Bayesian data analysis – reading instructions 11

Aki Vehtari

Chapter 11

Outline of the chapter 11

- Markov chain simulation: before section 11.1, pages 275-276
- 11.1 Gibbs sampler (an example of simple MCMC method)
- 11.2 Metropolis and Metropolis-Hastings (an example of simple MCMC method)
- 11.3 Using Gibbs and Metropolis as building blocks (can be skipped)
- 11.4 Inference and assessing convergence (important)
- 11.5 Effective number of simulation draws (important)
- 11.6 Example: hierarchical normal model (skip this)

R and Python demos at https://avehtari.github.io/BDA_course_Aalto/demos.html

- demo11_1: Gibbs sampling
- demo11_2: Metropolis sampling
- demo11_3: Convergence of Markov chain
- demo11_4: potential scale reduction $\hat{R}$

Find all the terms and symbols listed below. When reading the chapter, write down questions related to things unclear for you or things you think might be unclear for others.

- Markov chain
- Markov chain Monte Carlo
- random walk
- starting point
- transition distribution
- jumping / proposal distribution
- to converge, convergence, assessing convergence
- stationary distribution, stationarity
- effective number of simulations
- Gibbs sampler
- Metropolis sampling / algorithm
- Metropolis-Hastings algorithm
- acceptance / rejection rule
- acceptance / rejection rate
- within-sequence correlation, serial correlation
- warm-up / burn-in
- to thin, thinned
- overdispersed starting points
- mixing

- to diagnose convergence
- between- and within-sequence variances
- potential scale reduction, $\hat{R}$
- the variance of the average of a correlated sequence
- autocorrelation
- variogram
- n_{eff}

Basics of Markov chains

Slides by J. Virtamo for the course S-38.143 Queueing Theory have a nice review of the fundamental terms and Finnish translations for them (in English http://www.netlab.tkk.fi/opetus/s38143/luennot/english.shtml and in Finnish http://www.netlab.hut.fi/opetus/s38143/luennot/index.shtml). See specially the slides for the lecture 4. To prove that Metropolis algorithm works, it is sufficient to show that chain is irreducible, aperiodic and not transient.

Animations

Nice animations with discussion http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/

And just the animations with more options to experiment https://chi-feng.github.io/mcmc-demo/

Metropolis algorithm

There is a lot of freedom in selection of proposal distribution in Metropolis algorithm. There are some restrictions, but we don't go to the mathematical details in this course.

Don't confuse rejection in the rejection sampling and in Metropolis algorithm. In the rejection sampling, the rejected samples are thrown away. In Metropolis algorithm the rejected proposals are thrown away, but time moves on and the previous sample x(t) is also the sample x(t+1).

When rejecting a proposal, the previous sample is repeated in the chain, they have to be included and they are valid samples from the distribution. For basic Metropolis, it can be shown that optimal rejection rate is 55–77%, so that on even the optimal case quite many of the samples are repeated samples. However, high number of rejections is acceptable as then the accepted proposals are on average further away from the previous point. It is better to jump further away 23–45% of time than more often to jump really close. Methods for estimating the effective sample size are useful for measuring how effective a given chain is.

Transition distribution vs. proposal distribution

Transition distribution is a property of Markov chain. In Metropolis algorithm the transition distribution is a mixture of a proposal distribution and a point mass in the current point. The book uses also term jumping distribution to refer to proposal distribution.

Convergence

Theoretical convergence in an infinite time is different than practical convergence in a finite time. There is no exact moment when chain has converged and thus it is not possible to detect when the chain has converged (except for rare *perfect sampling* methods not discussed in BDA3). The convergence diagnostics can help to find out if the chain is unlikely to be representative of the target distribution. Furthermore, even if would be able to start from a independent sample from the posterior so that chain starts from the convergence, the mixing can be so slow that we may require very large number of samples before the samples are representative of the target distribution.

If starting point is selected at or near the mode, less time is needed to reach the area of essential mass, but still the samples in the beginning of the chain are not presentative of the true distribution unless the starting point was somehow samples directly from the target distribution.

$\widehat{R}$, effective sample size (ESS, previously n_{eff})

There are many versions of $\widehat{R}$ and effective sample size. Beware that some software packages compute $\widehat{R}$ using old inferior approaches.

The $\hat{R}$ and the approach to estimate effective sample size were updated in BDA3, and slightly updated version of this is described in Stan 2.18+ user guide. Since then we have developed even better $\hat{R}$, ESS (effective sample size with change from n : eff to ESS is due to improved consistency in the notation) in

- Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner (2019). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian analysis*, doi:10.1214/20-BA1221. https://projecteuclid.org/euclid.ba/1593828229.

New $\hat{R}$, ESS, and Monte Carlo error estimates are available in RStan `monitor` function in R, in `posterior` package in R, and in ArviZ package in Python.

Due to randomness in chains, $\hat{R}$ may get values slightly below 1.

Brief Guide to Stan's Warnings https://mc-stan.org/misc/warnings.html provides summary of available convergence diagnostics in Stan and how to interpret them.

Sometimes people write "the number of effective samples" which is wrong (it is possible that notation n_{eff} is partially to blame for this misconception). All the posterior draws in autocorrelated Markov chain are effective, but their efficiency for estimating an expectation depends on the autocorrelation. The effective sample size is not property of individual draws, but joint property of all draws in a sample. Effective sample size also depends on the functional and the effective sample size for a given dependent sample is often different when estimating, for example, $\mathrm{E}[\theta]$ or $\mathrm{E}[\theta^2]$. See more, for exampl, in https://projecteuclid.org/euclid.ba/1593828229.