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).

Demos

- 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 we can use MCMC and Stan instead. Hierarchical model exercises are made with Stan.

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.