Model Checking

Patrick Lam

Outline

Outline

Prediction

Prediction

Once we have a model and generated draws from our posterior distribution, we may want to predict future data points.

Prediction

Once we have a model and generated draws from our posterior distribution, we may want to predict future data points.

We may want to make predictions in order to:

Prediction

Once we have a model and generated draws from our posterior distribution, we may want to predict future data points.

We may want to make predictions in order to:

1. Predict how a system would behave in the future (substantive implications)

Prediction

Once we have a model and generated draws from our posterior distribution, we may want to predict future data points.

We may want to make predictions in order to:

1. Predict how a system would behave in the future (substantive implications)
2. Assess model accuracy (modeling implications)

Prediction

Once we have a model and generated draws from our posterior distribution, we may want to predict future data points.

We may want to make predictions in order to:

1. Predict how a system would behave in the future (substantive implications)
2. Assess model accuracy (modeling implications)

Through simulation, we can get a **posterior predictive distribution**.

Posterior Predictive Distribution

Posterior Predictive Distribution

Predicted distribution of some future data point(s) y^{rep} after having seen the data y.

Posterior Predictive Distribution

Predicted distribution of some future data point(s) y^{rep} after having seen the data y.

$$
\begin{aligned}
p(y^{\mathrm{rep}}|y) &= \int p(y^{\mathrm{rep}}, \theta|y) d\theta \\
&= \int p(y^{\mathrm{rep}}|\theta, y) p(\theta|y) d\theta
\end{aligned}
$$

Posterior Predictive Distribution

Predicted distribution of some future data point(s) y^{rep} after having seen the data y.

$$\begin{aligned} p(y^{\text{rep}}|y) &= \int p(y^{\text{rep}}, \theta|y) d\theta \\ &= \int p(y^{\text{rep}}|\theta, y) p(\theta|y) d\theta \end{aligned}$$

If we assume $y \perp y^{\text{rep}}|\theta$, then

Posterior Predictive Distribution

Predicted distribution of some future data point(s) y^{rep} after having seen the data y.

$$\begin{aligned}
p(y^{\text{rep}}|y) &= \int p(y^{\text{rep}}, \theta | y) d\theta \\
&= \int p(y^{\text{rep}}|\theta, y) p(\theta|y) d\theta
\end{aligned}$$

If we assume $y \perp y^{\text{rep}} | \theta$, then

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|y) p(\theta|y) d\theta$$

Posterior Predictive Distribution

Predicted distribution of some future data point(s) y^{rep} after having seen the data y.

$$\begin{aligned} p(y^{\mathrm{rep}}|y) &= \int p(y^{\mathrm{rep}}, \theta|y) d\theta \\ &= \int p(y^{\mathrm{rep}}|\theta, y) p(\theta|y) d\theta \end{aligned}$$

If we assume $y \perp y^{\mathrm{rep}}|\theta$, then

$$p(y^{\mathrm{rep}}|y) = \int p(y^{\mathrm{rep}}|y) p(\theta|y) d\theta$$

If y is a vector of n observations, then y^{rep} is also a vector of length n with covariates set at the observed (model checking) or hypothetical values (prediction)

Posterior Predictive Distribution

Predicted distribution of some future data point(s) y^{rep} after having seen the data y.

$$
\begin{aligned}
p(y^{\text{rep}}|y) &= \int p(y^{\text{rep}}, \theta|y) d\theta \\
&= \int p(y^{\text{rep}}|\theta, y) p(\theta|y) d\theta
\end{aligned}
$$

If we assume $y \perp y^{\text{rep}}|\theta$, then

$$
p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|y) p(\theta|y) d\theta
$$

If y is a vector of n observations, then y^{rep} is also a vector of length n with covariates set at the observed (model checking) or hypothetical values (prediction) and $p(y^{\text{rep}}|y)$ can be thought of as an n-variate distribution.

We can simulate the posterior predictive distribution.

We can simulate the posterior predictive distribution.

1. Sample m values of θ from our posterior.

We can simulate the posterior predictive distribution.

1. Sample m values of θ from our posterior.
2. For each posterior draw, sample a value (vector) of y^{rep} from our likelihood $p(y^{\mathrm{rep}}|\theta)$.

We can simulate the posterior predictive distribution.

1. Sample m values of θ from our posterior.
2. For each posterior draw, sample a value (vector) of y^{rep} from our likelihood $p(y^{\mathrm{rep}}|\theta)$.

The m values (vectors) of y^{rep} represent draws from the posterior predictive distribution $p(y^{\mathrm{rep}}|y)$.

We can simulate the posterior predictive distribution.

1. Sample m values of θ from our posterior.
2. For each posterior draw, sample a value (vector) of y^{rep} from our likelihood $p(y^{\mathrm{rep}}|\theta)$.

The m values (vectors) of y^{rep} represent draws from the posterior predictive distribution $p(y^{\mathrm{rep}}|y)$.

We can use the posterior predictive distribution to predict the future or assess model accuracy with posterior predictive checks.

Outline

Much of what we have done so far is based on a model that we specify, which may or may not be accurate.

Much of what we have done so far is based on a model that we specify, which may or may not be accurate.

Specifically, we make many assumptions with our model which may or may not be accurate (for example, independence across observations).

Much of what we have done so far is based on a model that we specify, which may or may not be accurate.

Specifically, we make many assumptions with our model which may or may not be accurate (for example, independence across observations).

We can attempt to check specific model assumptions with **posterior predictive checks**.

Posterior Predictive Checks

Posterior Predictive Checks

To conduct a posterior predictive check, do the following:

Posterior Predictive Checks

To conduct a posterior predictive check, do the following:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing.

Posterior Predictive Checks

To conduct a posterior predictive check, do the following:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing.
2. Calculate T for the observed data y:

Posterior Predictive Checks

To conduct a posterior predictive check, do the following:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing.
2. Calculate T for the observed data y: $T(y)$

Posterior Predictive Checks

To conduct a posterior predictive check, do the following:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing.
2. Calculate T for the observed data y: $T(y)$
3. Calculate T for each y^{rep} draw from the posterior predictive distribution:

Posterior Predictive Checks

To conduct a posterior predictive check, do the following:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing.
2. Calculate T for the observed data y: $T(y)$
3. Calculate T for each y^{rep} draw from the posterior predictive distribution: $T(y^{\mathrm{rep}}|y)$

Posterior Predictive Checks

To conduct a posterior predictive check, do the following:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing.
2. Calculate T for the observed data y: $T(y)$
3. Calculate T for each y^{rep} draw from the posterior predictive distribution: $T(y^{\mathrm{rep}}|y)$
4. Calculate the fraction of times $T(y^{\mathrm{rep}}|y) > T(y)$.

Posterior Predictive Checks

To conduct a posterior predictive check, do the following:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing.
2. Calculate T for the observed data y: $T(y)$
3. Calculate T for each y^{rep} draw from the posterior predictive distribution: $T(y^{\mathrm{rep}}|y)$
4. Calculate the fraction of times $T(y^{\mathrm{rep}}|y) > T(y)$. This is an estimate of the *posterior predictive p-value*.

The idea is that if our data violates one of our model assumptions, then our observed test statistic $T(y)$ should be significantly different than our model predicted test statistics $T(y^{\mathrm{rep}}|y)$.

The idea is that if our data violates one of our model assumptions, then our observed test statistic $T(y)$ should be significantly different than our model predicted test statistics $T(y^{\mathrm{rep}}|y)$.

If our posterior predictive p-value is close to 0 or 1 (say 0.05 or 0.95), then it suggests that our observed data has an extreme test statistic and that something in our model may be inadequate.

Possible Problems with Posterior Predictive Checks

Possible Problems with Posterior Predictive Checks

- Choice of test statistic is very important.

Possible Problems with Posterior Predictive Checks

- Choice of test statistic is very important.
  - Test statistic must be meaningful and pertinent to the assumption you want to test.

Possible Problems with Posterior Predictive Checks

- Choice of test statistic is very important.
  - Test statistic must be meaningful and pertinent to the assumption you want to test.
  - Test statistics often have low power (inability to find problems when problems exist)

Possible Problems with Posterior Predictive Checks

- Choice of test statistic is very important.
  - Test statistic must be meaningful and pertinent to the assumption you want to test.
  - Test statistics often have low power (inability to find problems when problems exist)
  - Test statistics should be not based on aspects of the data that are being explicit modeled (for example, the mean of y in a linear model).

Possible Problems with Posterior Predictive Checks

- Choice of test statistic is very important.
  - Test statistic must be meaningful and pertinent to the assumption you want to test.
  - Test statistics often have low power (inability to find problems when problems exist)
  - Test statistics should be not based on aspects of the data that are being explicit modeled (for example, the mean of y in a linear model).
- If the model passes posterior predictive check, it does not necessarily mean there are no problems with the model.

Possible Problems with Posterior Predictive Checks

- Choice of test statistic is very important.
  - Test statistic must be meaningful and pertinent to the assumption you want to test.
  - Test statistics often have low power (inability to find problems when problems exist)
  - Test statistics should be not based on aspects of the data that are being explicit modeled (for example, the mean of y in a linear model).
- If the model passes posterior predictive check, it does not necessarily mean there are no problems with the model.
  - Test statistic may have low power.

Possible Problems with Posterior Predictive Checks

- Choice of test statistic is very important.
  - Test statistic must be meaningful and pertinent to the assumption you want to test.
  - Test statistics often have low power (inability to find problems when problems exist)
  - Test statistics should be not based on aspects of the data that are being explicit modeled (for example, the mean of y in a linear model).
- If the model passes posterior predictive check, it does not necessarily mean there are no problems with the model.
  - Test statistic may have low power.
  - May be testing the wrong assumption.

Possible Problems with Posterior Predictive Checks

- Choice of test statistic is very important.
  - Test statistic must be meaningful and pertinent to the assumption you want to test.
  - Test statistics often have low power (inability to find problems when problems exist)
  - Test statistics should be not based on aspects of the data that are being explicit modeled (for example, the mean of y in a linear model).
- If the model passes posterior predictive check, it does not necessarily mean there are no problems with the model.
  - Test statistic may have low power.
  - May be testing the wrong assumption.
- It is not always clear how to correct the incorrect model assumptions.

Outline

Posterior Predictive Distribution

Posterior Predictive Checks

An Example

Bayes Factor

Running Example

Running Example

Time-series cross-sectional dataset on civil war onset from Fearon and Laitin.

Running Example

Time-series cross-sectional dataset on civil war onset from Fearon and Laitin.

```
> data <- read.table("FLdata.txt")
```

Running Example

Time-series cross-sectional dataset on civil war onset from Fearon and Laitin.

```
> data <- read.table("FLdata.txt")
```

Dependent variable: binary variable on civil war onset

Running Example

Time-series cross-sectional dataset on civil war onset from Fearon and Laitin.

```
> data <- read.table("FLdata.txt")
```

Dependent variable: binary variable on civil war onset

Independent variables: the normal set of independent variables predicting civil wars

Running Example

Time-series cross-sectional dataset on civil war onset from Fearon and Laitin.

```
> data <- read.table("FLdata.txt")
```

Dependent variable: binary variable on civil war onset

Independent variables: the normal set of independent variables predicting civil wars

Model: Bayesian logistic regression with binomial likelihood and multivariate Normal priors (using `MCMCpack`)

Running Example

Time-series cross-sectional dataset on civil war onset from Fearon and Laitin.

```
> data <- read.table("FLdata.txt")
```

Dependent variable: binary variable on civil war onset

Independent variables: the normal set of independent variables predicting civil wars

Model: Bayesian logistic regression with binomial likelihood and multivariate Normal priors (using `MCMCpack`)

```
> library(MCMCpack)
```

Posterior Predictive Distribution

Posterior Predictive Distribution

1. Create model matrix of covariates X.

Posterior Predictive Distribution

1. Create model matrix of covariates X.

Posterior Predictive Distribution

1. Create model matrix of covariates X.

2. Get linear predictors by multiplying X and our m draws from the posterior.

Posterior Predictive Distribution

1. Create model matrix of covariates X.
2. Get linear predictors by multiplying X and our m draws from the posterior.

Posterior Predictive Distribution

1. Create model matrix of covariates X.
2. Get linear predictors by multiplying X and our m draws from the posterior.
3. Convert linear predictors into probabilities with the inverse logit function.

Posterior Predictive Distribution

1. Create model matrix of covariates X.
2. Get linear predictors by multiplying X and our m draws from the posterior.
3. Convert linear predictors into probabilities with the inverse logit function.

Posterior Predictive Distribution

1. Create model matrix of covariates X.
2. Get linear predictors by multiplying X and our m draws from the posterior.
3. Convert linear predictors into probabilities with the inverse logit function.
4. Draw m samples of y^{rep} from the binomial likelihood.

Posterior Predictive Distribution

1. Create model matrix of covariates X.
2. Get linear predictors by multiplying X and our m draws from the posterior.
3. Convert linear predictors into probabilities with the inverse logit function.
4. Draw m samples of y^{rep} from the binomial likelihood.

Posterior Predictive Distribution

1. Create model matrix of covariates X.
2. Get linear predictors by multiplying X and our m draws from the posterior.
3. Convert linear predictors into probabilities with the inverse logit function.
4. Draw m samples of y^{rep} from the binomial likelihood.

The resulting posterior predictive distribution is an $n \times m$ matrix.

A Bad Test Statistic

A Bad Test Statistic

Let $T =$ the fraction of y's that take on the value of 1

A Bad Test Statistic

Let $T =$ the fraction of y's that take on the value of 1

What's wrong with this test statistic?

A Bad Test Statistic

Let $T =$ the fraction of y's that take on the value of 1

What's wrong with this test statistic?

- Unclear what assumption are we testing.

A Bad Test Statistic

Let $T =$ the fraction of y's that take on the value of 1

What's wrong with this test statistic?

- Unclear what assumption are we testing.
- The fraction of 1s is explicitly being modeled in the logit model.

A Bad Test Statistic

Let T = the fraction of y's that take on the value of 1

What's wrong with this test statistic?

- Unclear what assumption are we testing.
- The fraction of 1s is explicitly being modeled in the logit model.
  - The test will never show anything is wrong regardless of how bad our model is.

A Better Test Statistic

A Better Test Statistic

Assumption: No clustering within years

A Better Test Statistic

Assumption: No clustering within years

Test Statistic: $T =$ the variance of the number of 1s in each year

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing: $T =$ the variance of the number of 1s in each year

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing: $T =$ the variance of the number of 1s in each year

2. Calculate T for the observed data y:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing: $T =$ the variance of the number of 1s in each year
2. Calculate T for the observed data y: $T(y)$

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing: $T =$ the variance of the number of 1s in each year

2. Calculate T for the observed data y: $T(y)$

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing: $T =$ the variance of the number of 1s in each year

2. Calculate T for the observed data y: $T(y)$

3. Calculate T for each y^{rep} draw from the posterior predictive distribution:

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing: $T =$ the variance of the number of 1s in each year
2. Calculate T for the observed data y: $T(y)$
3. Calculate T for each y^{rep} draw from the posterior predictive distribution: $T(y^{\mathrm{rep}}|y)$

1. Come up with a test statistic T that has power to diagnose violations of whatever assumption you are testing: $T =$ the variance of the number of 1s in each year

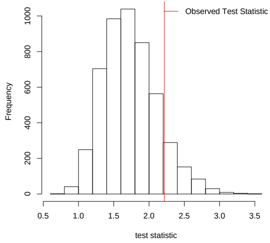
2. Calculate T for the observed data y: $T(y)$

3. Calculate T for each y^{rep} draw from the posterior predictive distribution: $T(y^{\mathrm{rep}}|y)$

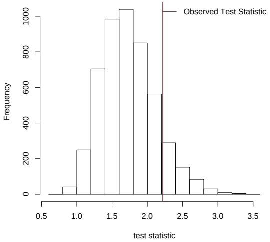
4. Calculate the fraction of times $T(y^{\mathrm{rep}}|y) > T(y)$.

4. Calculate the fraction of times $T(y^{\mathrm{rep}}|y) > T(y)$. This is an estimate of the *posterior predictive p-value*.

4. Calculate the fraction of times $T(y^{\mathrm{rep}}|y) > T(y)$. This is an estimate of the *posterior predictive p-value*.

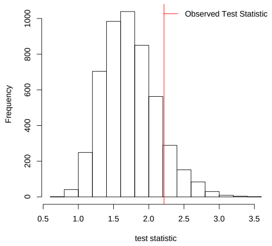


4. Calculate the fraction of times $T(y^{\mathrm{rep}}|y) > T(y)$. This is an estimate of the *posterior predictive p-value*.



Does this mean our assumption is correct?

4. Calculate the fraction of times $T(y^{\mathrm{rep}}|y) > T(y)$. This is an estimate of the *posterior predictive p-value*.



Does this mean our assumption is correct? Not necessarily (low power?)

Outline

Bayes Factor

Bayes Factor

Suppose we have two competing models/hypotheses:

Bayes Factor

Suppose we have two competing models/hypotheses: M_1 and M_2

Bayes Factor

Suppose we have two competing models/hypotheses: M_1 and M_2

Assuming that either M_1 or M_2 is true, we would like to know what the data $\mathbf{y}$ tell us about the probabilities of either model being true.

Bayes Factor

Suppose we have two competing models/hypotheses: M_1 and M_2

Assuming that either M_1 *or* M_2 *is true*, we would like to know what the data $\mathbf{y}$ tell us about the probabilities of either model being true.

That is, we would like to know $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$.

Bayes Factor

Suppose we have two competing models/hypotheses: M_1 and M_2

Assuming that either M_1 *or* M_2 *is true*, we would like to know what the data $\mathbf{y}$ tell us about the probabilities of either model being true.

That is, we would like to know $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$.

We can then compare $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$ to see which model fits the data better.

Bayes Factor

Suppose we have two competing models/hypotheses: M_1 and M_2

Assuming that either M_1 or M_2 is true, we would like to know what the data $\mathbf{y}$ tell us about the probabilities of either model being true.

That is, we would like to know $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$.

We can then compare $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$ to see which model fits the data better.

This is known as the Bayes factor approach and it is the Bayesian alternative to hypothesis testing in classical statistics.

Note that $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$ are posteriors of M_1 and M_2 respectively.

Note that $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$ are posteriors of M_1 and M_2 respectively.

In comparing the two models, we want to look at the posterior odds in favor of one model (say M_1):

Note that $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$ are posteriors of M_1 and M_2 respectively.

In comparing the two models, we want to look at the posterior odds in favor of one model (say M_1):

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{\frac{p(\mathbf{y}|M_1)p(M_1)}{p(\mathbf{y})}}{\frac{p(\mathbf{y}|M_2)p(M_2)}{p(\mathbf{y})}}$$

Note that $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$ are posteriors of M_1 and M_2 respectively.

In comparing the two models, we want to look at the posterior odds in favor of one model (say M_1):

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{\frac{p(\mathbf{y}|M_1)p(M_1)}{p(\mathbf{y})}}{\frac{p(\mathbf{y}|M_2)p(M_2)}{p(\mathbf{y})}}$$

$$= \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \frac{p(M_1)}{p(M_2)}$$

Note that $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$ are posteriors of M_1 and M_2 respectively.

In comparing the two models, we want to look at the posterior odds in favor of one model (say M_1):

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{\frac{p(\mathbf{y}|M_1)p(M_1)}{p(\mathbf{y})}}{\frac{p(\mathbf{y}|M_2)p(M_2)}{p(\mathbf{y})}}$$

$$= \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \frac{p(M_1)}{p(M_2)}$$

The term $\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$ is known as the **Bayes factor**.

Note that $p(M_1|\mathbf{y})$ and $p(M_2|\mathbf{y})$ are posteriors of M_1 and M_2 respectively.

In comparing the two models, we want to look at the posterior odds in favor of one model (say M_1):

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{\frac{p(\mathbf{y}|M_1)p(M_1)}{p(\mathbf{y})}}{\frac{p(\mathbf{y}|M_2)p(M_2)}{p(\mathbf{y})}} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}\frac{p(M_1)}{p(M_2)}$$

The term $\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$ is known as the **Bayes factor**.

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

If both models are equally likely *a priori* ($p(M_1) = p(M_2)$), then the posterior odds equal the Bayes factor.

If both models are equally likely *a priori* ($p(M_1) = p(M_2)$), then the posterior odds equal the Bayes factor.

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$$

If both models are equally likely *a priori* ($p(M_1) = p(M_2)$), then the posterior odds equal the Bayes factor.

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$$

Thus, the Bayes factor is a measure of how much the data supports one model relative to the other.

If both models are equally likely *a priori* ($p(M_1) = p(M_2)$), then the posterior odds equal the Bayes factor.

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$$

Thus, the Bayes factor is a measure of how much the data supports one model relative to the other.

The individual terms that make up the Bayes factor, $p(\mathbf{y}|M_1)$ and $p(\mathbf{y}|M_2)$, are known as **marginal likelihoods**

Marginal Likelihood

Marginal Likelihood

The marginal likelihood for model M_k is

$$p(\mathbf{y}|M_k) = \int_{\boldsymbol{\Theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k$$

Marginal Likelihood

The marginal likelihood for model M_k is

$$p(\mathbf{y}|M_k) = \int_{\boldsymbol{\Theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k$$

where $\boldsymbol{\theta}_k$ are the model parameters for model M_k.

Marginal Likelihood

The marginal likelihood for model M_k is

$$p(\mathbf{y}|M_k) = \int_{\boldsymbol{\Theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, M_k) \textcolor{red}{p(\boldsymbol{\theta}_k|M_k)} d\boldsymbol{\theta}_k$$

where $\boldsymbol{\theta}_k$ are the model parameters for model M_k.

Thus, the marginal likelihood is "marginal" because it is the likelihood of $\mathbf{y}$ under M_k *averaged* over the model parameters $\boldsymbol{\theta}_k$.

Marginal Likelihood

The marginal likelihood for model M_k is

$$p(\mathbf{y}|M_k) = \int_{\boldsymbol{\Theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k$$

where $\boldsymbol{\theta}_k$ are the model parameters for model M_k.

Thus, the marginal likelihood is "marginal" because it is the likelihood of $\mathbf{y}$ under M_k *averaged* over the model parameters $\boldsymbol{\theta}_k$.

Note that $p(\boldsymbol{\theta}_k|M_k)$ is just the prior for $\boldsymbol{\theta}$ under M_k.

Marginal Likelihood

The marginal likelihood for model M_k is

$$p(\mathbf{y}|M_k) = \int_{\mathbf{\Theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k$$

where $\boldsymbol{\theta}_k$ are the model parameters for model M_k.

Thus, the marginal likelihood is "marginal" because it is the likelihood of $\mathbf{y}$ under M_k *averaged* over the model parameters $\boldsymbol{\theta}_k$.

Note that $p(\boldsymbol{\theta}_k|M_k)$ is just the prior for $\boldsymbol{\theta}$ under M_k.

The marginal likelihood can be interpreted as

Marginal Likelihood

The marginal likelihood for model M_k is

$$p(\mathbf{y}|M_k) = \int_{\boldsymbol{\Theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k$$

where $\boldsymbol{\theta}_k$ are the model parameters for model M_k.

Thus, the marginal likelihood is "marginal" because it is the likelihood of $\mathbf{y}$ under M_k *averaged* over the model parameters $\boldsymbol{\theta}_k$.

Note that $p(\boldsymbol{\theta}_k|M_k)$ is just the prior for $\boldsymbol{\theta}$ under M_k.

The marginal likelihood can be interpreted as

- the normalizing constant of the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ given M_k

Marginal Likelihood

The marginal likelihood for model M_k is

$$p(\mathbf{y}|M_k) = \int_{\boldsymbol{\Theta}_k} p(\mathbf{y}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k$$

where $\boldsymbol{\theta}_k$ are the model parameters for model M_k.

Thus, the marginal likelihood is "marginal" because it is the likelihood of $\mathbf{y}$ under M_k *averaged* over the model parameters $\boldsymbol{\theta}_k$.

Note that $p(\boldsymbol{\theta}_k|M_k)$ is just the prior for $\boldsymbol{\theta}$ under M_k.

The marginal likelihood can be interpreted as

- the normalizing constant of the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ given M_k
- the expected value of the likelihood function taken over the prior density

Note that the marginal likelihood is just

$$p(\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

conditioned on M_k.

Note that the marginal likelihood is just

$$p(\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

conditioned on M_k.

Also recall Bayes rule:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

Note that the marginal likelihood is just

$$p(\mathbf{y}) = \int_{\mathbf{\Theta}} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

conditioned on M_k.

Also recall Bayes rule:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

This means that for model M_k, the marginal likelihood is just the normalizing constant for the posterior of $\boldsymbol{\theta}$!

Note that the marginal likelihood is just

$$p(\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

conditioned on M_k.

Also recall Bayes rule:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

This means that for model M_k, the marginal likelihood is just the normalizing constant for the posterior of $\boldsymbol{\theta}$!

Therefore, except in simple cases (such as conjugacy), the marginal likelihood usually has to be approximated.

With the (approximate) marginal likelihoods, the Bayes factor can be calculated:

With the (approximate) marginal likelihoods, the Bayes factor can be calculated:

$$\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} = \frac{\int_{\mathbf{\Theta}_1} p(\mathbf{y}|\boldsymbol{\theta}_1, M_1)p(\boldsymbol{\theta}_1|M_1)d\boldsymbol{\theta}_1}{\int_{\mathbf{\Theta}_2} p(\mathbf{y}|\boldsymbol{\theta}_2, M_2)p(\boldsymbol{\theta}_2|M_2)d\boldsymbol{\theta}_2}$$

With the (approximate) marginal likelihoods, the Bayes factor can be calculated:

$$\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} = \frac{\int_{\mathbf{\Theta}_1} p(\mathbf{y}|\boldsymbol{\theta}_1, M_1)p(\boldsymbol{\theta}_1|M_1)d\boldsymbol{\theta}_1}{\int_{\mathbf{\Theta}_2} p(\mathbf{y}|\boldsymbol{\theta}_2, M_2)p(\boldsymbol{\theta}_2|M_2)d\boldsymbol{\theta}_2}$$

Interpretation of the Bayes factor is somewhat arbitrary.

With the (approximate) marginal likelihoods, the Bayes factor can be calculated:

$$\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} = \frac{\int_{\boldsymbol{\Theta}_1} p(\mathbf{y}|\boldsymbol{\theta}_1, M_1)p(\boldsymbol{\theta}_1|M_1)d\boldsymbol{\theta}_1}{\int_{\boldsymbol{\Theta}_2} p(\mathbf{y}|\boldsymbol{\theta}_2, M_2)p(\boldsymbol{\theta}_2|M_2)d\boldsymbol{\theta}_2}$$

Interpretation of the Bayes factor is somewhat arbitrary.

Generally speaking, for the Bayes factor of M_1 over M_2, a Bayes factor

- less than $\frac{1}{100}$ leads us to reject M_1

With the (approximate) marginal likelihoods, the Bayes factor can be calculated:

$$\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} = \frac{\int_{\boldsymbol{\Theta}_1} p(\mathbf{y}|\boldsymbol{\theta}_1, M_1) p(\boldsymbol{\theta}_1|M_1) d\boldsymbol{\theta}_1}{\int_{\boldsymbol{\Theta}_2} p(\mathbf{y}|\boldsymbol{\theta}_2, M_2) p(\boldsymbol{\theta}_2|M_2) d\boldsymbol{\theta}_2}$$

Interpretation of the Bayes factor is somewhat arbitrary.

Generally speaking, for the Bayes factor of M_1 over M_2, a Bayes factor

- less than $\frac{1}{100}$ leads us to reject M_1
- greater than 1 leads us to accept M_1

One can also rely on tables such as the one given by Jeffreys for the Bayes factor of M_1 over M_2:

One can also rely on tables such as the one given by Jeffreys for the Bayes factor of M_1 over M_2:

Bayes Factor	Strength of Evidence for M_1
< 1	Negative (supports M_2)
1 to 3	Barely Worth Mentioning
3 to 10	Substantial
10 to 30	Strong
30 to 100	Very Strong
> 100	Decisive