STAT 677 Project

Patrick Ding

5/12/2019

Introduction

Variational inference is very popular in machine learning, in our view because of its speed and beacause there is less emphasis on uncertainty quantification in machine learning. One major shortcoming of variationl inference methods hampering its adoption in statistics is its tendency to underestimate posterior variance. The mean-field assumption for the approximate posterior also fails to capture correlations in the posterior. Recent work on expanding the variational family seeks to overcome these issues. Here we apply the Semi-Implicit Variational Inference (SIVI) approach of Yin and Zhou (2018) to a spatial Gaussian process regression model. In a simulation study we see that SIVI accurately captures posterior variance and correlation compared to MCMC.

Model

The spatial GP model we consider is

$$
\begin{aligned}
Y &\sim X\beta + w + \epsilon, \\
\beta &\sim N(0, 5I), \\
w &\sim GP(0, \rho(r)), \\
\epsilon &\sim N(0, \tau^2 I), \\
\rho(r) &= \exp(-r/\phi), \\
\phi &\sim U(.01, 10), \\
\tau^2 &\sim IG(2, 1).
\end{aligned}
$$

We are using an exponential covariance function. The parameters for which we want to infer the posterior are (β, ϕ, τ^2).

We impose the following approximate posterior model $q(\beta, \phi, \tau^2)$:

$$
\begin{aligned}
q(\beta|\mu_\beta, \Sigma_\beta) &\sim MultivariateNormal(\mu_\beta, \Sigma_\beta) \\
q(\phi|\mu_\phi, \sigma_\phi) &\sim LogNormal(\mu_\phi, \sigma_\phi) \\
q(\tau^2|\mu_{\tau^2}, \sigma_{\tau^2}) &\sim LogNormal(\mu_{\tau^2}, \sigma_{\tau^2}) \\
\psi = (\mu_\beta, \mu_\phi, \mu_{\tau^2}) &\sim q(\psi)
\end{aligned}
$$

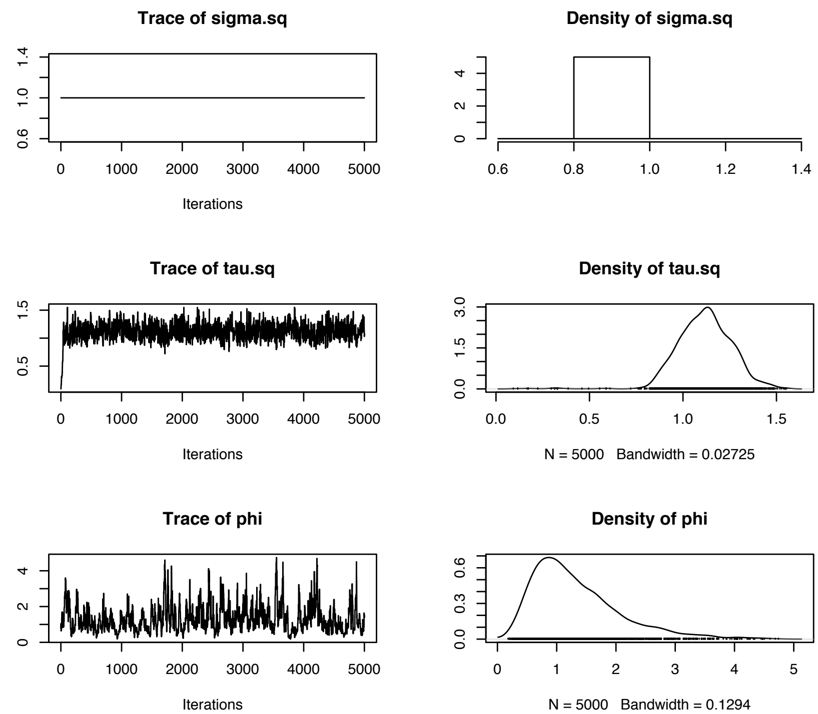
where $q(\psi)$ is an implicit distribution. Samples from $q(\psi)$ are obtained by passing $N(0, I_{50})$ noise through a 3 layer fully connected feed-forward neural network with layer sizes [100, 200, 100] and ReLU activations. The conditional approximate distribution variances $(\Sigma_\beta, \sigma_\phi, \sigma_{\tau^2})$ are variational parameters for which we compute point estimates. We choose log normal conditional approximate posteriors for the spatial GP variance parameters because we need the $q(\cdot|\psi)$ to be reparameterizable. In this case all of these distributions can be reparameterized in terms of a shift and scaling. This choice allows us to use the reparameterization trick (Kingma and Welling 2013) when taking Monte Carlo estimates of the gradient of the surrogate ELBO.

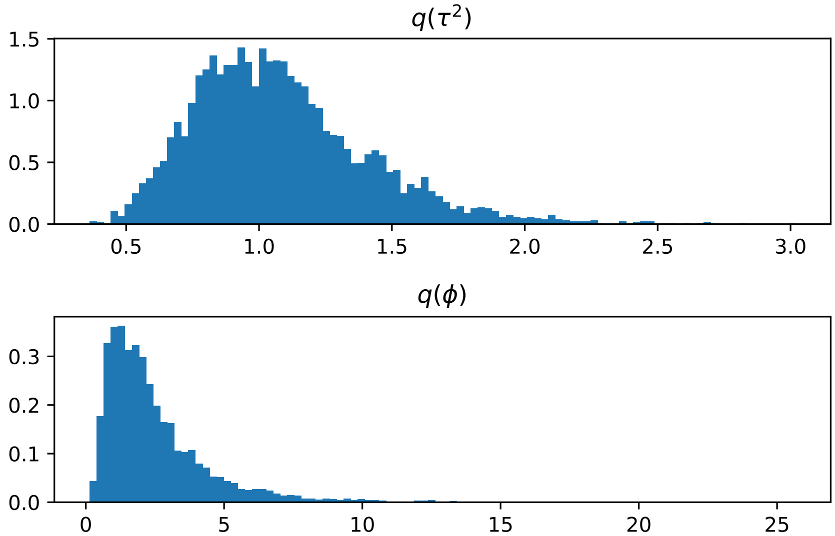
Simulation Result

For the simulation we generate data from the following settings of the parameters:

$$\tau^2 = 1,$$
$$\phi = 5,$$
$$\beta = (1, -5, 10).$$

We sample 200 locations uniformly from $[0,1]x[0,1]$. For MCMC we run the chain for 5000 steps. For SIVI we take 100 gradient updates, 20 outer Monte Carlo samples, and 5 inner Monte Carlo samples. We use the Adam optimizer in Pytorch to do automatic differentiation. Once we have learned the approximate posterior, we draw 5000 samples from it.





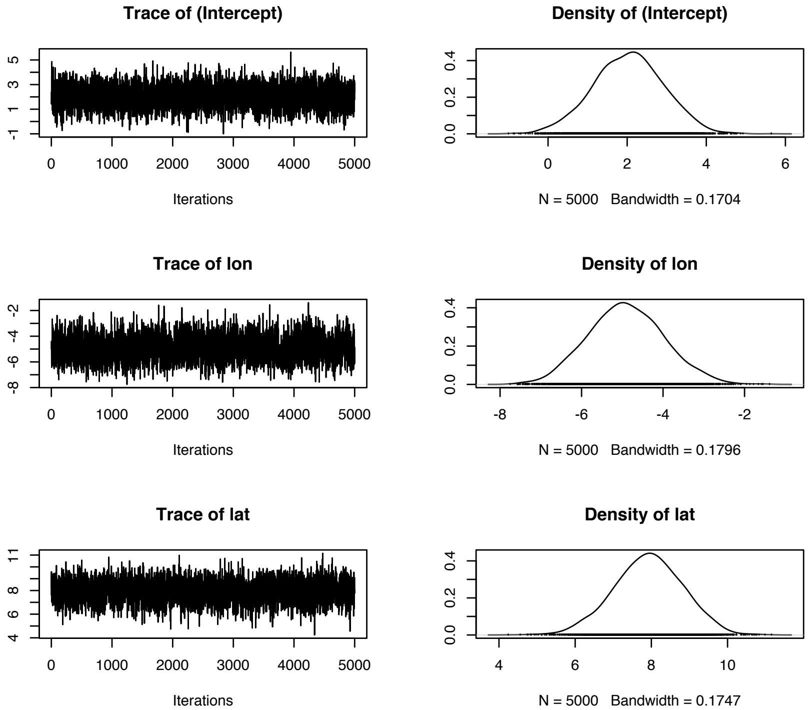
We see that the medians of the posteriors for the variance parameters from both methods are close:

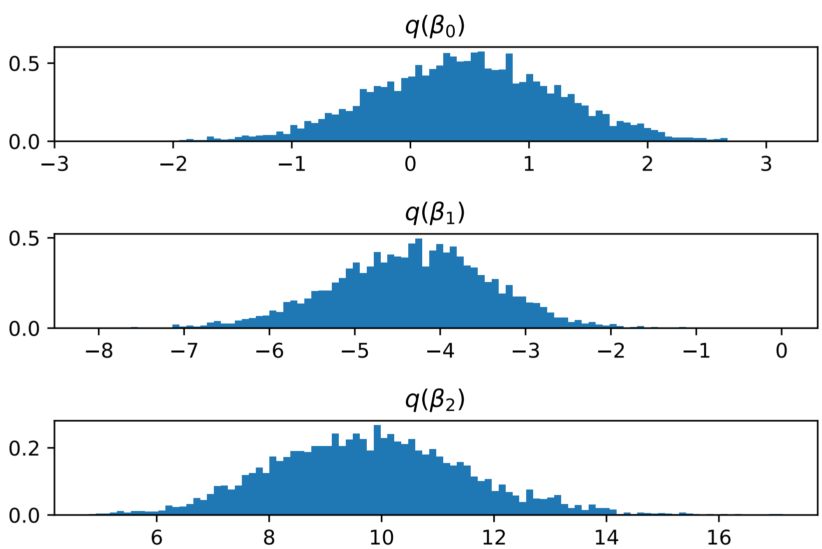
Parameter	MCMC	SIVI
τ^2	1.119	1.0351
ϕ	1.162	1.976

The posterior standard deviation of the variance parameters is larger for SIVI than for MCMC:

Parameter	MCMC	SIVI
τ^2	0.147	0.332
ϕ	0.743	2.001

–





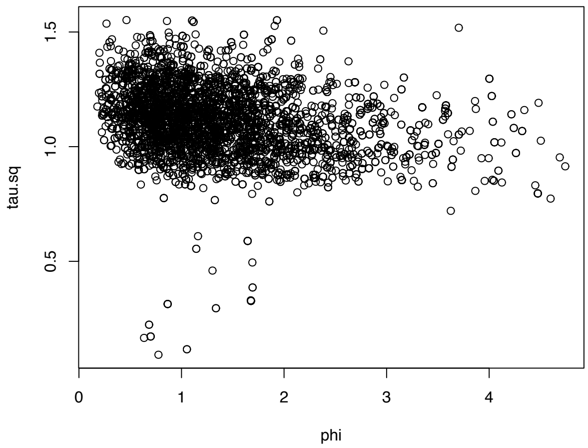
The posterior means of β are close for both MCMC and SIVI:

Parameter	MCMC	SIVI
β_0	2.035	0.471
β_1	-4.906	-4.330
β_2	7.920	9.800

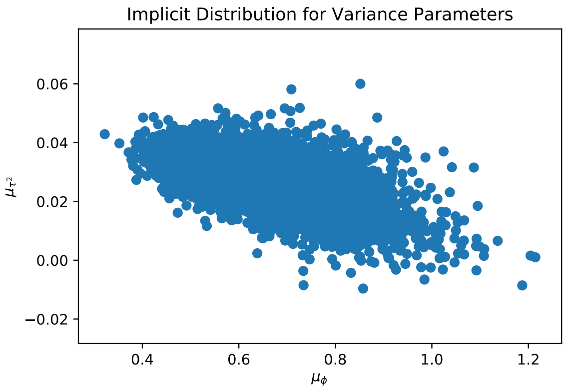
The posterior variances of β from SIVI are of comparable magnitude as that from MCMC. SIVI has managed to not underestimate the posterior variance:

Parameter	MCMC	SIVI
β_0	0.883	0.770
β_1	0.931	0.922
β_2	0.912	1.717

Comparing a scatter plot of τ^2 and ϕ samples from MCMC, we see there is some correlation between the two variables in the posterior:



Looking at the scatter plot of samples for the mean parameters for those two variables from the mixing distribution in SIVI, we see that SIVI has managed to capure this same correlation:



Discussion

I was pressed for time so could not run additional experiments. One issue was that had problems with the covariance matrix being not positive semi definite. Though SIVI is powerful it requires parameter tuning and a decent amount of coding.

References

Kingma, Diederik P., and Max Welling. 2013. “Auto-Encoding Variational Bayes.” *arXiv:1312.6114 [Cs, Stat]*, December. http://arxiv.org/abs/1312.6114.

Yin, Mingzhang, and Mingyuan Zhou. 2018. “Semi-Implicit Variational Inference.” In *Proceedings of the 35 Th International Conference on Machine Learning*, 16. Stockholm, Sweden.