, K$ from the proposal distribution developed in step 1.
3. **Likelihood Estimation:** Use the sample from step 2 to obtain an importance sampling estimate of $\int_{x_t} f(y_t|x_t;\zeta) f(x_t|y_{1:t-1};\xi) dx_t$ based on the proposal distribution developed in step 1.
4. **Calculate Sample Weights:** Assign each sample $x_t^{(k)}$ from step 2 a weight $w_t^{(k)}$ as follows (I'm not actually 100% sure that these are the right weights):

$$w_t^{(k)} \propto \frac{f(y_t|x_t^{(k)};\zeta) \sum_{i=1}^{K} w_{t-1}^{(k)} f(x_t|x_{t-1}^{(k)})}{\pi(x_t^{(k)}|y_{1:t-1};\xi)} \approx \frac{f(y_t|x_t^{(k)};\zeta) f(x_t^{(k)}|y_{1:t-1};\xi)}{\pi(x_t^{(k)}|y_{1:t-1};\xi)}$$

Note that in this scheme, there is no formal connection between the consecutive samples $x_{t-1}^{(k)}$ and $x_{t-1}^{(k)}$ for any particular k; this algorithm does not obtain a sample from the joint posterior of the states given the observed data. If we want to maintain chains of samples from the posterior, we could make appropriate adjustments to the proposals in step 1 (so that the proposal depends on $x_{t-1}^{(k)}$), the sampling in step 2, and the weight calculations in step 4. I don't think it's worth the effort and additional computational cost for our purpose though.

In Step 1a, it's important that this proposal distribution have heavy tails in order to ensure that the importance sampling estimate of the likelihood in step 3 below has small Monte Carlo variance. To ensure this, if we are using SVGD, we might want to use a large bandwidth and place normals or t's with a large bandwidth at each particle. If we use something like BBVI, we might want to estimate this with something like a mixture of t distributions, or use normals in the estimation and then replace them with t's. There will likely be a speed trade-off in the quality of this approximation and the number of samples we need to take from it. My guess is that we just need this estimate of the proposal to be roughly ok, and we want fast approximate methods in this step - so, maybe not SVGD.

It may also be that in step 1a, we can just get away with using the default proposal of $\pi(x_t|y_{1:t-1};\xi) = f(x_t|y_{1:t-1};\xi)$ in most cases, and only use an estimated optimal proposal distribution if we see signs that something is going wrong in the particle filter (such as a low effective sample size).

Literature/references to pursue. . . .

Quotes from https://statweb.stanford.edu/~owen/mc/Ch-var-is.pdf

Results from Sections 9.3, 9.10, and 9.11 seem like a potentially promising way forward: proposal is a mixture of distributions from VI and the density $f(x_t|y_{1:t-1};\xi)$. Each mixture component used as a control variate as described there might be too slow though?

“Mixtures have long been used for importance sampling. In a Bayesian context, if the posterior distribution . . . is multimodal then we might attempt to find all of the modes and approximate . . . by a mixture distribution with one component centered on each of those modes. Oh and Berger (1993) use a mixture of t distributions.

Defensive mixtures are from the dissertation of Hesterberg (1988), and are also in the article Hesterberg (1995)."