August 26

We specify the distribution for some or all the parameters. These parameters have a prior distribution. The summaries of the posterior distribution as known as Bayesian estimates.

In some cases, Bayesian estimates are easier for parameters that do not follow a normal distribution.

We can estimate different kind of models with Bayesian statistics. Then, we can compare them to choose the one that better represents the data. Some of the Bayesian analysis are computationally less intensive than other approaches.

Bayesian is good for small samples. Under the traditional approach, we assume that as the sample size increases, the precision does it, which is reflected in small standard errors.

MCMC (solves hard problems) is for Bayesian what ml is for traditional approaches:

1. Missing data
2. Lack of software capable of handling large sized analysis

Bayes tries to impute the latent variable. The posterior distribution aids to sample new values for the parameters in the solution.

Parameters > random > prior distribution.

Models specify likelihood function. Bayes adds one additional term that represents the prior distribution of each parameter.

Prior distributions express our beliefs or best guesses about the data. We combine the observed data with the prior distributions to form the posterior distribution. Posterior distributions will be different than prior distributions; they update our beliefs (beliefs and evidence)

The use of certain prior distributions can produce results that are biased. It is important to elicit what are our prior beliefs and justify them appropriately.

MCMC seeks for convergence near the mean of the posterior distribution. Its precision depends on the starting point.

The probability of a 1 is the parameter of interest (p1).

The data of the likelihood is:

Diagram

Description automatically generated with medium confidence

Based on the data we can get an estimate of the likelihood of the parameter we are interested in. The mean here:

Chart

Description automatically generated with low confidence

SD of the prior distribution expresses the strength of our beliefs. The values of Beta distribution affect SD estimation.

A picture containing diagram

Description automatically generated

August 31/22

Beta parameters:

Table

Description automatically generated with medium confidence

The variance of the prior distributions turns them in informative or uninformative according to their values (smaller is better, but we need more certainty as we decrease the variance).

The posterior distribution is also a Beta distribution, whose parameters use the hyperparameters alfa and beta. A prior distribution that leads to a posterior distribution of the same family is known as conjugate prior. However, this is not a prerequisite.

The posterior distribution is:

A picture containing diagram

Description automatically generated

Combining the info for the Beta mean and sd, and updating the Bayes’ theorem, we can get the following probabilities:

Graphical user interface, text, application

Description automatically generated

The posterior distribution updates the analysis for iterative steps. The new info affects the estimation of the future posterior distributions.

End of Intro to Bayesian Concept

September 07

Models = theoretical projections of reality. It aids to describe the reality.

Measurement models = they provide the link between the latent variable and the observed data. We need to know the distribution of some parameters as part of the model.

Structure models: link between one or more latent traits and indicators/items. Are models for the distribution of the latent trait.

IRT fits a very specific type of data with a very specific type of latent trait.

Measurement models need:

* Distributional assumptions about the data (with link functions) -> outcomes (response scores) are distributed in some specific way.
* A linear or non-linear form that predicts data from the trait(s) -> link functions that transform the conditional

The key: Observed data are being predicted by latent variable(s) -> we must check if the prior fit the data and if our model measures what it is intended to do. We compare the assumptions (definition of the relationship between the latent trait and the indicators) with the data. Under the Bayesian framework, it is the relationship between the a priori distribution of the parameter of interest and the data likelihood, which combine to produce the posterior distribution.

The classical Add Stuff Up models rely on parallel items: there is an implied measurement model with very strict assumptions (equal variances and covariances across items).

Latent variables can have different levels of measurement -> interval is common -> no absolute zero and equal distances across the range of values.

Measurement models can be wrong in the amount of error they include in the prior distributions.

Defining the metric is the first step in a latent variable model, called scale identification

We must decide the values of the mean and the standard deviation for the latent trait (scale identification -> factor standardization or item marker) (end of slide 25)

September 09

Distributions must be described explicitly in the syntaxis. The model must be translated into parameters and distributions.

CFA assumes that the latent trait and residuals, thus the response scores, follow a normal distribution.

Regardless of the scale identification for factors, the shape of the estimation will be the same.

If we assume a normal distribution for each outcome (response score), the data likelihood [p(x|0)] can assume that multivariate will work. We could also use a Bernoulli for each outcome.

If the posterior distribution of the latent trait changes, the factor loadings do the same. The mean is shifted, and the new estimations will be different. Bayes provides measures to test the uncertainty or precision achieved.

Distribution of data + link functions. (end of Lecture 2)

de Finetti’s theorem shows that, from an assumption of exchangeability, we can arrive at representation that is entirely consistent with the Bayesian approach to modeling in which the parameter θ is treated as random and modeled via a distribution.

A yellow sign with black text

Description automatically generated with medium confidence

What de Finetti’s theorem shows is that if we view the variables as exchangeable, we can parameterize this joint distribution—indeed, our beliefs—by specifying the variables as conditionally independent given a parameter, and a distribution for that parameter.

September 14

Data and model complexity affect the number of iterations needed to achieve convergence.

MCMC aids to form the density of the posterior distribution. Each density plot is analyzed individually.

Conjugate prior distributions are not needed to run Bayesian analysis. Sometimes, we do not have idea about the distribution the random variable follows.

Optimization -> optimize the most possible value for the parameter of interest.

“sketch” the posterior by sampling from it – then use that sketch to make inferences -> In part magic, in part maths.

Most of the time, posterior distributions are not easily obtainable. In such cases, we can simulate data using MCMC.

Each iteration produces one value for the parameters of interest. We start with a random value for each parameter, potentially based on previous information, which is updated in posterior simulations.

MCMC is useful to consider the dependency of the parameters on the posterior distribution.

Indirect algorithms: use a value from a proposal distribution. We test the acceptance of the candidate.

Efficiency: how quickly we go from the starting point towards the convergence/right value. How correlated the residual are across simulations?

Burnin -> the algorithm starts at certain point and then it changes. At some point, the sampling is not directly from the posterior. We do not consider these values at the end of the chain. It is like a resting period for the simulation.

At the end of the chains, we will have individual values for each parameter, considering the dependency across parameters. Once the simulation starts, it includes all the parameters of interest. This include the burnin period.

Data likelihood is fixed. What change across simulations are the values for the parameters we are trying to estimate.

Unconditionally normal is not the same than conditionally normal. The shape of the distribution can change once the predictors are included.

Bayesian analysis -> MCMC (6 steps)

September 16

Burn-in period: start point until some kind of convergence. We discard these iterations from the final summary estimates.

Tuning: most of the time is defined within the burn-in period. Accept-reject decisions. The proposed distribution could be linked to the posterior. We are tuning to achieve a stable point or convergence (less variability).

C++ is a compile language.

We can use prior distributions for the slopes.

The order of the parameters in the model defines their position in the iteration process. We can use seeds for reproducibility.

Vector -> array of floating numbers (it does not matter if the data contains only integers)

Missing values: stan does not model MD. The position of the MD is not important, we can use a vector to indicate their quantity.

Sigma -> sd. In the model part, we assign means and variances to the parameters:



Uniform goes from 0 to infinity, due to negative variances are inconvenient. Uniform assumes a continuous likelihood (flat line) for all the possible values for the variance. Same weight on each of the likely values.

StanData defines the objects that the model (model00) will use during the MCMC iteration.

Parallel uses a core to run a chain independently.

September 21

Posterior distributions are not always normal, they do not look like a bell.

ess: Hoy many parameters or draws are effectively independent. To reduce the dependency, we should change the prior distribution of the parameters and number of iterations.

Results interpretation: iterations from the warm-up period are discarded.

The hypothesis testing in Bayesian is different than in ML.

The posterior distribution of variance tends to be skewed. In those cases, we could report the posterior mode. In addition, we can report a range (quantiles)

The highest probability density (credible interval) defines the probability (95%) that the posterior parameter is within the limits of the range. Middle x-x percent

The summary of the posterior distribution comes from the valid draws (model00.samples$draws)

HDPI -> highest density posterior distribution. Smallest x-x interval.

Uniform distribution -> all possible values for the parameter have the same probability

End of lecture 3 (introduction to MCMC and Stan)

September 23

Variability is better represented in Bayesian analysis than in ML or OLS.

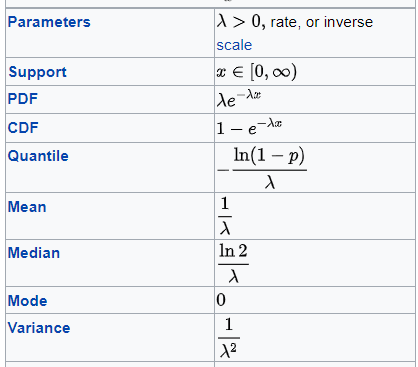
The estimation of the parameters is dynamic. If sigma wiggles, thus standard errors of the other parameters do the same.

Some of the parameters that are fixed in ML are randomly estimated in Bayes. If we fix sigma, so the variation in the other parameters will be drastically reduced.

In Bayes we don not have to rely on asymptotic convergence for standard errors.

The default distribution in Rstan is uniform. For sigma, we can use the exponential distribution (λ).

A picture containing diagram

Description automatically generated

Prior: as flat as possible, considering the mean and variance of the prior distribution.

September 28

The MCMC estimation method uses the posterior distribution once the chains converged to estimate the means of the parameters of interest. Thus, we have 4 parallel chains that converge and produce the estimates.

The more uninformative the priors, the greater the influence of the data likelihood. Doing this, the effect or influence of the prior on the posterior will be small.

Vector, array of numbers. Matrix, array of vectors (collection).

Stan <- data matrix and coefficients vector (optional, row/column vectors):

A picture containing graphical user interface

Description automatically generated

X`

September 30

For functions of parameters, such as R2 or interactions, they are computed at the end of each iteration (from piles of draws).

Easy way: Stan computes these values along together with the chains; Hard: we can take the posterior summary and calculate the value.

October 5

Formative assessment 6: generated quantities

No class next week

Does my data fit the data well? Simulated data should be similar to observed data (summary statistics)

Model fit tends to be good in univariate regression models

Select parameters from a single draw. Simulate data based on those parameters and sigma. Compute summary statistics with the simulated data (mean, sd). Compare the simulated values with the observed values.

We could calculate the sd of the simulated values in addition to the PPP. Btw, any approach can serve as an additional interpretation to the visualization inspection.

Generated quantities: rng can generate normal data (rnorm as usual). *After each sampling iteration is done, the specified code is run*. X\*beta produce the predicted values; beta (vector in parameters section) will change at the end of each sampling iteration, and it will be used in the generated quantities section.

DIC = deviance information criterion: penalty of the number of parameters based on the model average likelihood.

October 7

Relative fit: DIC, how well the model fits the data. We use it to compare models.

Expected number of parameters:



Pd = Average posterior model data likelihood - Posterior mean for all parameter estimates

We are looking for the lowest DIC:



The problem with this information criteria is its inconsistency with categorical predictors.

We want models that fit the data well and are parsimonious.

PPMC and PPP have no penalty for model parameters. Bayesian model fit takes into account parameter differences across models.

LOO tries to approximate the LOO cross-validation across the entire posterior distributions.

LOO gives warnings when less reliable to use. AIC and BIC do not provide standards errors for comparison.

Calculate the likelihood for each observation given parameter values (the log of the normal height of each person conditional on the parameter estimates from simulated data):

array[N] real log\_lik; // per person in the observed data set

for (person in 1:N){

log\_lik[person] = normal\_lpdf(y[person] | y\_pred[person], sigma);

log value of the normal PDF where observed value is, conditional on posterior estimates and variance.

LOO in Stan:

Text

Description automatically generated

We can observe what is the approximation of LOO CV to the entire data

Gamma = trending variance it seems like the residual standard deviations varies based on group.

Big picture: heterogeneity of variance makes the model fit the data better. In bayes we can build the models part by part.