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- a bit outdated and short (Bayesian Lasso is not a good idea), see more in lecture 9.3, https://avehtari.github.io/modelselection/ and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