Applied Probabilistic Programming & Bayesian Machine Learning

§3-1: Exchangeability

MIT IAP 2017 Max W. Shen

At this point, you should be comfortable using Stan as a tool for Bayesian inference. In this section, we dive into some theoretical considerations regarding assumptions that are implicitly made during Bayesian modeling, particularly with hierarchical models and regression.

An understanding of these topics will allow you to appreciate the potential and limits of Bayesian modeling, as well as recognize scenarios where Bayesian modeling is applicable and where it's not.

0.1 Complete Pooling and No Pooling

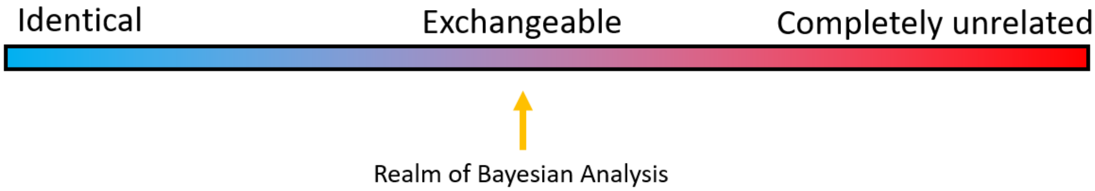

Figure 1: If a sequence of random variables is...

In figure 1, we consider how a sequence of random variables are related to each other.

Concretely, we can consider the example of performing a meta-analysis, pooling together 22 previously published studies. Our quantity of interest is the effect of having children on happiness. Each study has 100 different participants.

If we assume the studies are completely identical in all respects, then we should pool all the data together to get 22×100 samples. We can use this completely pooled data to estimate a single quantity of interest.

In contrast, if we assume the studies are too different to have anything in common, i.e. they are completely unrelated, then our analysis must infer 22 quantities of interest, one for each of the 22 studies.

In reality, our 22 studies are likely to be *similar*, not completely unrelated, but also not completely identical. **Exchangeability** is a formal definition of this notion of ”similar”.

In practice, if we assume our studies are **exchangeable**, then we can place a model over all of the studies. Our model includes the *random variation* between the studies, since they aren't completely identical, but in the end we can estimate a single quantity of interest, combining information contributed by each study.

1 Exchangeability

Theorem 1. *A sequence of random variables $\boldsymbol{X} = (X_1, ..., X_N)$ with associated realizations $(x_1, x_2, ..., x_n)$ are said to be **exchangeable** if:*

$$P(X_1 = x_1, X_2 = x_2, ..., X_N = x_n) = P(X_1 = \pi(x_1), X_2 = \pi(x_2), ..., X_N = \pi(x_n))$$

for all possible permutations given by a function $\pi()$ which permutes the sequence of observations $(x_1, x_2, ..., x_n)$.

Importantly, the order of random variables $(X_1, X_2, ..., X_N)$ is *not* permuted.

Also importantly, a sequence of random variables $\boldsymbol{X}$ can be raw data, or parameters in a model.

Example: Coin Flips

An example of an exchangeable sequence of random variables is a sequence of coin flips. All coin flips have the same probability distribution. The probability of an observation is independent of the position of the coin flip in the sequence of flips.

Example: Clinical Trial on Males and Females

For an example of a non-exchangeable sequence of random variables, consider the scenario where a sequence of measurements is made of a drug's impact, first on 10 females, and then on 10 males. In particular, the distribution of data for females is likely to be different than the distribution of data for males. **The key here is that we have additional information that distinguishes the sequence of random variables.** To restore exchangeability, we can include this additional information into the model, perhaps using regression or a mixture model.

1.1 List of Key Takeaways

We list the key takeaways of exchangeability to applied Bayesian modeling here, and in following sections we dive into each takeaway at a greater level of detail.

1. Placing a prior distribution over random variables (data or parameters) assumes exchangeability under that prior.
   - Specifically, "placing a prior" means representing the random variables as random samples from the prior.
2. More generally, placing a model over random variables assumes exchangeability under that model.
   - Specifically, "placing a model" means representing the random variables as random samples from the model.
3. When random variables are not exchangeable, they can often be made **conditionally exchangeable** by including additional information into the model. One common way to do this is by adding features in a regression model.

4. Training a Bayesian model on $(data_{old}, outcomes_{old})$ and using it to predict the outcomes for $data_{new}$ makes the assumption that $data_{old}$ and $data_{new}$ are exchangeable.

5. In regression applied to real-world data, it is nearly impossible to condition on enough information to achieve true exchangeability. There may always exist additional relevant information that distinguishes our data samples. Instead, we settle for a reasonable compromise in applied modeling.

2 Relation to IID

The IID property (independent, identically distributed) is sufficient but not necessary for exchangeability. That is, IID implies exchangeability, but exchangeability does not imply IID. Exchangeability is therefore more general than IID.

We show here that an IID sequence of random variables implies that it is also exchangeable:

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, ..., X_N = x_n) &= P(X_1 = x_1, X_1 = x_2, ..., X_1 = x_n) \\ &= \prod_{i=1}^{n} P(X_1 = x_i) \\ &= \prod_{i=1}^{n} P(X_1 = \pi(x_i)) \end{aligned} \tag{1}$$

Where to prove 1 we use the definition of identically distributed, the definition of independence, and finally the commutative property of multiplication.

However, exchangeability does not imply IID.

Example: Balls in an Urn

For one example, consider drawing without replacement from an urn with w white balls and b black balls. The joint probability distribution, where the sequence of random variables is an ordered sequence of draws from the urn, is invariant under permutation of the observation sequence. However, it is clear that drawing without replacement is not IID since the probability distribution for each draw depends on the previous draws.

Example: Symmetric 2D Gaussian

As another example, consider a 2D Gaussian defined for variables X, Y centered at $(x = 0, y = 0)$, with $\sigma_x = \sigma_y = 1$ and correlation $\rho \in (-1, 1)$. The probability density function is:

$$p(X = x, Y = y) \propto exp\Big(- \frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\Big) \tag{2}$$

From the probability density function, it is not too hard to see that X, Y are exchangeable. However, they are not IID unless $\rho = 0$. A graphical intuition that may be helpful is that exchangeability demands symmetric probability density about the diagonal $x = y$ line of reflection.

3 Exchangeability and Priors/Models

Representing a sequence of random variables $\boldsymbol{X}$ as random samples from a shared prior distribution D makes the assumption that $\boldsymbol{X}$ is **conditionally IID** given D.

By extending proof 1, it follows that placing a prior distribution over $\boldsymbol{X}$ make the assumption that $\boldsymbol{X}$ is exchangeable under the prior distribution D.

Similarly, representing a sequence of random variables $\boldsymbol{X}$ as random samples from a shared model M makes the assumption that $\boldsymbol{X}$ is **conditionally IID** given M. And by the same reasoning, placing a model over $\boldsymbol{X}$ make the assumption that $\boldsymbol{X}$ is exchangeable under the model M.

A hierarchical model makes many nested assumptions about exchangeability - the data is exchangeable given the entire model, and each layer of parameters is exchangeable given the submodel that generates them.

4 The General Representation Theorem

This theorem, loosely stated, has *de Finetti's Theorem* at its core, which will not be discussed here.

Theorem 2. *If $\boldsymbol{X} = (X_1, X_2, ..., X_N)$ is an exchangeable sequence of random variables, then* ***there exists*** *a parametric model $p(x|\theta)$ with parameters $\theta \in \Theta$, and* ***there exists*** *a probability distribution $p(\theta)$, with density $d(\theta)$ such that*

$$p(x_1, x_2, ..., x_N) = \int_{\Theta} \prod_{i=1}^{N} p(x_i|\theta) p(\theta) d(\theta)$$

Note: Here we return to canonical shortened notation, where $p(X_i = x_i)$ is denoted $p(x_i)$.

Theorem 2 has powerful implications for applied Bayesian modeling - it states that exchangeability implies that our sequence of random variables X are random samples from a shared parameterized model (parameterized by θ) and also implies the existence of a prior distribution over θ. Theoretically, this is exciting because it links the somewhat weak assumption of exchangeability to the existence of a true underlying Bayesian model, providing a motivation for the approach of Bayesian modeling!

Put differently, $\boldsymbol{X}$ is exchangeable under some true underlying model M^* which we define to represent the true parametric model $p(x|\theta)$ and true prior $p(\theta)$. This model M^* can be a distribution, or it could be a hierarchical model consisting of multiple distributions, each with its own parameters. All the parameters are bundled together into θ.

When we place a user-chosen model $\tilde{M}$ over $\boldsymbol{X}$, we make the assumption that $\boldsymbol{X}$ is exchangeable under $\tilde{M}$. However, $\boldsymbol{X}$ is not truly exchangeable under our chosen model $\tilde{M}$ unless $\tilde{M}$ is equivalent to M^*,

Intuitively, the act of choosing a model is a "guess" at the true underlying model - it makes the assumption that $\boldsymbol{X}$, data or parameters, is exchangeable under our chosen model, which may or may not actually be true.

Similarly, the act of choosing a prior is a "guess" at the true underlying prior - it makes the assumption that $\boldsymbol{X}$, in this case representing parameters, is exchangeable under our chosen prior, which may or may not actually be true.

5 Conditional Exchangeability and Modeling

Exchangeability is broken when you have information that can be used to distinguish your measurements. However, exchangeability can be restored by conditioning on the information that distinguishes the measurements.

Data/parameters can be **conditionally exchangeable**, which means they are exchangeable only after conditioning on some information. In practice, this means including additional information into your model.

5.1 Clinical Trial Example

Consider the example from the beginning of this lecture, where a sequence of measurements is made of a drug's impact, first on 10 females, and then on 10 males. This data is unlikely to be exchangeable.

To perform Bayesian analysis, we need to build a model where these random variables are exchangeable. We achieve this by including gender information into the model.

Note: Mixture models and regression are discussed in greater depth in later material.

Example Approach 1: Mixture Model

Here, it may make sense to model the females and males separately, and also model the probability of a study participant's gender. This could look like:

$$\begin{aligned} y_i &\sim N(\mu_{z_i}, \sigma) \\ z_i &\sim bernoulli(\lambda) \\ \lambda &\sim weakly-informative() \end{aligned} \tag{3}$$

where y_i is the treatment outcome, and z_i is a binary variable indicating gender (male is 0, female is 1), and λ has a weakly informative prior that is updated based on the ratio of females to males observed in the data. We have two different μ values, μ_0 and μ_1, indicating a difference in the mean treatment effect of males and females. We assume here that σ is the same for both genders.

Under this model, the data is exchangeable.

Example Approach 2: Regression

The ordinary Bayesian regression model is:

$$Y \sim N(XW, \sigma^2 I) \tag{4}$$

For any Y_i, $E[Y_i] = X_i W$ - the mean of Y_i depends linearly on its associated features X_i. The observations Y_i are normally distributed about this mean, with variance σ^2.

We can include gender as an additional binary feature in X. After inference, the weight assigned to the gender feature would inform us of the difference in treatment effect that gender is responsible for.

5.2 Assumptions in Modeling

Interpreting modeling results at face value makes several crucial assumptions:

1. All sources of variation that are relevant for predicting the observation are accounted for in the model
    - Variation that is unassociated with the prediction task is commonly captured by some σ in the model.
2. The specified model $\tilde{M}$ is equivalent to the true model M^*

Point 2 may not be true if, for example, you assume linearity in regression ($E[Y] \sim XW$) but the true relationship is non-linear.

In practice, including all sources of relevant variation (achieving true conditional exchangeability) can be challenging to satisfy. As a direct result, it is ill-advised to interpret modeling results directly at face value, especially in regions of input space that are poorly represented in your training data. All models are wrong, only some are useful - it is useful to have doubt in your model and understand its failure modes.

In practice, it is preferred to err on the side of including many features rather than too few. For instance, a well-built Bayesian regression model will inform you when features are irrelevant to the prediction task by shrinking the weights of those features to 0, see discussion on sparsity-inducing priors in section 1 and Bayesian variable selection in section 4.

Misspecified models can be viewed as asserting some untrue exchangeability assumption about your data or parameters. Under some conditions, fixing exchangeability problems / fixing a misspecified model comes down to including additional information in the model. Depending on your application, this can require costly data collection.

Gelman provides another perspective on the same ideas in Bayesian Data Analysis 3:

> ”When making predictions we are assuming that the old and new observations are exchangeable given the same feature values x, so that that the vector x contains all the information we have to distinguish the new observation from the old (this includes, for example, the assumption that time of observation is irrelevant if it is not encoded in x).”
>
> ”As with exchangeability in general, it is not required that all the data samples be identical or even similar, just that all relevant knowledge about them be included in x. The more similar the units are, the lower the variance of the regression will be, but that is an issue of precision, not validity.”

6 Diagnosing Model Misspecification

In general, if your residuals have structure, your model is likely to be misspecified. Checking residual plots can be useful during model checking.

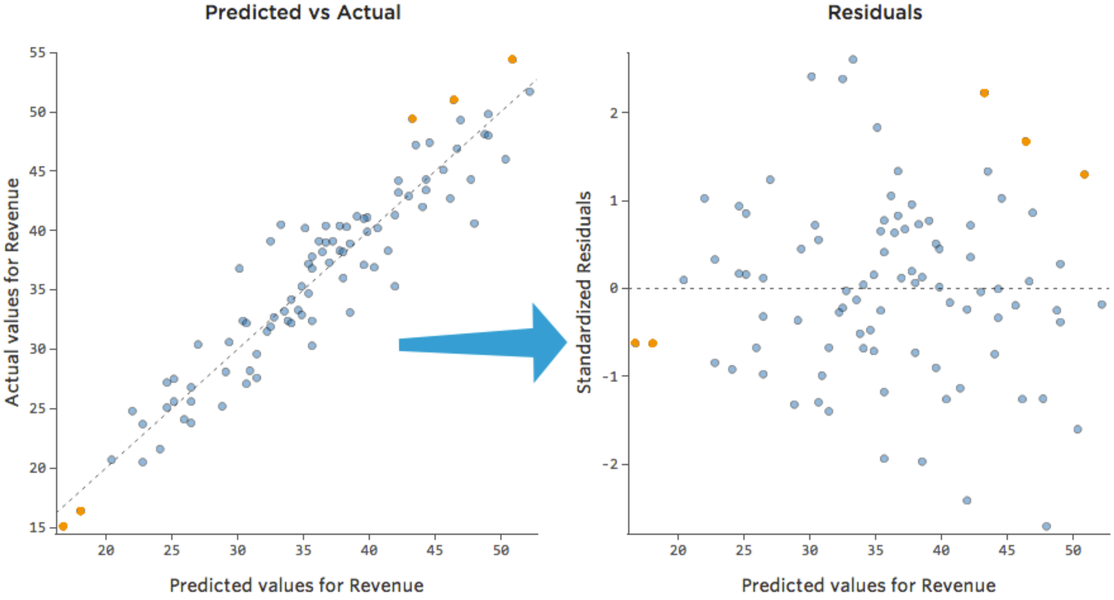

Figure 2: Residual Plot

Exercise 1

Making residual plots is a useful tool for quickly uncovering structure that may hint towards model misspecifications. If you are given a feature x_i with true "answer" y_i, and your model makes prediction $\hat{y}_i$, then a residual plot has $\hat{y}_i$ on the x-axis, and $y - \hat{y}_i$ on the y axis.

With a perfect model, $y - \hat{y}_i = 0$ for all i, so your residual plot is the horizontal line at 0.

Consider observed data X drawn uniformly between 0 and 10. The true model is:

$$y \sim N(x^2, 0.1) \tag{5}$$

You have observed data X and observed outcomes Y from the true model, and you use them to train your model:

$$\begin{aligned} y &\sim N(wx, \sigma) \\ w &\sim \text{weakly-informative}() \\ \sigma &\sim \text{weakly-informative}() \end{aligned} \tag{6}$$

Draw on paper the expected residual plot. Describe one false exchangeability assumption asserted by your model 6.

7 Bibliography and Additional Resources

Chapter 5 of Bayesian Data Analysis 3 discusses exchangeability, and Chapter 14 discusses Bayesian regression.

The Stan reference is available here: http://mc-stan.org/documentation/

See this paper for more information on the theory of exchangeability.

`http://www.uv.es/∼bernardo/Exchangeability.pdf`