STAT 576 Bayesian Analysis

Lecture 6: Model Checking

Chencheng Cai

Washington State University

Model Checking Methods

Goal:

- Assess the fit of the model to the data.
- Assess the fit of the model to our substantive knowledge.
- Assess the adequacy/robusteness of the model.

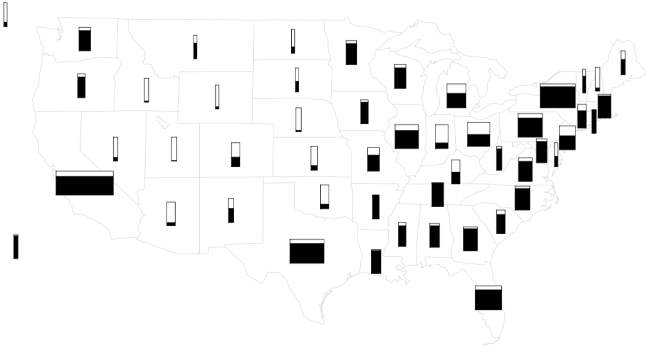
Methods:

- Sensitivity Analysis.
  - Check whether other models generate a similar posterior.
- External Validation.
  - Posterior predictive checking.
- Internal Validation.
  - Cross-validation predictive checking.

Sensitivity Analysis

- How the results are affected by different choices of the model structure?
  - different models (binomial v.s. Poisson, normal v.s. t)
  - different priors
  - different structures (hierarchical v.s. separate)
  - different distribution families (Gaussian v.s. mixed Gaussian)
- Compare the sensitivity of essential inference quantities.
  - extreme quantities v.s. mean/median.
  - extrapolation v.s. interpolation.

Example: Election Prediction



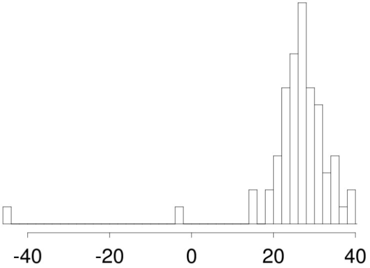
- Posterior winning probability of Bill Clinton at each state in Oct. 1992.
- Hierarchical linear regression model.
- The model seems wrong at Texas and Florida.
- It is much easier to evaluate the performance afterwards.

Posterior Predictive Checking

- Idea: check the discrepancy between the predicted values and the observed values.
- Procedure:
  - Generate simulated samples from the **joint posterior predictive distribution**
  - Compare the samples with the observed data.
  - Systematic differences imply the failings of the model.

Example: Light Speed Experiment

- Simon Newcomb set up an experiment in 1882 to measure the light speed.
- The travel time of light was recorded for the round-trip between
  - his lab on the Potomac river
  - a mirror at the base of the Washington Monument
- The total travel distance is 7422 meters.
- The measurement was repeated $n = 66$ times.



Histogram for deviations from 24800 ns

Example: Light Speed Experiment

- We model the travel time by a normal distribution:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

- We can choose a noninformative prior for μ and σ^2:

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

- Recall our previous results for multiparameter Bayesian inference. The marginal posterior for μ is

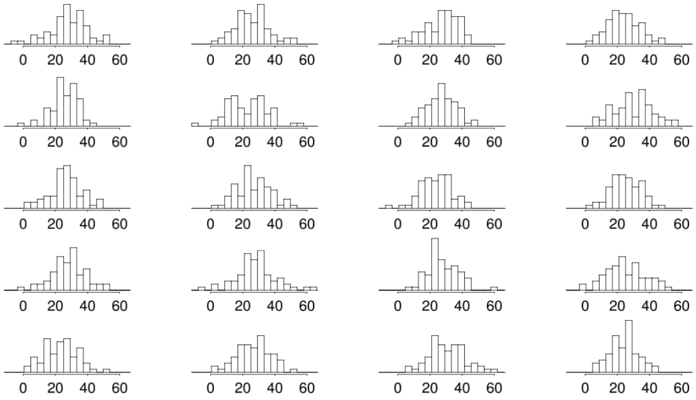
$$\mu \mid y \sim t_{66}\left(\bar{y}, \frac{65}{66^2} s^2\right)$$

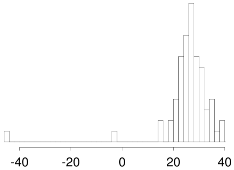
- A 95% credible interval is $[23.6, 28.8]$.
- We know the true value should be around 33.0.

Example: Light Speed Experiment

Generate posterior predictive replicates y^{rep}

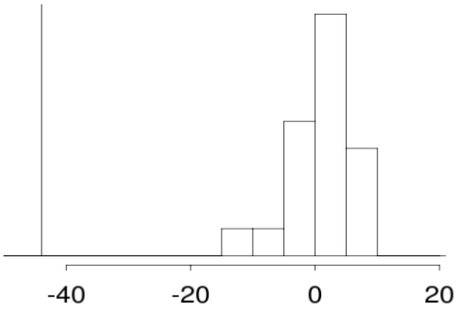
- Draw $\mu^{(s)}, \sigma^{2(s)}$ from the joint posterior distribution $p(\mu, \sigma^2 \mid y)$.
- Draw $y^{rep(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{2(s)})$.
- Repeat the drawing to get n replicates of y^{rep}.





Example: Light Speed Experiment

We get the histogram of the **smallest** travel time for all replicates.



- Can hardly observe an occurrence that is less than -20.
- Decide: whether the **data** was wrong or the **model** was wrong?
- The model was wrong: should use heavy-tailed distribution or contaminated normal (mixed Gaussian).

Posterior Predictive Checking

- Replicated datasets:

$$p(y^{rep} \mid y) = \int \underbrace{p(y^{rep} \mid \theta)}_{obs.\ model} \underbrace{p(\theta \mid y)}_{posterior} d\mu(\theta)$$

- **Test quantity** (or discrepancy measure) $T(y, \theta)$
  - Summary quantity for the observed data $T(y, \theta)$
  - Summary quantity for a replicated data $T(y^{rep}, \theta)$.
- The frequentist counter-part is known as **test statistics** $T(y)$, which only depends on the data.
- In the light speed example, we choose $T(y, \theta) = \min(y)$ (also a test statistic).

Posterior Predictive Checking

Classical p-values:

$$p_C = \mathbb{P}[T(y^{rep}) \geq T(y) \mid \theta]$$

Posterior predictive p-values:

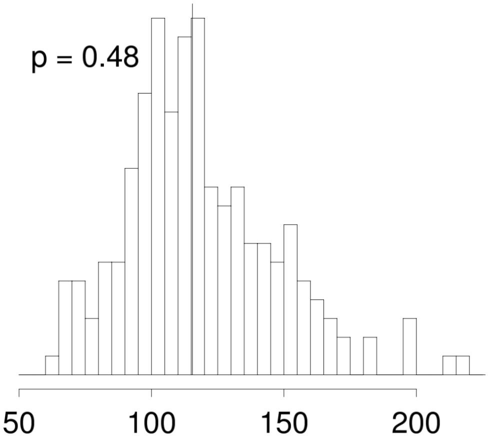
$$p_B = \mathbb{P}[T(y^{rep}, \theta) \geq T(y, \theta) \mid y]$$

- The classical p-values measure how likely the data is coming from the null model.
- The posterior predictive p-values measure how likely the data is similar to the postetior predictive replicates.
- In Bayesian, θ is also random. p_B can be estimated by joint samples of (y^{rep}, θ).

$$\begin{aligned} p_B &= \iint \mathbb{I}\{T(y^{rep}, \theta) \geq T(y, \theta)\} p(y^{rep} \mid \theta) p(\theta \mid y) d\mu(\theta) d\mu(y^{rep}) \\ &\approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\{T(y^{rep(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)})\} \end{aligned}$$

Example: Light Speed Experiment

If we use the sample variance as the test quantity:



Cannot tell the discrepancy — because the sample variance is a sufficient statistics.

Posterior Predictive Checking

- A **good** test statistic is ancilliary
  - ancilliary: depends on the observed data but independent of the parameters.
- A **bad** test statsistic is highly dependent of the parameters.
  - i.e. sufficient statistics.
- If we have multiple test statistics, we do not conduct p-value justification.
  - See the smoking example in the textbook.
- An extreme p-value often suggests the weakness of the current model. The next step is to revise the model.

Example: Educational Testing

Data: the effects of coaching programs for the SAT-V scores for students in 8 schools.

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18

Example: Educational Testing

Separate estimation:

- Some schools have moderate effects (18-28).
- Most schools have small effects (0-12).
- Two have negative effects.
- Difficult to distinguish because of large variance.

Pooled estimation:

- All schools have identical effect θ.
- Use noninformative prior.
- Posterior mean: 7.7 with s.e. 4.1

Hierarchical model:

- $\theta_1, \ldots, \theta_8 \sim \mathcal{N}(\mu, \tau^2)$ i.i.d.
- $y_j \mid \theta_j \sim (\theta_j, \sigma_j^2)$ independent.
- choose flat prior $p(\mu, \tau) \propto 1$.

Example: Educational Testing

Hierarchical model:

- By drawing posterior samples:
  - draw $\mu^{(s)}, \tau^{(s)}$ from $p(\mu, \tau \mid y)$
  - draw $\theta_1^{(s)}, \dots, \theta_8^{(s)}$ from $p(\theta_1, \dots, \theta_8 \mid \mu^{(s)}, \tau^{(s)}, y)$
- we have the posterior quantiles for each school:

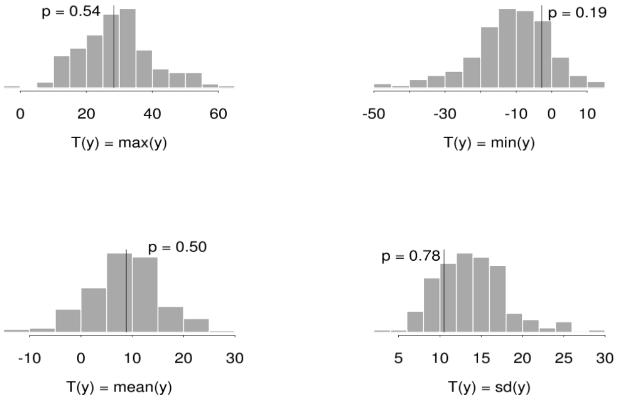
School	Posterior quantiles				
	2.5%	25%	median	75%	97.5%
A	−2	7	10	16	31
B	−5	3	8	12	23
C	−11	2	7	11	19
D	−7	4	8	11	21
E	−9	1	5	10	18
F	−7	2	6	10	28
G	−1	7	10	15	26
H	−6	3	8	13	33

Example: Educational Testing — Model Checking

- Assumptions:
  - normality of y_j.
  - exchangeability of the priors for θ_j's.
  - normality of prior of θ_j.
  - flat hyperprior.
- Comparing posterior inferences to substantive knowledge:
  - Individual effects between 5 and 10 seems reasonable.
  - Some lower bounds go to negative.

Example: Educational Testing — Model Checking

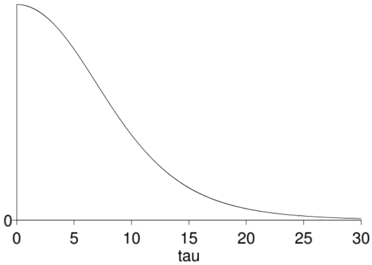
- Posterior predictive checking.
  - $y^{rep} = (y_1^{rep}, \ldots, y_8^{rep})$
  - Test statistics: max, min, mean, s.d.



Example: Educational Testing — Model Checking

Sensitivity Analysis:

- Uniform prior for τ: the marginal posterior for τ
  — no significant change if we multiply it by another prior



- normality of $y_j \mid \theta_j, \sigma_j$: ensured by experimental designa and CLT.
- normality of the prior for θ_j's:
  One may consider other heavy-tailed distributions. But needs advanced sampling techniques.

Model Evaluation

- We need certain criterion in evaluating a model.
- — provide a "perfomance measure" of the model
- — provide a standard for comparing models
- A very intuitive way is to compare the predicted values with the true values.

Prediction Accuracy

Compare y_i (observation) with prediction:

- if the prediction is a **point prediction** $\hat{y}_i$:
  - mean squared error: $n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - mean absolute error: $n^{-1}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- if the prediction is a **probabilistic prediction** $p(y_i \mid \theta)$:
  - log-predictive density (lpd): $n^{-1}\sum_{i=1}^{n}\log p(y_i \mid \theta)$

Justification

If we have the true distribution F (with density f) such that $y_1, \ldots, y_n \sim F,\ i.i.d.$.
Then

$$\text{lpd} = \frac{1}{n}\sum_{i=1}^{n}\log p(y_i \mid \theta) \xrightarrow{a.s.} \mathbb{E}_F[\log p(y_i \mid \theta)] = \int f(y)\log p(y \mid \theta)d\mu(y)$$

$$= \underbrace{\int f(y)\log f(y)d\mu(y)}_{\text{neg. entropy of } F} - \underbrace{\int f(y)\log\frac{f(y)}{p(y \mid \theta)}d\mu(y)}_{\text{Kullback-Leibler divergence } \mathrm{KL}(f||p_\theta)}$$

Prediction Accuracy — Log-Predictive Density

Notation:

- y: observed data
- $\tilde{y}$: a new data
- F: the true model of y with density f.

The **posterior predictive density** for $\tilde{y}_i$ is

$$p(\tilde{y}_i \mid y) = \int p(\tilde{y}_i \mid \theta) \underbrace{p(\theta \mid y)}_{\text{posterior}} d\mu(\theta) = \mathbb{E}_{\text{post}}[p(\tilde{y}_i \mid \theta)] = p_{\text{post}}(\tilde{y}_i)$$

- $\mathbb{E}_{\text{post}}$ is the expectation is taken for θ w.r.t. the posterior.
- $p_{\text{post}}(\tilde{y}_i)$ is the predictive density for $\tilde{y}_i$ induced from the posterior $p_{\text{post}}(\theta)$.

The **expected predictive density** for $\tilde{y}_i$ is

$$\text{elpd} = \mathbb{E}_F[\log p_{\text{post}}(\tilde{y}_i)] = \int f(\tilde{y}_i) \log p_{\text{post}}(\tilde{y}_i) d\mu(\tilde{y}_i)$$

Prediction Accuracy — Log-Predictive Density

Bayesian version: the expected predictive density:

$$\mathbb{E}_F[\log p_{\textsf{post}}(\tilde{y}_i)] = \int f(\tilde{y}_i) \log p(\tilde{y}_i \mid y) d\mu(\tilde{y}_i)$$

Frequentist version: the expected predictve density given $\hat{\theta}$:

$$\mathbb{E}_F[\log p(\tilde{y}_i \mid \hat{\theta})] = \int f(\tilde{y}_i) \log p(\tilde{y}_i \mid \hat{\theta}) d\mu(\tilde{y}_i)$$

The connection is given by

$$p(\tilde{y}_i \mid y) = \int p(\tilde{y}_i \mid \theta) p(\theta \mid y) d\mu(\theta)$$

Prediction Accuracy — Evaluation

- In practice, we do not know $\theta \longrightarrow$ we cannot calculate $\log p(y_i \mid \theta)$.
- Instead, we work with an averaged version w.r.t. $\theta \sim p(\theta \mid y)$ (the posterior).
- We summarize the predictive accuracy of the fitted model to data by the **log pointwise predictive density**:

$$\mathrm{lppd} = \log \prod_{i=1}^{n} p_{\mathsf{post}}(y_i) = \sum_{i=1}^{n} \log \int p(y_i \mid \theta) p_{\mathsf{post}}(\theta) d\mu(\theta)$$

- It is called "pointwise" because we ignore any dependence structure between the observations and only compute the marginal.
- If we don't have a closed-form for the integral, we can draw $\theta^{(1)}, \ldots, \theta^{(S)} \sim p_{\mathsf{post}}(\theta)$ i.i.d., and

$$\widehat{\mathrm{lppd}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \theta^{(s)}) \right)$$

Prediction Accuracy — Estimation

- We want to estimat the expected predictive accuracy using **out-of-sample** data.
- Several methods can be used to estimate the out-of-sample predictive accuracy by the existing data.
  - **Within-sample predictive accuracy**: use the log predicitive density on the training data.
  - **Adjusted within-sample predictive accuracy**: adjust the within-sample predictive accuracy by the expected overestimation. Also known as **information criterion**.
  - **Cross-validation**: split training and testing data and estimate the predictive accuracy on the testing data.

Akaike Information Criterion (AIC)

In classical inference (frequentist version), the goal is to estimate the expected out-of-sample predictive accuracy conditioned on $\hat{\theta}$:

$$\text{epld} = \mathbb{E}_F[\log p(\tilde{y} \mid \hat{\theta})]$$

It is estimated by

$$\widehat{\text{epld}}_{\mathsf{AIC}} = \log p(y \mid \hat{\theta}_{\mathsf{mle}}) - k$$

where k is the number of parameters in the model. Or equivalently, we define

$$\text{AIC} = -2 \log p(y \mid \hat{\theta}_{\mathsf{mle}}) + 2k$$

Why $-k$ in estimated epld (or $2k$ in AIC)?

- Overestimation from using in-sample data

$$\log p(y \mid \hat{\theta}_{\mathsf{mle}}) - \frac{k}{2} \approx \mathbb{E}_F[\log p(\tilde{y} \mid \theta_0)] \approx \mathbb{E}_F[\log p(\tilde{y} \mid \hat{\theta}_{\mathsf{mle}})] + \frac{k}{2}$$

Deviance Information Criterion (DIC)

DIC is a Bayesian version of AIC:

$$\widehat{\mathsf{epld}}_{\mathsf{DIC}} = \log p(y \mid \hat{\theta}_{\mathsf{Bayes}}) - p_{\mathsf{DIC}}$$

where p_{DIC} is the effective number of parameters:

$$p_{\mathsf{DIC}} = 2\left(\log p(y \mid \hat{\theta}_{\mathsf{Bayes}}) - \mathbb{E}_{\mathsf{post}}[\log p(y \mid \theta)]\right)$$

Equivlantly, DIC is defined as

$$\mathrm{DIC} = -2\log p(y \mid \hat{\theta}_{\mathsf{Bayes}}) + 2p_{\mathsf{DIC}}$$

Watanabe-Akaike Information Criterion (WAIC)

WAIC revises DIC in two ways:

- replace $\hat{\theta}_{\text{Bayes}}$ by an average over $p_{\text{post}}(\theta)$.
- replace the joint predictive density by the point-wise version.

The effective number of parameters in WAIC is

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n}\left(\log \mathbb{E}_{\text{post}}[p(y_i \mid \theta)] - \mathbb{E}_{\text{post}}[\log p(y_i \mid \theta)]\right)$$

The estimated expected log pointwise predict density is

$$\widehat{\text{elppd}}_{\text{WAIC}} = \text{lppd} - p_{\text{WAIC}} = \sum_{i=1}^{n} \log \mathbb{E}_{\text{post}}[p(y_i \mid \theta)] - p_{\text{WAIC}}$$

Similarly, we define WAIC by

$$\text{WAIC} = -2\text{lppd} + 2p_{\text{WAIC}}$$

Comparison

- All estimators are equivelent asymptotically.
- AIC and DIC require a point estimator. WAIC does not.
- The integrals involved in DIC and WAIC need Monte Carlo simulation.
- WAIC requires a partition of the data.
- AIC and DIC requires independent errors in the observations.
- Only WAIC is fully Bayesian.

- Bayesian information criterion (BIC) has a different goal and therefore is not discussed here.

Leave-One-Out Cross Validation (LOO-CV)

The Bayesian LOO-CV estimate of out-of-sample predictive fit is

$$\text{lppd}_{\text{loo-cv}} = \sum_{i=1}^{n} \log p_{\text{post}(-i)}(y_i) = \sum_{i=1}^{n} \log \int p(y_i \mid \theta) \underbrace{p(\theta \mid y \setminus \{y_i\})}_{\text{posterior with all obs. except } y_i} d\mu(\theta)$$

- In practice, the above integral can be replaced by Monte Carlo sample mean.
- $\text{lppd}_{\text{loo-cv}}$ underestimates the predictive accuracy because it uses $n-1$ observations instead of n.
- The bias can be estimated by

$$b = \text{lppd} - \overline{\text{lppd}}_{-i}$$

where

$$\overline{\text{lppd}}_{-i} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \log p_{\text{post}(-i)}(y_j)$$

Leave-One-Out Cross Validation (LOO-CV)

The **bias-corrected** Bayesian LOO-CV is then

$$\text{lppd}_{\textsf{cloo-cv}} = \text{lppd}_{\textsf{loo-cv}} + b$$

If we compare the formula to other methods, we can the effective numbers of parameters are

$$p_{\textsf{loo-cv}} = \text{lppd} - \text{lppd}_{\textsf{loo-cv}}$$
$$p_{\textsf{cloo-cv}} = \overline{\text{lppd}}_{-i} - \text{lppd}_{\textsf{loo-cv}}$$

Example: SAT-V Score

		No pooling ($\tau = \infty$)	Complete pooling ($\tau = 0$)	Hierarchical model (τ estimated)
AIC	$-2\,\mathrm{lpd} = -2 \log p(y \mid \hat{\theta}_{\mathrm{mle}})$	54.6	59.4	
	k	8.0	1.0	
	$\mathrm{AIC} = -2\,\widehat{\mathrm{elpd}}_{\mathrm{AIC}}$	70.6	61.4	
DIC	$-2\,\mathrm{lpd} = -2 \log p(y \mid \hat{\theta}_{\mathrm{Bayes}})$	54.6	59.4	57.4
	p_{DIC}	8.0	1.0	2.8
	$\mathrm{DIC} = -2\,\widehat{\mathrm{elpd}}_{\mathrm{DIC}}$	70.6	61.4	63.0
WAIC	$-2\,\mathrm{lppd} = -2 \sum_i \log p_{\mathrm{post}}(y_i)$	60.2	59.8	59.2
	$p_{\mathrm{WAIC}\,1}$	2.5	0.6	1.0
	$p_{\mathrm{WAIC}\,2}$	4.0	0.7	1.3
	$\mathrm{WAIC} = -2\,\widehat{\mathrm{elppd}}_{\mathrm{WAIC}\,2}$	68.2	61.2	61.8
LOO-CV	$-2\,\mathrm{lppd}$		59.8	59.2
	$p_{\mathrm{loo-cv}}$		0.5	1.8
	$-2\,\mathrm{lppd}_{\mathrm{loo-cv}}$		60.8	62.8

Baysian Hypothesis Testing

Suppose we have two competing models H_1 and H_2. We put the testing in a Bayesian framework:

- Prior $p(H_1)$ and $p(H_2)$ with $p(H_1) + p(H_2) = 1$
- Likelihood: $p(y \mid H_1)$ and $p(y \mid H_2)$
- Posterior:

$$p(H_i \mid y) = \frac{p(H_i)p(y \mid H_i)}{p(H_1)p(y \mid H_1) + p(H_2)p(y \mid H_2)}, \quad i = 1, 2$$

  It is easy to verify $p(H_1 \mid y) + p(H_2 \mid y) = 1$.

- To decide, we look at the posterior ratio:

$$\frac{p(H_2 \mid y)}{p(H_1 \mid y)} = \frac{p(H_2)}{p(H_1)} \times \underbrace{\frac{p(y \mid H_2)}{p(y \mid H_1)}}_{\text{Bayes Factor}(H_2;H_1)}$$

  The decision depends on the magnitude of the Bayes Factor of the two models.

Baysian Hypothesis Testing

Common decisions based on the Bayes Factor:

Bayes factor	1 to 3.2	3.2 to 10	10 to 100	> 100
Decision	a bare mention	substantial	strong	decisive

- H_1 and H_2 are symmetric.
- When H_i is a composite assumption on θ, we have

$$p(y \mid H_i) = \int p(y \mid \theta) p(\theta \mid H_i) d\mu(\theta)$$

- There is no Type I error to control.
- The posterior directly gives the probability of hypotheses after observing the data.
- Bayes factor works better for discrete models than continuous models.