

Assignment II: Advanced simulation techniques.

Computational Statistics
Instructor: Luiz Max de Carvalho

November 12, 2020

Hand-in date: 23/11/2020.

General guidance

- State and prove all non-trivial mathematical results necessary to substantiate your arguments;
- Do not forget to add appropriate scholarly references *at the end* of the document;
- Mathematical expressions also receive punctuation;
- All computational implementations must be “from scratch”, i.e., you may not employ a ready-made package to implement the technique in question. You may, however (a) employ pre-packaged routines for things like random variate generation and MCMC diagnostics and (b) use a package implementation against which to check your own.
- Please hand in a single PDF file as your final main document. Code appendices are welcome, *in addition* to the main PDF document.

Background

We have by now hopefully acquired a solid theoretical understanding of simulation techniques, including Markov chain Monte Carlo (MCMC). In this assignment, we shall re-visit some of the main techniques in the field of Simulation. The goal is to broaden your knowledge of the field by implementing one of the many variations on the general theme of simulation algorithms.

Each method/paper brings its own advantages and pitfalls, and each explores a slightly different aspect of Computational Statistics. You should pick **one** of the listed papers and answer the associated questions.

In what follows, ESS stands for effective sample size, and is similar to n_{eff} we have encountered before: it measures the number of effectively uncorrelated samples in a given collection of random variates.

Paper 1: MCMC using Hamiltonian dynamics (Neal et al., 2011)

As discussed in class, as the dimensionality of the space over which integrals need to be taken grows, performance suffers massively – this is the so-called “curse of dimensionality”. In our quest to compute expectations efficiently, we might want to draw on all of the available information in order to find pockets of high probability mass.

Clever proposal mechanisms in MCMC use local information, usually in the form of gradients of the (log) target. In this seminal 2011 review, Radford Neal lays out a complete treatment of a technique known as Hybrid or Hamiltonian Monte Carlo (HMC), which works by constructing a Markov chain on an augmented state-space where one considers potentials and momenta.

1. Describe how to apply Hamiltonian dynamics to MCMC;
2. Implementation: reproduce Figure 6 of Neal et al. (2011)
 - (a) Supplement the analyses presented therein with ESS/hour computations in order to gauge the real gain of applying HMC. *Hint:* Use the function `effectiveSize()` from the `coda` package in R (Plummer et al., 2006);
3. Why does HMC avoid random walk behaviour? What advantages are there of such an algorithm?

Paper 2: Bootstrap (Efron and Tibshirani, 1986)

In orthodox (frequentist) Statistics, it is common to want to ascertain long run (frequency) properties of estimators, including coverage of confidence intervals and standard errors. Unfortunately, for the models of interest in actual practice, constructing confidence intervals directly (exactly) is difficult. The bootstrap method is a re-sampling technique that allows for a simple yet theoretically grounded way of constructing confidence intervals and assessing standard errors in quite complex situations.

For this assignment, you are encouraged to consult the seminal 1986 review by stellar statisticians Bradley Efron and Robert Tibshirani (Efron and Tibshirani, 1986).

Hint: Brush off on your Normal theory before delving in. The book by Schervish (2012) – specially Chapter 5 – is a great resource.

1. Define and explain the bootstrap technique;
2. Define and explain the jackknife technique;
3. Implementation:
 - (a) Reproduce the results in Table I of Efron and Tibshirani (1986);
 - (b) Show what happens if one increases/decreases the value of B ;
4. Why is it important to draw exactly n samples in each bootstrap iteration? Can this be relaxed?

5. (bonus) Propose an alternative bootstrap method to the one proposed in the paper and discuss the situations where the new method is expected to perform better.

Paper 3: Blocked Gibbs sampling (Tan and Hobert, 2009)

The so-called Gibbs sampler is a work horse of Computational Statistics. It depends on decomposing a target distribution into conditional densities from which new values of a given coordinate can be drawn.

One of the difficulties one might encounter with the Gibbs sampler is that it might be slow to converge, specially in highly-correlated targets. In Statistics, multilevel models (also called hierarchical or random effects) are extremely useful in modelling data coming from stratified structures (e.g. individuals within a city and cities within a state) and typically present highly correlated posterior distributions.

One way to counteract the correlation between coordinates in the Gibbs sampler is to **block** them together, and sample correlated coordinates jointly.

For this assignment you are referred to the 2009 *Journal of Computational and Graphical Statistics* paper by Tan and Hobert (Tan and Hobert, 2009).

1. Precisely describe the so-called blocked Gibbs sampler; *Hint*: you do not need to describe theoretical properties of the algorithm given in this paper; a general description of the algorithm should suffice.
2. Explain the advantages – both theoretical and practical – of a clever blocking scheme;
3. Would it be possible to apply the “simple” Gibbs sampler in this example? Why?
4. Implementation:
 - (a) Implement the blocked Gibbs sampler discussed in the paper in order to fit the model of Section 1 of Tan and Hobert (2009) to the data described in Section 5 therein.
 - (b) Assess convergence (or lack thereof) and mixing of the resulting chain.
 - (c) Confirm your results agree with those given by the original authors up to Monte Carlo error.
5. Comment on the significance of geometric ergodicity for the blocked Gibbs sampler proposed by Tan and Hobert (2009).

Paper 4: Approximate Bayesian computation (Beaumont et al., 2002)

Bayesian inference relies on computing a posterior distribution of a set of unknowns, θ conditional on the observed data, \mathbf{x} . This posterior distribution,

$p(\boldsymbol{\theta} \mid \boldsymbol{x})$, is proportional to a likelihood function times a prior distribution, i.e.,

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto l(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}). \quad (1)$$

In many situations, however, our models are so complex that the likelihood function in (1) might be either very costly to compute or computationally intractable. Examples of models which fall onto this class include Epidemiological models, Population Genetics models and Gibbs random fields.

In such cases, one can use the so-called likelihood-free methods, which either replace the “true” likelihood function with a surrogate or eschew computing it altogether. The so-called Approximate Bayesian Computation (ABC) class of algorithms has enjoyed great success in recent years because it allows inference about very complex stochastic models that are inaccessible to other methods.

Hint: I strongly suggest you consult the recent review of Beaumont (2019) for extra details.

1. Describe the (basic) ABC rejection algorithm;
2. Implementation:
 - (a) Suppose one has data $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ on a binary outcome, i.e., $x_i \in \{0, 1\}$. Suppose further we choose to model these data as independent $x_i \sim \text{Bernoulli}(\theta)$ and pick a Beta prior for θ with hyperparameters $\alpha > 0$ and $\beta > 0$. Implement an ABC scheme to sample (approximately) from the corresponding posterior, $p(\theta \mid \boldsymbol{x})$.
 - (b) Implement a Metropolis-Hastings scheme to sample from $p(\theta \mid \boldsymbol{x})$;
 - (c) Compare the results of the previous two items to the exact posterior distribution (we derived this in class): how well does ABC fare for a range of true values of θ ? Does performance change if one changes the sample size (n)?
3. Is it possible to employ improper priors with ABC?¹
4. What is the role of sufficient statistics in ABC? *Hint:* Take a look at Robert et al. (2011).

¹Recall that a prior distribution, $\pi(\theta)$ is said to be *improper* if $\int_{\Theta} \pi(t) d\mu(t) = \infty$.

Bibliography

- Beaumont, M. A. (2019). Approximate Bayesian computation. *Annual review of statistics and its application*, 6:379–403.
- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.
- Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.
- Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov chain Monte Carlo*, 2(11):2.
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.
- Robert, C., Marin, J.-M., and Pillai, N. S. (2011). Why approximate Bayesian computational (ABC) methods cannot handle model choice problems. *arXiv preprint arXiv:1101.5091*.
- Schervish, M. J. (2012). *Theory of Statistics*. Springer Science & Business Media.
- Tan, A. and Hobert, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. *Journal of Computational and Graphical Statistics*, 18(4):861–878.