ST433/533 Applied Spatial Statistics

Lab activity for 9/16/2020

A. CLARIFICATION QUESTIONS

(1) Is the code amcmc <- list("n.batch"=n.samples/100, "batch.length"=100, "accept.rate"=0.43) the same as the accept rate in Metroplis sampling?

Yes, these commands say to adapt the algorithm every 100 iterations in an attempt to get the Metropolis acceptance rate to be approximately 43%.

(2) In Spatial prediction video lec07, the Covariance function is mentioned several times. Which part of the formulas shown in lec7 is the covariance function?

The spatial covariance function is used to fill in the elements of the covariance matrix Sigma. For example, if we decide on an exponential covariance function and the distance between points i and j is d_{ij}, then the [i,j] element of Sigma is sigma2*exp(d_{ij}/phi).

(3) It was not clear for me how MCMC estimated the uncertainty of the carbon data.

MCMC gives draws form the posterior distribution, which represents uncertainty about parameters/predictions after observing the data. So rather than a single estimate/prediction, you get the whole uncertainty distribution.

(4) How can we have previous information to set the priors for MCMC?. How much affect the model if we don't have them. That can cause poor convergence?

Usually we don't have prior information and things work out fine. Potential sources of prior information include previous and related studies or expert opinion. But 99% of the time you do not have solid prior information except maybe rough bounds on parameters, like maybe the spatial range is between 0 and the maximum distance between sites in the study area.

(5) The convergence performance is only evaluated visually? there is not a performance score or any other criteria?

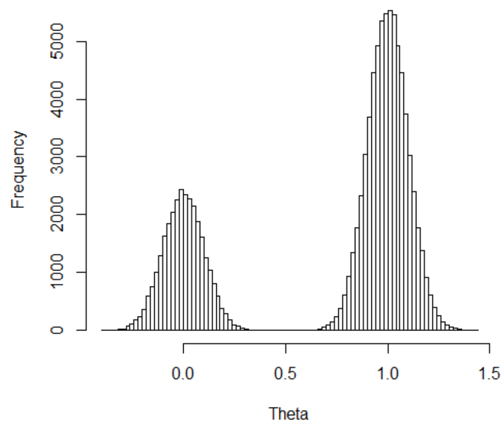
There are a gazillion formal diagnostics, but honestly, you can usually tell by looking at the trace plots whether convergence has been achieved. These formal diagnostics are useful when there are many parameters or when you are trying to automate the process.

(6) What to do if only half parameters have good convergence? what means if a parameter converges but has a large range?

All parameters need to converge before you are done sampling. If the posterior range is wide it likely just means that the data do not provide a lot of information about this parameter, e.g., the data is small or the parameter really isn't that important for model fit.

(7) In lecture, Dr. Reich stated that MCMC is guaranteed to find the initial distribution if it is allowed to run long enough but would that be true for a split distribution where half the values are centered around 0 and the other half centered a distant, different value? Wouldn't it depend on the initial value location for which distribution half it converged to? Or are such distributions not applicable to this example?

Below is an example of this pathological case. There are some technical conditions (reversible chains etc) that must be satisfied, but say Metropolis sampling with Gaussian candidate distribution always satisfies these conditions in theory. In practice though, MCMC would not work super well in this case.



(8) In lecture 8, Dr. Reich states that when prior information is weak; MLE and Bayesian give similar results but that the interpretation is different. Could you clarify?

The posterior is the post(theta) = likelihood(theta)*prior(theta) so if the prior is say uniform, prior(theta) = c for all theta, then post(theta) = likelihood(theta)*c and so maximizing the likelihood and maximizing the posterior give the same result. Stat theory also shows that in this case the posterior mean \approx MLE and the posterior SD \approx standard error.

(9) Generally speaking, when MCMC returns the posterior parameter values to you (say, theta), does MCMC always return the parameters in their alphabetical order? For example from the Bayesian geostatistical analysis for the air pollution data, in this case, 1st column: sigma.sq; 2nd column: tau.sq; 3rd column: phi?

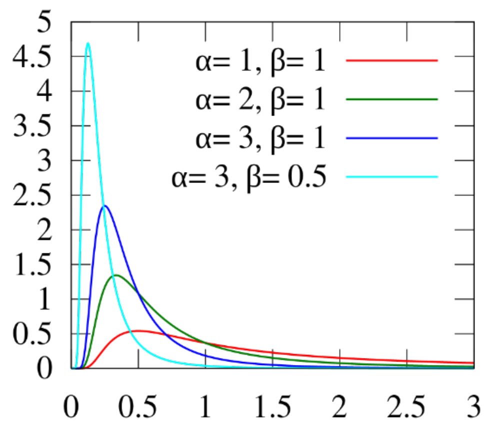
The order of parameters is arbitrary in MCMC, but some packages may choose to alphabetize the parameters for convenience.

(10) Does MCMC convergence assume that the posterior follows a particular family of distribution (i.e. normal)? Or can it converge to any distribution without being told prior information about the posterior. If not, is there a way to check for the presence of a non-normal distribution based on "poor" MCMC results that look for a normal distribution?

No, MCMC does not assume normality. The posterior can be skewed, bimodal (as above) or even discrete (only taking a few values). This is one of the big advantages of MCMC over some frequentist methods that are only able to approximate the standard error if the sample size is large enough to evoke the central limit theorem and thus assume normality.

(11) Are there other reasons to choose the Inverse Gamma distribution for the variance priors other than the fact that the distribution forces the values to be positive?

Yes, in some case it is a conjugate prior, but we won't get into this level of detail. I should probably say a bit more about the inverse gamma prior though. If Y ~ Gamma(a,b) then X=1/Y is InvGamma(a,b). The two parameters are the shape, a, and the scale, b. Here are some examples of this from the wiki page.



(12) What is the benefit of using Bayesian Kriging when were just drawing from the posterior dataset as our prior knowledge, as done in the example?

Let's address this one in the student discussion questions.

(13) What does the c^2*I term represent in the normal function? What is an example?

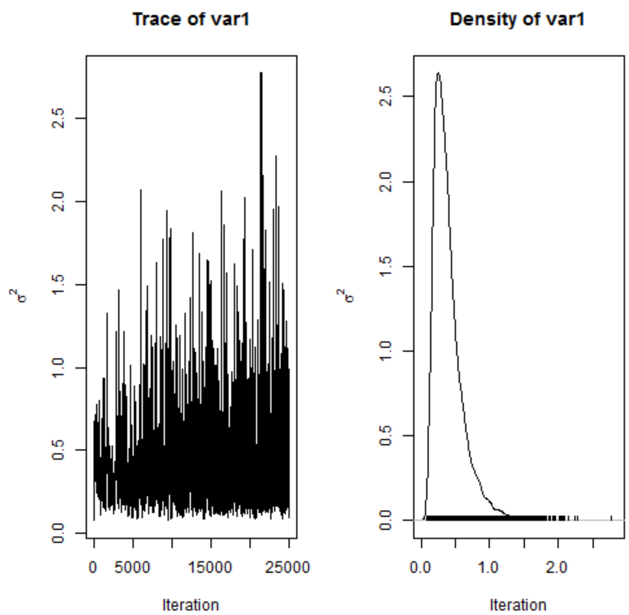
This means that the covariance is the scalar c^2 times the identity matrix. The identify matrix has ones on the diagonal and zeros everywhere else, so c^2*I has c^2 on the diagonal and zeros everywhere else. In other words, the variance of all observations is c^2 and all correlations are zero.

(14) It looks like we decide the number of iterations for MCMC based on the trace plot, but I was wondering if there is a way to decide the number of iterations beforehand? How did we decide to use 25000 in the example? Will it depend on the number of parameters we are considering, the complexity of the posterior distribution?

It's hard to set a firm number before doing any preliminary MCMC runs because time to convergence depends on the model and data in ways that are hard to predict. But tens of thousands is usually a good starting point.

(15) At the end of lecture 9, we look at an example wherein we see trace plots and density plots. I was a little confused with the density plot. It seemed to be labelled as \sigma^2 vs iterations, shouldn't it be frequency vs \sigma^2?

Here is the image. Yes, the x-axis should be labeled “sigma^2” and the y-axis should be “Density”.



(16) Can "Halton Draw" be used instead of MCMC draws to find prediction/estimations from the posterior distribution?

I’m not sure what this is, sorry.

(17) We take n draws and have the mean of these n draws as the estimated value from the posterior distribution. How many draws is the optimum number of draws to have representative/good estimates from the posterior distribution?

Optimal is n=infinity, so you have to balance time versus approximation accuracy. A simple check is to run MCMC a few times and see how much the results change.

(18) How can we know if it's proper to use Bayesian analysis or not in that situation?

It is always “proper”, but as we’ll discuss (below), there are cases where it is more beneficial than others.

(19) What do the sigma, tau, phi refer to for the Bayesian estimate?

These are the model parameters, the partial sill, nugget and spatial range.

(20) If we have time, could we review how to interpret the parameters in HW3 in the context of the data? Like we were asked to do in HW3 Q2. I feel like I have a good understanding of how to calculate the parameters, how to "read" them from a variogram, and how they affect the predictions/estimates relatively (like increases or decreases) but not directly.

I just mean to say ~ "The spatial range estimate is 10. This means that mean precipitation is uncorrelated after a distance of 30 kilometers. The partial sill is 5 times as large as the nugget. This means most of the residual variation is spatial and opposed to independent/measurement error."

(21) We talked about how typically with Bayes, we only have one parameter to estimate. What are some methods to try when you have many parameters to estimate?

MCMC can be used with multiple parameters, as in the Bayes Kriging sample code.

B. STUDENT DISCUSSION QUESTIONS

(1) What are the benefits to using normal Kriging and what are the benefits of using Bayesian Kriging? When would we rather use the classical Kriging method over the Bayesian method?

Advantages of Bayesian Kriging are that it can be used for non-Gaussian data and uncertainty quantification. Can incorporate priors for new situations/small datasets.

Advantages of classical Kriging are that it's faster and doesn't require a prior.

(2) You mentioned GLMs are one reason to use Bayesian models. What are some examples of actual data that have non-Gaussian distributions in their outcomes, that have spatial dependence?

Basketball shots made (1) and missed (0), kids getting sick (1) or healthy (0), maybe precip is very right skewed not Gaussian.

(3) Let's say we measure air pollution once each day at sensors across the country. We use Bayesian Kriging to get posterior distributions and credible intervals for our spatial predictions using data from day d. Would it be reasonable (and possible) to use those posterior distributions from day d as our prior distributions when we go to estimate posterior distributions using data from day d+1?

Yes! (BR: sequential analysis; Kalman filter)

(4) Which factors that affect the posterior distribution's sensitivity to the prior distribution, e.g. magnitude of parameters versus similarity to true distribution, etc.

When the prior very different from the likelihood; when the small size is small; when the number parameters is large.

(5) Does the prior distribution in the Bayesian method matters very much? It seems to work well when the prior distribution is uninformative.

Yes, the prior distribution matters, especially for informative priors. Good idea to try different priors to see how much the prior matters.

(6) When n -> infinity (data is large enough) the MLE is almost the same as the estimate in the Bayesian method. And, when n->infinity, every kind of nugget is similar to normal distribution. In this case, should we prefer the MLE method? Also, in this case, the MLE method and Bayesian method work slowly, what other choices can we use?

Yes, MLE is faster. Wait for mid-term 2 for big n methods.

(7) If MCMC algorithm does not converge on the parameter, what else can we do other than changing the initial values of the parameters, increasing the iteration runs, and switching to a simpler model?

BR: These are all great suggestions! I'd add that you could change the prior to be more informative.

(8) If your MCMC is not converging, what information can you use to choose a better initial value?

BR: MLE or variogram

(9) If we use cross validation to compare models, what kind of coverage value is acceptable? In other words, if a model has smallest MSE and AIC, BIC, but the coverage is only about 70%, should we select this model as a good and acceptable one?

BR: This may be depend a bit on your objectives, but I wouldn’t trust a model with 70% coverage.

C. BRIAN'S DISCUSSION QUESTIONS

(1) How might you get informative priors for the analysis in your mid-term exam? What might you use for uninformative priors?

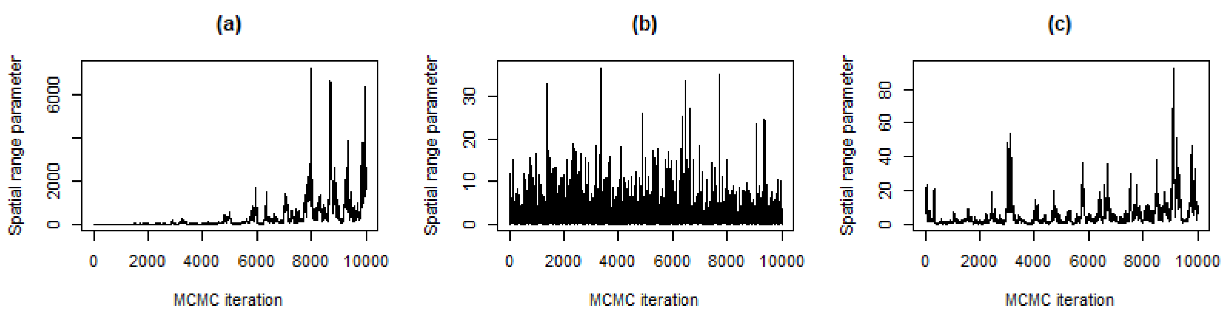
Get informative prior from other states or other years (fit the model for other states, use the posterior from this analysis as the prior for your analysis). Variogram (empirical Bayes).

We could use uninformative prior as in the example from class.

(2) How might you test if the results are sensitive to the choice of prior? Why is this important?

Try 3-5 different priors and see how the posterior changes. This is important because selecting the prior is somewhat subjective and so you want to understand how the subjectivity of the analyst affects the results.

(3) Assess convergence of each MCMC chain. For those that don't converge well, what would you do?

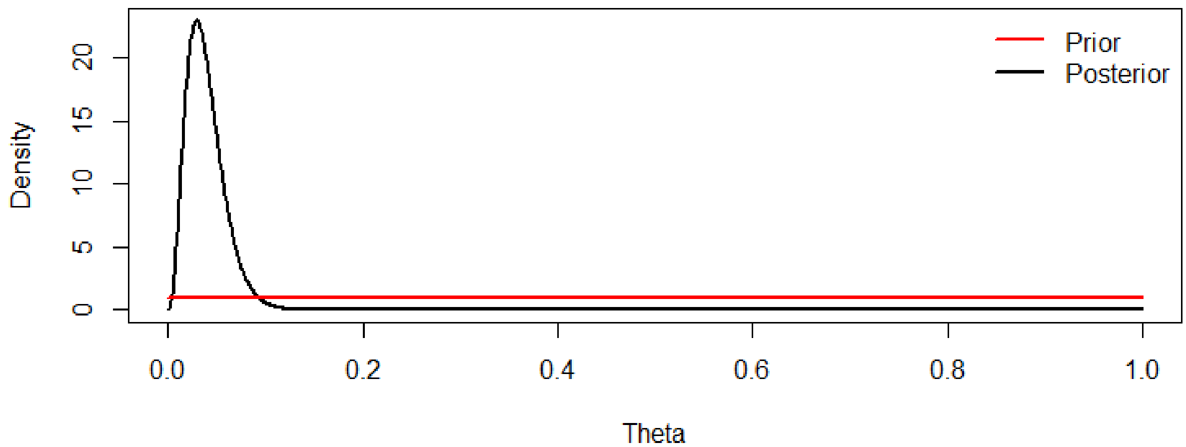


(a) Maybe is started to converge at 8K iterations, run longer.

(b) Looks OK.

(c) Maybe? Needs more iteration.

(4) Say a vaccine will be approved only if it reduces the infection rate by 50% or more. It is known that 20% of the unvaccinated population will become infected. For this study, researchers give a random sample of n=100 individuals a vaccine and find that Y=3 become infected. A Bayesian analysis of these data is Y ~ Binomial(n,theta) and theta ~ Uniform(0,1). The prior and posterior are plotted below.



(a) Summarize (approximately) the posterior of the infection rate, theta, in layman's terms.

Centered at 5% and most likely it is between 0% and 10%.

(b) Does the vaccine reduce the infection rate by 50%? How would you perform this test?

Compute the probability that theta is less than 0.1, and if this probability is greater than 0.95, we conclude that the vaccine is effective.

(5) Let Y be the number of major forest fires in NC in the past N=10 years. Since Y is a count, we assume Y follows a Poisson distribution. We fit the model Y ~ Poisson(N*theta) where theta is the expected number of fires per day. Because theta must be positive, we assume a gamma prior distribution.

(a) The gamma prior has two parameters, a and b. Explore the gamma distribution using these apps

https://shiny.stat.ncsu.edu/bjreich/Poisson_PMF/

https://shiny.stat.ncsu.edu/bjreich/Gamma_PDF/

Select values of a and b that give:

(i) An uninformative (i.e., with large variance) prior distribution for lambda

(ii) An informative prior distribution that restricts there to be between two and four fires per year

(b) Now we observe Y=18 fires in the past decade. Conduct a Bayesian analysis with Poisson likelihood and gamma prior using this app

https://shiny.stat.ncsu.edu/bjreich/PoissonGamma/

For each prior in (a), what the posterior mean and 95% credible set (approximately)

(i)

(ii)