Bayesian data analysis – reading instructions

Aki Vehtari

Chapter 1 – outline

Outline of the chapter 1

- 1.1-1.3 important terms, especially 1.3 for the notation
- 1.4 an example related to the first excerise, and another practical example
- 1.5 foundations
- 1.6 good example related to visualisation exercise
- 1.7 example which can be skipped
- 1.8 background material, good to read before doing the first exercises
- 1.9 background material, good to read before doing the second exercises
- 1.10 a point of view for using Bayesian inference

Chapter 1 – most important terms

Find all the terms and symbols listed below. Note that some of the terms are now only briefly introduced and will be covered later in more detail. When reading the chapter, write down questions related to things unclear for you or things you think might be unclear for others.

- full probability model
- posterior distribution
- potentially observable quantity
- quantities that are not directly observable
- exchangeability
- independently and identically distributed
- $\theta, y, \tilde{y}, x, X, p(\cdot|\cdot), p(\cdot), \Pr(\cdot), \sim, H$
- sd, E, var
- Bayes rule
- prior distribution
- sampling distribution, data distribution
- joint probability distribution
- posterior density
- probability
- density
- distribution
- $p(y|\theta)$ as a function of y or θ
- likelihood
- posterior predictive distribution
- probability as measure of uncertainty
- subjectivity and objectivity

- transformation of variables
- simulation
- inverse cumulative distribution function

Proportional to

The symbol $\propto$ means *proportional to*, which means left hand side is equal to right hand size given a constant multiplier. For instance if $y = 2x$, then $y \propto x$. It's `\propto` in LaTeX. See https://en.wikipedia.org/wiki/Proportionality_(mathematics).

Model and likelihood

Term $p(y|\theta, M)$ has two different names depending on the situation. Due to the short notation used, there is possibility of confusion.

1) Term $p(y|\theta, M)$ is called a *model* (sometimes more specifically *observation model* or *statistical model*) when it is used to describe uncertainty about y given θ and M. Longer notation $p_y(y|\theta, M)$ shows explicitly that it is a function of y.

2) In Bayes rule, the term $p(y|\theta, M)$ is called *likelihood function*. Posterior distribution describes the probability (or probability density) for different values of θ given a fixed y, and thus when the posterior is computed the terms on the right hand side (in Bayes rule) are also evaluated as a function of θ given fixed y. Longer notation $p_\theta(y|\theta, M)$ shows explicitly that it is a function of θ. Term has it's own name (likelihood) to make the differene to the model. The likelihood function is unnormalized probability distribution describing uncertainty related to θ (and that's why Bayes rule has the normalization term to get the posterior distribution).

Two types of uncertainty

Epistemic and aletory uncertainty are reviewed nicely in the article: Tony O'Hagan, "Dicing with unknown" Significance 1(3):132-133, 2004. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2004.00050.x/abstract

Transformation of variables

- BDA3 p. 21

Ambiguous notation in statistics

In $p(y|\theta)$

- y can be variable or value
  we could clarify by using $p(Y|\theta)$ or $p(y|\theta)$
- θ can be variable or value
  we could clarify by using $p(y|\Theta)$ or $p(y|\theta)$
- p can be a discrete or continuous function of y or θ
  we could clarify by using P_Y, P_Θ, p_Y or p_Θ

- $P_Y(Y|\Theta = \theta)$ is a probability mass function, sampling distribution, observation model
- $P(Y = y|\Theta = \theta)$ is a probability
- $P_\Theta(Y = y|\Theta)$ is a likelihood function (can be discrete or continuous)
- $p_Y(Y|\Theta = \theta)$ is a probability density function, sampling distrbution, observation model
- $p(Y = y|\Theta = \theta)$ is a density
- $p_\Theta(Y = y|\Theta)$ is a likelihood function (can be discrete or continuous)
- y and θ can also be mix of continuous and discrete
- Due to the sloppines sometimes likelihood is used to refer $P_{Y,\theta}(Y|\Theta)$, $p_{Y,\theta}(Y|\Theta)$

Exchangeability

You don't need to understand or use the term exchangeability before Chapter 5 and Lecture 7. At this point and until Chapter 5 and Lecture 7, it is sufficient that you know that 1) independence is stronger condition than exchangeability, 2) independence implies exchangeability, 3) exchangeability does not imply independence, 4) exchangeability is related to what information is available instead of the properties of unknown underlying data generating mechanism. If you want to know more about exchangeability right now, then read BDA Section 5.2 and BDA_notes_ch5.

Bayesian data analysis – reading instructions 2

Aki Vehtari

Chapter 2 – outline

Outline of the chapter 2

- 2.1 Binomial model (e.g. biased coin flipping)
- 2.2 Posterior as compromise between data and prior information
- 2.3 Posterior summaries
- 2.4 Informative prior distributions (skip exponential families and sufficient statistics)
- 2.5 Gaussian model with known variance
- 2.6 Other single parameter models
  - in this course the normal distribution with known mean but unknwon variance is the most important
  - glance through Poisson and exponential
- 2.7 glance through this example, which illustrates benefits of prior information, no need to read all the details (it's quite long example)
- 2.8 Noninformative and weakly informative priors

Laplace's approach for approximating integrals is discussed in more detail in Chapter 4.

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

- demo2_1: Binomial model and Beta posterior.
- demo2_2: Comparison of posterior distributions with different parameter values for the Beta prior distribution.
- demo2_3: Use samples to plot histogram with quantiles, and the same for a transformed variable.
- demo2_4: Grid sampling using inverse-cdf method.

Chapter 2 – most important terms

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. See also the additional comments below.

- binomial model
- Bernoulli trial
- exchangeability
- Bin, $\binom{n}{y}$
- Laplace's law of succession
- think which expectations in eqs. 2.7-2.8
- summarizing posterior inference
- mode, mean, median, standard deviation, variance, quantile
- central posterior interval

- highest posterior density interval / region
- uninformative / informative prior distribution
- principle of insufficient reason
- hyperparameter
- conjugacy, conjugate family, conjugate prior distribution, natural conjugate prior
- nonconjugate prior
- normal distribution
- conjugate prior for mean of normal distribution with known variance
- posterior for mean of normal distribution with known variance
- precision
- posterior predictive distribution
- normal model with known mean but unknown variance
- proper and improper prior
- unnormalized density
- Jeffreys' invariance principle
- note non-uniqueness of noniformative priors for the binomial parameter
- difficulties with noninformative priors
- weakly informative priors

Integration over Beta distribution

Chapter 2 has an example of analysing the ratio of girls born in Paris 1745–1770. Laplace used binomial model and uniform prior which produces Beta distribution as posterior distribution. Laplace wanted to calculate $p(\theta \geq 0.5)$, which is obtained as

$$\begin{aligned} p(\theta \geq 0.5) &= \int_{0.5}^{1} p(\theta|y,n,M)d\theta \\ &= \frac{493473!}{241945!251527!}\int_{0.5}^{1}\theta^{y}(1-\theta)^{n-y}d\theta \end{aligned}$$

Note that $\Gamma(n) = (n-1)!$. Integral has a form which is called *incomplete Beta function*. Bayes and Laplace had difficulties in computing this, but nowadays there are several series and continued fraction expressions. Furthermore usually the normalisation term is computed by computing $\log(\Gamma(\cdot))$ directly without explicitly computing $\Gamma(\cdot)$. Bayes was able to solve integral given small n and y. In case of large n and y, Laplace used Gaussian approximation of the posterior (more in chapter 4). In this specific case, R pbeta gives the same results as Laplace's result with at least 3 digit accuracy.

Numerical accuracy

Laplace calculated

$$p(\theta \geq 0.5|y,n,M) \approx 1.15 \times 10^{-42}.$$

Correspondingly Laplace could have calculated

$$p(\theta \geq 0.5|y,n,M) = 1 - p(\theta \leq 0.5|y,n,M),$$

which in theory could be computed in R with `1-pbeta(0.5,y+1,n-y+1)`. In practice this fails, due to the limitation in the floating point representation used by the computers. In R the largest floating point number which is smaller than 1 is about 1-eps/4, where eps is about 2.22×10^{-16} (the smallest floating point number larger than 1 is 1+eps). Thus the result from `pbeta(0.5,y+1,n-y+1)` will be rounded to 1 and $1 - 1 = 0 \neq 1.15 \times 10^{-42}$. We can compute $p(\theta \geq 0.5|y, n, M)$ in R with `pbeta(0.5, y+1, n-y+1, lower.tail=FALSE)`.

Highest Posterior Density interval

HPD interval is not invariant to reparametrization. Here's an illustrative example (using R and package `HDInterval`):

```
> r <- exp(rnorm(1000))
> quantile(log(r),c(.05, .95))
       5%       95%
-1.532931  1.655137
> log(quantile(r,c(.05, .95)))
       5%       95%
-1.532925  1.655139
> hdi(log(r), credMass = 0.9)
    lower     upper
-1.449125  1.739169
attr(,"credMass")
[1] 0.9
> log(hdi(r, credMass = 0.9))
    lower     upper
-2.607574  1.318569
attr(,"credMass")
[1] 0.9
```

Gaussian distribution in more complex models and methods

Gaussian distribution is commonly used in mixture models, hierarchical models, hierarchical prior structures, scale mixture distributions, Gaussian latent variable models, Gaussian processes, Gaussian random Markov fields, Kalman filters, proposal distribution in Monte Carlo methods, etc.

Predictive distribution

Often the predictive distribution is more interesting than the posterior distribution. The posterior distribution describes the uncertainty in the parameters (like the proportion of red chips in the bag), but the predictive distribution describes also the uncertainty about the future event (like which color is picked next). This difference is important, for example, if we want to what could happen if some treatment is given to a patient.

In case of Gaussian distribution with known variance σ^2 the model is

$$y \sim \mathrm{N}(\theta, \sigma^2),$$

where σ^2 describes aleatoric uncertainty. Using uniform prior the posterior is

$$p(\theta|y) \sim \mathrm{N}(\theta|\bar{y}, \sigma^2/n),$$

where σ^2/n described epistemic uncertainty related to θ. Using uniform prior the posterior predictive distribution for new $\tilde{y}$ is

$$p(\tilde{y}|y) \sim \mathrm{N}(\tilde{y}|\bar{y}, \sigma^2 + \sigma^2/n),$$

where the uncertainty is sum of epistemic (σ^2/n) and aleatoric uncertainty (σ^2).

Weakly informative priors

Our thinking has advanced since section 2.9 was written. See the Prior Choice Wiki (https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations for more recent general discussion and model specific recommendations.

Should we worry about rigged priors?

Andrew Gelman's blog post answering worries that data analyst would choose a too optimistic prior http://andrewgelman.com/2017/10/04/worry-rigged-priors/.

Exchangeability

You don't need to understand or use the term exchangeability before Chapter 5 and Lecture 7. At this point and until Chapter 5 and Lecture 7, it is sufficient that you know that 1) independence is stronger condition than exchangeability, 2) independence implies exchangeability, 3) exchangeability does not imply independence, 4) exchangeability is related to what information is available instead of the properties of unknown underlying data generating mechanism. If you want to know more about exchangeability right now, then read BDA Section 5.2 and BDA_notes_ch5.

Bayesian data analysis – reading instructions 3

Aki Vehtari

Chapter 3

Outline of the chapter 3

- 3.1 Marginalisation
- 3.2 Normal distribution with a noninformative prior (very important)
- 3.3 Normal distribution with a conjugate prior (very important)
- 3.4 Multinomial model (can be skipped)
- 3.5 Multivariate normal with known variance (needed later)
- 3.6 Multivariate normal with unknown variance (glance through)
- 3.7 Bioassay example (very important, related to one of the exercises)
- 3.8 Summary (summary)

Normal model is used a lot as a building block of the models in the later chapters, so it is important to learn it now. Bioassay example is good example used to illustrate many important concepts and it is used in several exercises over the course.

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

- demo3_1: visualise joint density and marginal densities of posterior of normal distribution with unknown mean and variance
- demo3_2: visualise factored sampling and corresponding marginal and conditional density
- demo3_3: visualise marginal distribution of mu as a mixture of normals
- demo3_4: visualise sampling from the posterior predictive distribution
- demo3_5: visualise Newcomb's data
- demo3_6: visualise posterior in bioassay example

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. See also the additional comments below.

- marginal distribution/density
- conditional distribution/density
- joint distribution/density
- nuisance parameter
- mixture
- normal distribution with a noninformative prior
- normal distribution with a conjugate prior
- sample variance
- sufficient statistics
- μ,σ^2,$\bar{y}$,s^2
- a simple normal integral

- Inv-χ^2
- factored density
- t_{n-1}
- degrees of freedom
- posterior predictive distribution
- to draw
- N-Inv-χ^2
- variance matrix Σ
- nonconjugate model
- generalized linear model
- exchangeable
- binomial model
- logistic transformation
- density at a grid

Conjugate prior for Gaussian distribution

BDA p. 67 (BDA3) mentions that the conjugate prior for Gaussian distribution has to have a product form $p(\sigma^2)p(\mu|\sigma^2)$. The book refers to (3.2) and the following discussion. As additional hint is useful to think the relation of terms $(n-1)s^2$ and $n(\bar{y}-\mu)^2$ in 3.2 to equations 3.3 and 3.4.

Trace of square matrix

Trace of square matrix, trace, $\mathrm{tr}\,A$, $\mathrm{trace}(A)$, $\mathrm{tr}(A)$, is the sum of diagonal elements. To derive equation 3.11 the following property has been used $\mathrm{tr}(ABC) = \mathrm{tr}(CAB) = \mathrm{tr}(BCA)$.

History and naming of distributions

See *Earliest Known Uses of Some of the Words of Mathematics* http://jeff560.tripod.com/mathword.html.

Using Monte Carlo to obtain draws from the posterior of generated quantities

Chapter 3 discusses closed form posteriors for binomial and normal models given conjugate priors. These are also used as part of the assignment. The assignment also requires forming a posterior for derived quantities, and these posterior don't have closed form (so no need to try derive them). As we know how to sample from the posterior of binomial and normal models, we can use these posterior draws to get draws from the posterior of derived quantity.

For example, given posteriors $p(\theta_1|y_1)$ and $p(\theta_2|y_2)$ we want to find the posterior for the difference $p(\theta_1 - \theta_2|y_1, y_2)$.

1. Sample θ_1^s from $p(\theta_1|y_1)$ and θ_2^s from $p(\theta_2|y_2)$, we can compute posterior draws for the derived quantity as $\delta^s = \theta_1^s - \theta_2^s$ $(s = 1, \ldots, S)$.

2. δ^s are then draws from $p(\delta^s|y_1, y_2)$, and they can be used to illustrate the posterior $p(\delta^s|y_1, y_2)$ with histogram, and compute posterior mean, sd, and quantiles.

This is one reason why Monte Carlo approaches are so commonly used.

The number of required Monte Carlo draws

This will discussed in chapter 10. Meanwhile, e.g., 1000 draws is sufficient.

Bioassay

Bioassay example is is an example of very common statistical inference task typical, for example, medicine, pharmacology, health care, cognitive science, genomics, industrial processes etc.

The example is from Racine et al (1986) (see ref in the end of the BDA3). Swiss company makes classification of chemicals to different toxicity categories defined by authorities (like EU). Toxicity classification is based on lethal dose 50% (LD50) which tells what amount of chemical kills 50% of the subjects. Smaller the LD50 more lethal the chemical is. The original paper mentions "1983 Swiss poison Regulation" which defines following categories for chemicals orally given to rats (mg/ml)

Class	LD50
1	<5
2	5-50
3	50-500
4	500-2000
5	2000-5000

To reduce the number of rats needed in the experiments, the company started to use Bayesian methods. The paper mentions that in those days use of just 20 rats to define the classification was very little. Book gives LD50 in log(g/ml). When the result from demo3_6 is transformed to scale mg/ml, we see that the mean LD50 is about 900 and $p(500 < \text{LD50} < 2000) \approx 0.99$. Thus, the tested chemical can be classified as category 4 toxic.

Note that the chemical testing is moving away from using rats and other animals to using, for example, human cells grown in chips, tissue models and human blood cells. The human-cell based approaches are also more accurate to predict the effect for humans.

logit transformation can be justified information theoretically when binomial likelihood is used.

Example codes in demo3_6 can be helpful in exercises related to Bioassay example.

Bayesian vs. frequentist statements in two group comparisons

When asking to compare groups, some students get confused as the frequentist testing is quite different. The frequentist testing is often focusing on a) differently named tests for different models and b) null hypothesis testing. In Bayesian inference a) the same Bayes rule and investigation of posterior is used for all models, b) null hypothesis testing is less common. We come later to decision making given posterior and utility/ cost function (Lecture 10.1) and more about null hypothesis testing (Lecture 12.1). Now it is assumed you will report the posterior (e.g. histogram), possible summaries, and report what you can infer from that. Specifically as in this assignment the group comparisons are based on continuous model parameter, the probability of 0 difference is 0 (later lecture 12.1 covers null hypothesis testing). Instead of forcing dichotomous answer (yes/no) about whether there is difference, report the whole posterior that tells also how big that difference might be. What big means depends on the application, which brings us back to the fact of importance of domain expertise. You are not experts on the application examples used in the assignment, but you can think how would you report what you have learned to a domain expert.

Frank Harrell's recommendations how to state results in two group comparisons are excellent http://www.fharrell.com/2017/10/bayesian-vs-frequentist-statements.html.

Bayesian data analysis – 4

Aki Vehtari

Chapter 4

Outline of the chapter 4

- 4.1 Normal approximation (Laplace's method)
- 4.2 Large-sample theory
- 4.3 Counter examples
- 4.4 Frequency evaluation (not part of the course, but interesting)
- 4.5 Other statistical methods (not part of the course, but interesting)

Normal approximation is used often used as part of posterior computation (more about this in Ch 13, which is not a part of the course).

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

- demo4_1: Bioassay example

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.

- sample size
- asymptotic theory
- normal approximation
- quadratic function
- Taylor series expansion
- observed information
- positive definite
- why $\log \sigma$?
- Jacobian of the transformation
- point estimates and standard errors
- lower-dimnsional approximations
- large-sample theory
- asymptotic normality
- consistency
- underidentified
- nonidentified
- number of parameters increasing with sample size
- aliasing
- unbounded likelihood
- improper posterior
- edge of parameter space
- tails of distribution

Normal approximation

Other Gaussian posterior approximations are discussed in Chapter 13. For example, variational and expectation propagation methods improve the approximation by global fitting instead of just the curvature at the mode. The Gaussian approximation at the mode is often also called the Laplace method, as Laplace used it first.

Several researchers have provided partial proofs that posterior converges towards Gaussian distribution. In the mid 20th century Le Cam was first to provide a strict proof.

Observed information

When $n \to \infty$, the posterior distribution approaches Gaussian distribution. As the log density of the Gaussian is a quadratic function, the higher derivatives of the log posterior approach zero. The curvature at the mode describes the information only in the case if asymptotic normality. In the case of the Gaussian distribution, the curvature describes also the width of the Gaussian. Information matrix is a *precision matrix*, which inverse is a covariance matrix.

Aliasing

In Finnish: valetoisto.

Aliasing is a special case of under-identifiability, where likelihood repeats in separate points of the parameter space. That is, likelihood will get exactly same values and has same shape although possibly mirrored or otherwise projected. For example, the following mixture model

$$p(y_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = \lambda \,\mathrm{N}(\mu_1, \sigma_1^2) + (1 - \lambda)\,\mathrm{N}(\mu_2, \sigma_2^2),$$

has two Gaussians with own means and variances. With a probability λ the observation comes from $\mathrm{N}(\mu_1, \sigma_1^2)$ and a probability $1 - \lambda$ from $\mathrm{N}(\mu_2, \sigma_2^2)$. This kind of model could be used, for example, for the Newcomb's data, so that the another Gaussian component would model faulty measurements. Model does not state which of the components 1 or 2, would model good measurements and which would model the faulty measurements. Thus it is possible to interchange values of (μ_1, μ_2) and (σ_1^2, σ_2^2) and replace λ with $(1 - \lambda)$ to get the equivalent model. Posterior distribution then has two modes which are mirror images of each other. When $n \to \infty$ modes will get narrower, but the posterior does not converge to a single point.

If we can integrate over the whole posterior, the aliasing is not a problem. However aliasing makes the approximative inference more difficult.

Frequency property vs. frequentist

Bayesians can evaluate frequency properties of Bayesian estimates without being frequentist. For Bayesians the starting point is the Bayes rule and decision theory. Bayesians care more about efficiency than unbiasedness. For frequentists the starting point is to find an estimator with desired frequency properties and quite often unbiasedness is chosen as the first restriction.

Transformation of variables

See p. 21 for the explanation how to derive densities for transformed variables. This explains, for example, why uniform prior $p(\log(\sigma^2)) \propto 1$ for $\log(\sigma^2)$ corresponds to prior $p(\sigma^2) = \sigma^{-2}$ for σ^2.

On derivation

Here's a reminder how to integrate with respect to $g(\theta)$. For example

$$\frac{d}{d\log\sigma}\sigma^{-2} = -2\sigma^{-2}$$

is easily solved by setting $z = \log\sigma$ to get

$$\frac{d}{dz}\exp(z)^{-2} = -2\exp(z)^{-3}\exp(z) = -2\exp(z)^{-2} = -2\sigma^{-2}.$$

Bayesian data analysis – reading instructions ch 5

Aki Vehtari

Chapter 5

Outline of the chapter 5

- 5.1 Lead-in to hierarchical models
- 5.2 Exchangeability (a useful theoretical concept)
- 5.3 Bayesian analysis of hierarchical models
- 5.4 Hierarchical normal model
- 5.5 Example: parallel experiments in eight schools (uses hierarchical normal model, details of computation can be skipped)
- 5.6 Meta-analysis (can be skipped in this course)
- 5.7 Weakly informative priors for hierarchical variance parameters

The hierarchical models in the chapter are simple to keep computation simple. More advanced computational tools are presented in Chapters 10-12 (part of the course) and 13 (not part of the course).

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

- demo5_1: Rats example
- demo5_2: SAT example

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.

- population distribution
- hyperparameter
- overfit
- hierarchical model
- exchangeability
- invariant to permutations
- independent and identically distributed
- ignorance
- the mixture of independent identical distributions
- de Finetti's theorem
- partially exchangeable
- conditionally exchangeable
- conditional independence
- hyperprior
- different posterior predictive distributions
- the conditional probability formula

Computation

Examples in Sections 5.3 and 5.4 continue computation with factorization and grid, but there is no need to go deep in to computational details as in the assignments you will use MCMC and Stan instead.

Exchangeability vs. independence

Exchangeability and independence are two separate concepts. Neither necessarily implies the other. Independent identically distributed variables/parameters are exchangeable. Exchangeability is less strict condition than independence. Often we may assume that observations or unobserved quantities are in fact dependent, but if we can't get information about these dependencies we may assume those observations or unobserved quantities as exchangeable. "Ignorance implies exchangeability."

In case of exchangeable observations, we may sometimes act *as if* observations were independent if the additional potential information gained from the dependencies is very small. This is related to de Finetti's theorem (p. 105), which applies formally only when $J \to \infty$, but in practice difference may be small if J is finite but relatively large (see examples below).

- If no other information than data y is available to distinguish θ_j from each other and parameters can not be ordered or grouped, we may assume symmetry between parameters in their prior distribution
- This symmetry can be represented with exchangeability
- Parameters $\theta_1, \ldots, \theta_J$ are exchangeable in their joint distribution if $p(\theta_1, \ldots, \theta_J)$ is invariant to permutation of indexes $(1, \ldots, J)$

Here are some examples you may consider.

Ex 5.1. Exchangeability with known model parameters: For each of following three examples, answer: (i) Are observations y_1 and y_2 exchangeable? (ii) Are observations y_1 and y_2 independent? (iii) Can we act *as if* the two observations are independent?

1. A box has one black ball and one white ball. We pick a ball y_1 at random, put it back, and pick another ball y_2 at random.

2. A box has one black ball and one white ball. We pick a ball y_1 at random, we do not put it back, then we pick ball y_2.

3. A box has a million black balls and a million white balls. We pick a ball y_1 at random, we do not put it back, then we pick ball y_2 at random.

Ex 5.2. Exchangeability with unknown model parameters: For each of following three examples, answer: (i) Are observations y_1 and y_2 exchangeable? (ii) Are observations y_1 and y_2 independent? (iii) Can we act *as if* the two observations are independent?

1. A box has n black and white balls but we do not know how many of each color. We pick a ball y_1 at random, put it back, and pick another ball y_2 at random.

2. A box has n black and white balls but we do not know how many of each color. We pick a ball y_1 at random, we do not put it back, then we pick ball y_2 at random.

3. Same as (b) but we know that there are many balls of each color in the box.

Note that for example in opinion polls, balls i.e. humans are not put back and there is a large but finite number of humans.

Following complements the divorce example in the book by discussing the effect of the additional observations

- Example: divorce rate per 1000 population in 8 states of the USA in 1981
  - without any other knowledge $y_1, \ldots, y_8$ are exchangeable
  - it is reasonable to assume a prior independence given population density $p(y_i|\theta)$
- Divorce rate in first seven are $5.6, 6.6, 7.8, 5.6, 7.0, 7.2, 5.4$
  - now we have some additional information, but still changing the indexing does not affect the joint distribution. For example, if we were told that divorce rate were not for the first seven but last seven states, it does not change the joint distribution, and thus $y_1, \ldots, y_8$ are exchangeable
  - sensible assumption is a prior independence given population density $p(y_i|\theta)$
  - if "true" θ_0 were known, $y_1, \ldots, y_8$ were independent given "true" θ_0
  - since θ is estimated using observations, y_i are a posterior dependent, which is obvious, e.g., from the predictive density $p(y_8|y_1, \ldots, y_7, M)$, i.e. e.g. if $y_1, \ldots, y_7$ are large then probably y_8 is large
  - if we were told that given rates were for the last seven states, then $p(y_1|y_2, \ldots, y_8, M)$ would be exactly same as $p(y_8|y_1, \ldots, y_7, M)$ above, i.e. changing the indexing does not have effect since $y_1, \ldots, y_8$ are exchangeable
- Additionally we know that y_8 is Nevada and rates of other states are $5.6, 6.6, 7.8, 5.6, 7.0, 7.2, 5.4$
  - based on what we were told about Nevada, predictive density s $p(y_8|y_1, \ldots, y_7, M)$ should take into account that probability $p(y_8 > \max(y_1, \ldots, y_7)|y_1, \ldots, y_7)$ should be large
  - if we were told that, Nevada is y_3 (not y_8 as above), then new predictive density $p(y_8|y_1, \ldots, y_7, M)$ would be different, because $y_1, \ldots, y_8$ are not anymore exchangeable

What if observations are not exchangeable

Often observations are not fully exchangeable, but are partially or conditionally exchangeable. Two basic cases

1) If observations can be grouped, we may make hierarchical model, were each group has own subpart, but the group properties are unknown. If we assume that group properties are exchangeable we may use common prior for the group properties.
2) If y_i has additional information x_i, then y_i are not exchangeable, but (y_i, x_i) still are exchangeable, then we can be make joint model for (y_i, x_i) or conditional model $(y_i|x_i)$.

Here are additional examples (Bernardo & Smith, Bayesian Theory, 1994), which illustrate the above basic cases. Think of old fashioned thumb pin. This kind of pin can stay flat on it's base or slanting so that the pin head and the edge of the base touch the table. This kind of pin represents realistic version of "unfair" coin.

1) Let's throw pin n times and mark $x_i = 1$ when pin stands on it's base. Let's assume, that throwing conditions stay same all the time. Most would accept throws as exchangeable.
2) Same experiment, but odd numbered throws will be made with full metal pin and even numbererd throws with plastic coated pin. Most would accept exchangeability for all odd and all even throws separately, but not necessarily for both series combined. Thus we have partial exchangeability.
3) Laboratory experiments $x_1, ..., x_n$, are real valued measurements about the chemical property of some substance. If all experiments are from the same sample, in the same laboratory with same

procedure, most would accept exchangeability. If experiments were made, for example, in different laboratories we could assume partial exchangeability.

4) $x_1, ..., x_n$ are real valued measurements about the physiological reactions to certain medicine. Different test persons get different amount of medicine. Test persons are males and females of different ages. If the attributes of the test persons were known, most would not accept results as exchangeable. In a group with certain dose, sex and age, the measurements could be assumed exchangeable. We could use grouping or if the doses and attributes are continuous we could use regression, i.e. assume conditional independence.

Weakly informative priors for hierarchical variance parameters

Our thinking has advanced since section 5.7 was written. Section 5.7 (p. 128–) recommends use of half-Cauchy as weakly informative prior for hierarchical variance parameters. More recent recommendation is half-normal if you have substantial information on the high end values, or or half-t_4 if you there might be possibility of surprise. Often we don't have so much prior information that we would be able to well define the exact tail shape of the prior, but half-normal produces usually more sensible prior predictive distributions and is thus better justified. Half-normal leads also usually to easier inference.

See the Prior Choice Wiki (https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations for more recent general discussion and model specific recommendations.

Bayesian data analysis – reading instructions 6

Aki Vehtari

Chapter 6

Outline of the chapter 6

- 6.1 The place of model checking in applied Bayesian statistics
- 6.2 Do the inferences from the model make sense?
- 6.3 Posterior predictive checking (p-values can be skipped)
- 6.4 Graphical posterior predictive checks
- 6.5 Model checking for the educational testing example

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

- demo6_1: Posterior predictive checking - light speed
- demo6_2: Posterior predictive checking - sequential dependence
- demo6_3: Posterior predictive checking - poor test statistic
- demo6_4: Posterior predictive checking - marginal predictive p-value

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.

- model checking
- sensitivity analysis
- external validation
- posterior predictive checking
- joint posterior predictive distribution
- marginal (posterior) predictive distribution
- self-consistency check
- replicated data
- y^{rep}, $\tilde{y}$, $\tilde{x}$
- test quantities
- discrepancy measure
- tail-area probabilities
- classical p-value
- posterior predictive p-values
- multiple comparisons
- marginal predictive checks
- cross-validation predictive distributions
- conditional predictive ordinate

Replicates vs. future observation

Predictive $\tilde{y}$ is the next not yet observed possible observation. y^{rep} refers to replicating the whole experiment (with same values of x) and obtaining as many replicated observations as in the original data.

Posterior predictive p-values

Section 6.3 discusses posterior predictive p-values, which we don't recommend any more especially in a form of hypothesis testing.

Prior predictive checking

Prior predictive checking using just the prior predictive distributions is very useful tool for assessing the sensibility of the model and priors even before observing any data or before doing the posterior inference. See additional reading below for examples.

Additional reading

The following article has some useful discussion and examples also about prior and posterior predictive checking.

- Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman (2018). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society Series A*, , 182(2):389-402. https://doi.org/10.1111/rssa.12378.
- Video of the paper presentation https://www.youtube.com/watch?v=E8vdXoJId8M

And some useful demos

- Graphical posterior predictive checks using the bayesplot package http://mc-stan.org/bayesplot/articles/graphical-ppcs.html
- Another demo demos_rstan/ppc/poisson-ppc.Rmd

Bayesian data analysis – reading instructions 7

Aki Vehtari

Chapter 7

Outline

- 7.1 Measures of predictive accuracy
- 7.2 Information criteria and cross-validation (read instead the article mentioned below)
- 7.3 Model comparison based on predictive performance (read instead the article mentioned below)
- 7.4 Model comparison using Bayes factors
- 7.5 Continuous model expansion / sensitivity analysis
- 7.5 Example (may be skipped)

Instead of Sections 7.2 and 7.3 it's better to read

- Aki Vehtari, Andrew Gelman and Jonah Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**(5):1413-1432, doi:10.1007/s11222-016-9696-4. arXiv preprint arXiv:1507.04544.
- LOO package glossary summarises many important terms used in the assignments.

In Sections 7.2 and 7.3 of BDA, for historical reasons there is a multiplier -2 used. After the book was published, we have concluded that it causes too much confusion and recommend not to multiply by -2. The above paper is not using -2 anymore.

Extra material

The following article provides excellent discussion about "How should I evaluate my modelling choices?" from a scientific perspective.

- Danielle J. Navarro (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior* **2**:28–34. Online.

There is extra material at https://avehtari.github.io/modelselection/

- Videos, slides, notebooks, references
- Sections 1 and 5 (less than 3 pages) of "Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison" clarify how to interpret standard error in model comparison
- Cross-validation FAQ https://avehtari.github.io/modelselection/CV-FAQ.html answers many frequently asked questions.

Important terms

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

- predictive accuracy/fit/error

- external validation
- cross-validation
- information criteria
- overfitting
- measures of predictive accuracy
- point prediction
- scoring function
- mean squared error
- probabilistic prediction
- scoring rule
- logarithmic score
- log-predictive density
- out-of-sample predictive fit
- elpd, elppd, lppd
- deviance
- within-sample predictive accuracy
- adjusted within-sample predictive accuracy
- AIC, DIC, WAIC (less important)
- effective number of parameters
- singular model
- BIC (less important)
- leave-one-out cross-validation
- evaluating predictive error comparisons
- bias induced by model selection
- Bayes factors
- continuous model expansion
- sensitivity analysis

Additional reading

More theoretical details can be found in

- Aki Vehtari and Janne Ojanen (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. In Statistics Surveys, 6:142-228. http://dx.doi.org/10.1214/12-SS102

See more experimental comparisons in

- Juho Piironen and Aki Vehtari (2017). Comparison of Bayesian predictive methods for model selection. Statistics and Computing, 27(3):711-735. doi:10.1007/s11222-016-9649-y. http://link.springer.com/article/10.1007/s11222-016-9649-y

Posterior probability of the model vs. predictive performance

Gelman: "To take a historical example, I don't find it useful, from a statistical perspective, to say that in 1850, say, our posterior probability that Newton's laws were true was 99%, then in 1900 it was 50%, then

by 1920, it was 0.01% or whatever. I'd rather say that Newton's laws were a good fit to the available data and prior information back in 1850, but then as more data and a clearer understanding became available, people focused on areas of lack of fit in order to improve the model."

Newton's laws are still sufficient for prediction in specific contexts (non-relative speeds and differences in gravity, non-significant effects of air resistance or other friction). See more in the course video 1.1 Introduction to uncertainty and modelling https://aalto.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=d841f429-9c3d-4d24-8228-a9f400efda7b.

Bayesian data analysis – reading instructions 8

Aki Vehtari

Chapter 8

In the earlier chapters it was assumed that the data collection is ignorable. Chapter 8 explains when data collection can be ignorable and when we need to model also the data collection. We don't have time to go through chapter 8 in BDA course at Aalto, but it is highly recommended that you would read it in the end or after the course. Most important parts are 8.1, 8.5, pp 220–222 of 8.6, and 8.8, and you can get back to the other sections later.

Outline of the chapter 8 (* denotes the most important parts)

- 8.1 Bayesian inference requires a model for data collection (*)
- 8.2 Data-collection models and ignorability
- 8.3 Sample surveys
- 8.4 Designed experiments
- 8.5 Sensitivity and the role of randomization (*)
- 8.6 Observational studies (* pp 220–222)
- 8.7 Censoring and truncation (*)

Most important terms in the chapter

- observed data
- complete data
- missing data
- stability assumption
- data model
- inclusion model
- complete data likelihood
- observed data likelihood
- finite-population and superpopulation inference
- ignorability
- ignorable designs
- propensity score
- sample surveys
- random sampling of a finite population
- stratified sampling
- cluster sampling
- designed experiments
- complete randomization
- randomized blocks and latin squares
- sequntial designs
- randomization given covariates

- observational studies
- censoring
- truncation
- missing completely at random

Bayesian data analysis – reading instructions 9

Aki Vehtari

Chapter 9

Outline of the chapter 9

- 9.1 Context and basic steps (most important part)
- 9.2 Example
- 9.3 Multistage decision analysis (you may skip this example)
- 9.4 Hierarchical decision analysis (you may skip this example)
- 9.5 Personal vs. institutional decision analysis (important)

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.

- decision analysis
- steps of Bayesian decision analysis 1–4 (p. 238)
- decision
- outcome
- utility function
- expected utility
- decision tree
- summarizing inference
- model selection
- individual decision problem
- institutional decision problem

Simpler examples

The lectures have simpler examples and discus also some challenges in selecting utilities or costs.

Model selection as a decision problem

Chapter 7 discusses how model selection con be considered as a decision problem.

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 computer 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.

Stan User Guide Chapter 15 https://mc-stan.org/docs/2_26/stan-users-guide/floating-point-arithmetic.html discusses floating point arithmetic in context of Stan.

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 a 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 $\approx$ 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.

The derivation for the effective sample size and Monte Carlo standard error (MCSE) for importance sampling can be found, for example, in Chapter 9 of *Monte Carlo theory, methods and examples* by Art B. Owen https://statweb.stanford.edu/~owen/mc/.

Buffon's needles

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

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 $\widehat{R}$

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, $\widehat{R}$
- 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 $\widehat{R}$ 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 $\widehat{R}$, 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 (2020). 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 $\widehat{R}$, 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, $\widehat{R}$ 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.

Bayesian data analysis – reading instructions 12

Aki Vehtari

Chapter 12

Outline of the chapter 12

- 12.1 Efficient Gibbs samplers (not part of the course)
- 12.2 Efficient Metropolis jump rules (not part of the course)
- 12.3 Further exetensions to Gibbs and Metropolis (not part of the course)
- 12.4 Hamiltonian Monte Carlo (used in Stan)
- 12.5 Hamiltonian dynamics for a simple hierarchical model (read through)
- 12.6 Stan: developing a computing environment (read through)

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

- See rstan_demo.Rmd, pystan_demo.py, pystan_demo.ipynb for demos how to use Stan from R/Python and several model examples

There is only 8 pages to read (sections 12.4-12.6) what is inside Stan.

MCMC animations

These don't include the specific version of dynamic HMC in Stan, but are useful illustrations anyway.

- Markov Chains: Why Walk When You Can Flow?
  http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/
- MCMC animation site by Chi Feng https://chi-feng.github.io/mcmc-demo/

Hamiltonian Monte Carlo

An excellent review of static HMC (the number of steps in dynamic simulation are not adaptively selected) is

- Radford Neal (2011). MCMC using Hamiltonian dynamics. In Brooks et al (ed), *Handbook of Markov Chain Monte Carlo*, Chapman & Hall / CRC Press. Preprint https://arxiv.org/pdf/1206.1901.pdf.

Stan uses a variant of dynamic Hamiltonian Monte Carlo (using adaptive number of steps in the dynamic simulation), which has been further developed since BDA3 was published. The first dynamic HMC variant was

- Matthew D. Hoffman, Andrew Gelman (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *JMLR*, 15:1593–1623 http://jmlr.org/papers/v15/hoffman14a.html.

The No-U-Turn Sampler gave the name NUTS which you can see often associated with Stan, but the current dynamic HMC variant implemented in Stan has some further developments described (mostly) in

- Michael Betancourt (2018). A Conceptual Introduction to Hamiltonian Monte Carlo. arXiv preprint arXiv:1701.02434 https://arxiv.org/abs/1701.02434.

Instead of reading all above, you can also watch a video

- Scalable Bayesian Inference with Hamiltonian Monte Carlo by Michael Betancourt https://www.youtube.com/watch?v=jUSZboSq1zg

Divergences and BFMI

Divergences and Bayesian Fraction of Missing Information (BFMI) are HMC specific convergence diagnostics developed by Michael Betancourt after BDA3 was published.

- Divergence diagnostic checks whether the discretized dynamic simulation has problems due to fast varying density. See more in a case study http://mc-stan.org/users/documentation/case-studies/divergences_and_bias.html.
- BFMI checks whether momentum resampling in HMC is sufficiently efficient. See more in https://arxiv.org/abs/1604.00695
- 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.

Further information about Stan

- http://mc-stan.org/ & http://mc-stan.org/documentation/
  - I recommend to start with these
    - Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2015) In press for Journal of Statistical Software. Stan: A Probabilistic Programming Language. http://www.stat.columbia.edu/~gelman/research/published/stan-paper-revision-feb2015.pdf
    - Andrew Gelman, Daniel Lee, and Jiqiang Guo (2015) Stan: A probabilistic programming language for Bayesian inference and optimization. In press, Journal of Educational and Behavior Science. http://www.stat.columbia.edu/~gelman/research/published/stan_jebs_2.pdf
  - Basics of Bayesian inference and Stan, parts 1+2 Jonah Gabry & Lauren Kennedy https://www.youtube.com/playlist?list=PLuwyh42iHquU4hUBQs20hkBsKSMrp6H0J
  - Modeling Language User's Guide and Reference Manual (more complete reference with lot's of examples) https://github.com/stan-dev/stan/releases/download/v2.17.0/stan-reference-2.17.0.pdf

Compiler and transpiler

This is a minor comment on the terminology. As a shorthand it's common to see mentioned just Stan compiler, but sometimes the transpiler term is also mentioned as in the slides for this part.

Wikipedia (https://en.wikipedia.org/wiki/Source-to-source_compiler):

A source-to-source translator, source-to-source compiler (S2S compiler), transcompiler, or transpiler[1] is a type of translator that takes the source code of a program written in a programming language as

its input and produces an equivalent source code in the same or a different programming language. A source-to-source translator converts between programming languages that operate at approximately the same level of abstraction, while a traditional compiler translates from a higher level programming language to a lower level programming language.

So it is more accurate to say that the Stan model code is first transpiled to a C++ code, and then that C++ code is compiled to machine code to create an executable program. Cool thing about the new stanc3 transpiler (https://github.com/stan-dev/stanc3) is that it can create also, for example, LLVM IR or Tensorflow code.

Using transpiler and compiler allows to develop Stan language to be good for writing models, but get the benefit of speed and external libraries of C++, Tensorflow, and whatever comes in the future.

Bayesian data analysis – reading instructions 13

Aki Vehtari

Chapter 13: Modal and distributional approximations

Chapter 4 presented normal distribution approximation at the mode (aka Laplace approximation). Chapter 13 discusses more about distributional approximations.

Outline of the chapter 13

13.1 Finding posterior modes

- Newton's method is very fast if the distribution is close to normal and the computation of the second derivatives is fast
- Stan uses limited-memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) which is a quasi-Newton method which needs only the first derivatives (provided by Stan autodiff). L-BFGS is known for good performance for wide variety of functions.

13.2 Boundary-avoiding priors for modal summaries

- Although full integration is preferred, sometimes optimization of some parameters may be sufficient and faster, and then boundary-avoiding priors maybe useful.

13.3 Normal and related mixture approximations

- Discusses how the normal approximation can be used to approximate integrals of a a smooth function times the posterior.
- Discusses mixture and t approximations.

13.4 Finding marginal posterior modes using EM

- Expectation maximization is less important in the time of efficient probabilistic programming frameworks, but can be sometimes useful for extra efficiency.

13.5 Conditional and marginal posterior approximations

- Even in the time of efficient probabilistic programming, the methods discussed in this section can produce very big speedups for a big set of commonly used models. The methods discussed are important part of popular INLA software and are coming also to Stan to speedup latent Gaussian variable models.

13.6 Example: hierarchical normal model

13.7 Variational inference

- Variational inference (VI) is very popular in machine learning, and this section presents it in terms of BDA. Auto-diff variational inference in Stan was developed after BDA3 was published.

13.8 Expectation propagation

- Practical efficient computation for expectation propagation (EP) is applicable for more limited set of models than post-BDA3 black-box VI, but for those models EP provides better posterior approximation. Variants of EP can be used for parallelization of any Bayesian computation for hierarchical models.

13.9 Other approximations

- Just brief mentions of INLA (uses methods discussed in 13.5), CCD (deterministic adaptive quadrature approach) and ABC (inference when you can only sample from the generative model).

13.10 Unknown normalizing factors

- Often the normalizing factor is not needed, but it can be estimated using importance, bridge or path sampling.

Bayesian data analysis – reading instructions Part IV

Aki Vehtari

Part IV, Chapters 14–18 discuss basics of linear and generalized linear models with several examples. The parts discussing computation can be useful to provide additional insight on these models or sometimes for actual computation, it's likely that most of the readers will use some probabilistic programming framework for computation. Regression and other stories (ROS) by Gelman, Hill and Vehtari discusses linear and generalized linear models from the modeling perspective more thoroughly.

Chapter 14: Introduction to regression models

Outline of the chapter 14:

14.1 Conditional modeling

- formal justification of conditional modeling
- if joint model factorizes $p(y, x|\theta, \phi) = p(y|x, \theta)p(x|\phi)$
  we can model just $p(y|x, \theta)$

14.2 Bayesian analysis of classical regression

- uninformative prior on β and σ^2
- connection to multivariate normal (cf. Chapter 3) is useful to understand as it then reveals what would be the conjugate prior
- closed form posterior and posterior predictive distribution
- these properties are sometimes useful and thus good to know, but with probabilistic programming less often needed

14.3 Regression for causal inference: incumbency and voting

- Modelling example with bit of discussion on causal inference (see more in ROS Chs. 18-21)

14.4 Goals of regression analysis

- discussion of what we can do with regression analysis (see more in ROS)

14.5 Assembling the matrix of explanatory variables

- transformations, nonlinear relations, indicator variables, interactions (see more in ROS)

14.6 Regularization and dimension reduction

- a bit outdated and short (Bayesian Lasso is not a good idea), see more in lecture 9.3, https://avehtari.github.io/modelselection/ and https://betanalpha.github.io/assets/case_studies/bayes_sparse_regression.html)

14.7 Unequal variances and correlations

- useful concept, but computation is easier with probabilistic programming frameworks

14.8 Including numerical prior information

- useful conceptually, but easy computation with probabilistic programming frameworks makes it easier to define prior information as the prior doesn't need to be conjugate
- see more about priors in https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Chapter 15 Hierarchical linear models

Chapter 15 combines hierarchical models from Chapter 5 and linear models from Chapter 14. The chapter discusses some computational issues, but probabilistic programming frameworks make computation for hierarchical linear models easy.

Outline of the chapter 15:

15.1 Regression coefficients exchangeable in batches

- exchangeability of parameters
- the discussion of fixed-, random- and mixed-effects models is incomplete
  - we don't recommend using these terms, but they are so popular that it's useful to know them
  - a relevant comment is *The terms 'fixed' and 'random' come from the non-Bayesian statistical tradition and are somewhat confusing in a Bayesian context where all unknown parameters are treated as 'random' or, equivalently, as having fixed but unknown values.*
  - often fixed effects correspond to population level coefficients, random effects correspond to group or individual level coefficients, and mixed model has both

`y ~ 1 + x`	fixed / population effect; pooled model
`y ~ 1 + (0 + x | g)`	random / group effects
`y ~ 1 + x + (1 + x | g)`	mixed effects; hierarchical model

15.2 Example: forecasting U.S. presidential elections

- illustrative example

15.3 Interpreting a normal prior distribution as extra data

- includes very useful interpretation of hierarchical linear model as a single linear model with certain design matrix

15.4 Varying intercepts and slopes

- extends from hierarchical model for scalar parameter to joint hierarchical model for several parameters

15.5 Computation: batching and transformation

- Gibbs sampling part is mostly outdated
- transformations for HMC is useful if you write your own models, but the section is quite short and you can get more information from Stan user guide 21.7 Reparameterization and https://mc-stan.org/users/documentation/case-studies/divergences_and_bias.html

15.6 Analysis of variance and the batching of coefficients

- ANOVA as Bayesian hierarchical linear model
- rstanarm and brms packages make it easy to make ANOVA

15.7 Hierarchical models for batches of variance components

- more variance components

Chapter 16 Generalized linear models

Chapter 16 extends linear models to have non-normal observation models. Model in Bioassay example in Chapter 3 is also generalized linear model. Chapter reviews the basics and discusses some computational issues, but probabilistic programming frameworks make computation for generalized linear models easy (especially with rstanarm and brms). Regression and other stories (ROS) by Gelman, Hill and Vehtari discusses generalized linear models from the modeling perspective more thoroughly.

Outline of the chapter 16:

16 Intro: Parts of generalized linear model (GLM):

1. The linear predictor $\eta = X\beta$
2. The link function $g(\cdot)$ and $\mu = g^{-1}(\eta)$
3. Outcome distribution model with location parameter μ
   - the distribution can also depend on dispersion parameter ϕ
   - originally just exponential family distributions (e.g. Poisson, binomial, negative-binomial), which all have natural location-dispersion parameterization
   - after MCMC made computation easy, GLM can refer to models where outcome distribution is not part of exponential family and dispersion parameter may have its own latent linear predictor

16.1 Standard generalized linear model likelihoods

- section title says "likelihoods", but it would be better to say "observation models"
- continuous data: normal, gamma, Weibull mentioned, but common are also Student's t, log-normal, log-logistic, and various extreme value distributions like generalized Pareto distribution
- binomial (Bernoulli as a special case) for binary and count data with upper limit
  - Bioassay model uses binomial observation model
- Poisson for count data with no upper limit
  - Poisson is useful approximation of Binomial when the observed counts are much smaller than the upper limit

16.2 Working with generalized linear models

- bit of this and that information on how think about GLMs (see ROS for more)
- normal approximation to the likelihood is good for thinking how much information non-normal observations provide, can be useful for someone thinking about computation, but easy computation with probabilistic programming frameworks means not everyone needs this

16.3 Weakly informative priors for logistic regression

- an excellent section although the recommendation on using Cauchy has changed (see https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations)
- the problem of separation is useful to understand
- computation part is outdated as probabilistic programming frameworks make the computation easy

16.4 Overdispersed Poisson regression for police stops

- an example

16.5 State-level opinions from national polls

- another example

16.6 Models for multivariate and multinomial responses

- extension to multivariate responses
- polychotomous data with multivariate binomial or Poisson
- models for ordered categories

16.7 Loglinear models for multivariate discrete data

- multinomial or Poisson as loglinear models

Chapter 17 Models for robust inference

Chapter 17 discusses over-dispersed observation models. The discussion is useful beyond generalized linear models. The computation is outdated. See Regression and other stories (ROS) by Gelman, Hill and Vehtari for more examples.

Outline of the chapter 17:

17.1 Aspects of robustness

- overdispersed models are often connected to robustness of inferences to outliers, but the observed data can be overdispersed without any observation being outlier
- outlier is sensible only in the context of the model, being something not well modelled or something requiring extra model component
- switching to generic overdispersed model can help to recognize problem in the non-robust model (sensitivity analysis), but it can also throw away useful information in the "outliers" and it would be useful to think what is the generative mechanism for observations which are not like others

17.2 Overdispersed versions of standard models

normal	$\rightarrow$	t-distribution
Poisson	$\rightarrow$	negative-binomial
binomial	$\rightarrow$	beta-binomial
probit	$\rightarrow$	logistic / robit

17.3 Posterior inference and computation

- computation part is outdated as probabilistic programming frameworks and MCMC make the computation easy
- posterior is more likely to be multimodal

17.4 Robust inference for the eight schools

- eight schools example is too small too see much difference

17.5 Robust regression using t-distributed errors

- computation part is outdated as probabilistic programming frameworks and MCMC make the computation easy
- posterior is more likely to be multimodal

Chapter 18 Models for missing data

Chapter 18 extends the data collection modelling from Chapter 8. See Regression and other stories (ROS) by Gelman, Hill and Vehtari for more examples.

Outline of the chapter 18:

18.1 Notation

- Missing completely at random (MCAR)
  missingness does not depend on missing values or other observed values (including covariates)
- Missing at random (MAR)
  missingness does not depend on missing values but may depend on other observed values (including covariates)
- Missing not at random (MNAR)
  missingness depends on missing values

18.2 Multiple imputation

1. make a model predicting missing data
2. sample repeatedly from the missing data model to generate multiple imputed data sets
3. make usual inference for each imputed data set
4. combine results

- discussion of computation is partially outdated

18.3 Missing data in the multivariate normal and t models

- a special continuous data case computation, which can still be useful as fast starting point

18.4 Example: multiple imputation for a series of polls

- an example

18.5 Missing values with counted data

- discussion of computation for count data (ie computation in 18.3 is not applicable)

18.6 Example: an opinion poll in Slovenia

- another example