Bayesian data analysis – reading instructions ch 10

Aki Vehtari

Chapter 10

Outline of the chapter 10

- 10.1 Numerical integration (overview)
- 10.2 Distributional approximations (overview, more in Chapter 4 and 13)
- 10.3 Direct simulation and rejection sampling (overview)
- 10.4 Importance sampling (used in PSIS-LOO discussed later)
- 10.5 How many simulation draws are needed? (Important! Ex 10.1 and 10.2)
- 10.6 Software (can be skipped)
- 10.7 Debugging (can be skipped)

Sections 10.1-10.4 give overview of different computational methods. Some of then have been already used in the book.

Section 10.5 is very important and related to the exercises.

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

- demo10_1: Rejection sampling
- demo10_2: Importance sampling
- demo10_3: Sampling-importance resampling

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.

- unnormalized density
- target distribution
- log density
- overflow and underflow
- numerical integration
- quadrature
- simulation methods
- Monte Carlo
- stochastic methods
- deterministic methods
- distributional approximations
- crude estimation
- direct simulation
- grid sampling
- rejection sampling
- importance sampling
- importance ratios/weights

Numerical accuracy of compute arithmetic

Many models use continuous real valued parameters. Computers have finite memory and thus the continuous values are also presented with finite number of bits and thus with finite accuracy. Most commonly used presentations are floating-point presentations that try to have balanced accuracy over the range of values where it mostly matters. As the the presentation has finite accuracy there are limitations, for example, with IEC 60559 floating-point (double precision) arithmetic used in current R

- the smallest positive floating-point number x such that $1 + x \neq 1$ is $2.220446 \cdot 10^{-16}$
- the smallest non-zero normalized floating-point number is $2.225074 \cdot 10^{-308}$
- the largest normalized floating-point number $1.797693 \cdot 10^{308}$
- the largest integer which can be represented is $2^{31} - 1 = 2147483647$
- see more at https://stat.ethz.ch/R-manual/R-devel/library/base/html/zMachine.html

Article by Goldberg (1991) "What Every Computer Scientist Should Know About Floating-Point Arithmetic" https://docs.oracle.com/cd/E19957-01/806-3568/ncg_goldberg.html provides nice overvew of floating-point arithmetic and how the computations should be arranged for improved accuracy.

Lecture notes by Geyer (2020) "Stat 3701 Lecture Notes: Computer Arithmetic" https://stat.umn.edu/geyer/3701/notes/arithmetic.html provide code examples in R illustrating the most common issues in floating-point arithmetic including examples similar shown in the BDA course lecture.

Draws and sample

A group of draws is a sample. A sample can consist of one draw, and thus some people use the word sample for both single item and for the group. For clarity, we prefer separate words for single item (draw) and for the group (sample).

How many digits should be displayed

- Too many digits make reading of the results slower and give false impression of the accuracy.
- Don't show digits which are just random noise. You can use Monte Carlo standard error estimates to check how many digits are likely to stay the same if the sampling would be continued.
- Show meaningful digits given the posterior uncertainty. You can compare posterior standard error or posterior intervals to the mean value. Posterior interval length can be used to determine also how many digits to show for the interval endpoints.
- Example: The mean and 90% central posterior interval for temperature increase C°/century (see the slides for the example) based on posterior draws:
  - 2.050774 and [0.7472868 3.3017524] (too many digits)
  - 2.1 and [0.7 3.3] (good compared to the interval length)
  - 2 and [1 3] (depends on the context)
- Example: The probability that temp increase is positive
  - 0.9960000 (too many digits)

  - 1.00 (depends on the context. 1.00 hints it's not exactly 1, but larger than 0.99)
  - With 4000 draws MCSE ≈ 0.002. We could report that probability is very likely larger than 0.99, or sample more to justify reporting three digits
  - For probabilities close to 0 or 1, consider also when the model assumption justify certain accuracy

- When reporting many numbers in table, for aesthetics reasons, it may be sometimes better for some numbers to show one extra or one too few digits compared to the ideal.

- Often it's better to plot the whole posterior density in addition of any summaries, as summaries always loose some information content.

- For your reports: Don't be lazy and settle for the default number of digits in R or Python. Think for each reported value how many digits is sensible.

Quadrature

Sometimes 'quadrature' is used to refer generically to any numerical integration method (including Monte Carlo), sometimes it is used to refer just to deterministic numerical integration methods.

Rejection sampling

Rejection sampling is mostly used as a part of fast methods for univariate sampling. For example, sampling from the normal distribution is often made using Ziggurat method, which uses a proposal distribution resembling stairs.

Rejection sampling is also commonly used for truncated distributions, in which case all draws from the truncated part are rejected.

Importance sampling

Popularity of importance sampling is increasing. It is used, for example, as part of other methods as particle filters and pseudo marginal likelihood approaches, and to improve distributional approximations (including variational inference in machine learning).

Importance sampling is useful in importance sampling leave-one-out cross-validation. Cross-validation is discussed in Chapter 7 and importance sampling leave-one-out cross-validation is discussed in the article

- Aki Vehtari, Andrew Gelman and Jonah Gabry (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. In Statistics and Computing, 27(5):1413–1432. arXiv preprint arXiv:1507.04544 <http://arxiv.org/abs/1507.04544>

After the book was published, we have developed Pareto smoothed importance sampling which is more stable than plain importance sampling and has very useful Pareto-k diagnostic to check the reliability

- Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry (2019). Pareto smoothed importance sampling. arXiv preprint arXiv:1507.02646. <http://arxiv.org/abs/1507.02646>

Importance resampling with or without replacement

BDA3 p. 266 recommends importance resampling without replacement. At the time of writing that in 2013, we had less experience with importance sampling and there were some reasonable papers showing reduced variance doing resampling without replacement. We don't recommend this anymore as Pareto smoothed importance sampling works better and is also applicable when the resample sample size is equal to the original sample size.

Importance sampling effective sample size

BDA3 1st (2013) and 2nd (2014) printing have an error for $\tilde{w}(\theta^s)$ used in the effective sample size equation 10.4. The normalized weights equation should not have the multiplier S (the normalized weights should sum to one). Errata for the book http://www.stat.columbia.edu/~gelman/book/errata_bda3.txt.

The effective sample size estimate mentioned in the book is generic approximation, and more accurate effective sample size estimate would take into account also the functional. For example, importance sampling effective sample size can be different when estimating $\mathrm{E}[\theta]$ or $\mathrm{E}[\theta]^2$. If you are interested see more details, for example, in our Pareto importance sampling paper https://arxiv.org/abs/1507.02646.

Buffon's needles

Computer simulation of Buffon's needle dropping method for estimating the value of π https://mste.illinois.edu/activity/buffon/.