Hierarchical Models & Bayesian Model Selection

Geoffrey Roeder

Departments of Computer Science and Statistics
University of British Columbia

Jan. 20, 2016

Contact information

Please report any typos or errors to geoff.roeder@gmail.com

Outline

1. Hierarchical Bayesian Modelling
   - Coin toss redux: point estimates for θ
   - Hierarchical models
   - Application to clinical study

2. Bayesian Model Selection
   - Introduction
   - Bayes Factors
   - Shortcut for Marginal Likelihood in Conjugate Case

Outline

1. Hierarchical Bayesian Modelling
   - Coin toss redux: point estimates for θ
   - Hierarchical models
   - Application to clinical study

2. Bayesian Model Selection
   - Introduction
   - Bayes Factors
   - Shortcut for Marginal Likelihood in Conjugate Case

Coin toss: point estimates for θ

Probability model

- Consider the experiment of tossing a coin n times. Each toss results in heads with probability θ and tails with probability $1 - \theta$

Coin toss: point estimates for θ

Probability model

- Consider the experiment of tossing a coin n times. Each toss results in heads with probability θ and tails with probability $1-\theta$
- Let Y be a random variable denoting number of observed heads in n coin tosses. Then, we can model $Y \sim Bin(n, \theta)$, with probability mass function

$$p(Y = y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \tag{1}$$

Coin toss: point estimates for θ

Probability model

- Consider the experiment of tossing a coin n times. Each toss results in heads with probability θ and tails with probability $1 - \theta$
- Let Y be a random variable denoting number of observed heads in n coin tosses. Then, we can model $Y \sim Bin(n, \theta)$, with probability mass function
$$p(Y = y \mid \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y} \tag{1}$$
- We want to estimate the parameter θ

Coin toss: point estimates for θ

Maximum Likelihood

- By interpreting $p(Y = y|\theta)$ as a function of θ rather than y, we get the likelihood function for θ

Coin toss: point estimates for θ

Maximum Likelihood

- By interpreting $p(Y = y|\theta)$ as a function of θ rather than y, we get the likelihood function for θ
- Let $\ell\ell(\theta|y) := log\ p(y|\theta)$, the log-likelihood. Then,

$$\hat{\theta}_{ML} = \underset{\theta}{\text{argmax}} \ell\ell(\theta|y) = \underset{\theta}{\text{argmax}}\ ylog(\theta) + (n - y)log(1 - \theta) \quad (2)$$

Coin toss: point estimates for θ

Maximum Likelihood

- By interpreting $p(Y = y|\theta)$ as a function of θ rather than y, we get the likelihood function for θ
- Let $\ell\ell(\theta|y) := log\ p(y|\theta)$, the log-likelihood. Then,

$$\hat{\theta}_{ML} = \underset{\theta}{\text{argmax}} \ell\ell(\theta|y) = \underset{\theta}{\text{argmax}}\ ylog(\theta) + (n-y)log(1-\theta) \tag{2}$$

- Since the log likelihood is a concave function of θ,

$$\begin{aligned} \underset{\theta}{\text{argmax}}\ \ell(\theta|y) &\Leftrightarrow 0 = \left.\frac{\partial \ell(\theta|y)}{\partial \theta}\right|_{\hat{\theta}_{ML}} \\ &\Leftrightarrow 0 = \frac{y}{\hat{\theta}_{ML}} - \frac{n-y}{1-\hat{\theta}_{ML}} \\ &\Leftrightarrow \hat{\theta}_{ML} = \frac{y}{n} \end{aligned} \tag{3}$$

- What if sample size is small?

- What if sample size is small?
- Asymptotic result that this approaches true parameter

Coin toss: point estimates for θ

Maximum A Posteriori

Alternative analysis: reverse the conditioning with Bayes' Theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4}$$

- Lets us encode our prior beliefs or knowledge about θ in a prior distribution for the parameter, $p(\theta)$

Coin toss: point estimates for θ

Maximum A Posteriori

Alternative analysis: reverse the conditioning with Bayes' Theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4}$$

- Lets us encode our prior beliefs or knowledge about θ in a prior distribution for the parameter, $p(\theta)$
- Recall that if $p(y|\theta)$ is in the exponential family, there exists a conjugate prior $p(\theta)$ s.t. if $p(\theta) \in \mathcal{F}$, then $p(y|\theta)p(\theta) \in \mathcal{F}$

Coin toss: point estimates for θ

Maximum A Posteriori

Alternative analysis: reverse the conditioning with Bayes' Theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{4}$$

- Lets us encode our prior beliefs or knowledge about θ in a prior distribution for the parameter, $p(\theta)$
- Recall that if $p(y|\theta)$ is in the exponential family, there exists a conjugate prior $p(\theta)$ s.t. if $p(\theta) \in \mathcal{F}$, then $p(y|\theta)p(\theta) \in \mathcal{F}$
- Saw last time that binomial is in the exponential family, and $\theta \sim Beta(\alpha, \beta)$ is a conjugate prior.

$$p(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{5}$$

- Moreover, for any given realization y of Y, the marginal distribution $p(y) = \int p(y|\theta')p(\theta')d\theta'$ is a constant

Coin toss: point estimates for θ

Maximum A Posteriori

- Moreover, for any given realization y of Y, the marginal distribution $p(y) = \int p(y|\theta')p(\theta')d\theta'$ is a constant

- Thus, $p(\theta|y) \propto p(y|\theta)p(\theta)p(y)$ so that

$$
\begin{aligned}
\hat{\theta}_{MAP} &= \underset{\theta}{\text{argmax}}\ p(\theta|y) \\
&= \underset{\theta}{\text{argmax}}\ p(y|\theta)p(\theta) \\
&= \underset{\theta}{\text{argmax}}\ log\ \ p(y|\theta)p(\theta)
\end{aligned}
$$

- Moreover, for any given realization y of Y, the marginal distribution $p(y) = \int p(y|\theta')p(\theta')d\theta'$ is a constant

- Thus, $p(\theta|y) \propto p(y|\theta)p(\theta)p(y)$ so that

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}}\ p(\theta|y) \\ &= \underset{\theta}{\operatorname{argmax}}\ p(y|\theta)p(\theta) \\ &= \underset{\theta}{\operatorname{argmax}}\ log\ \ p(y|\theta)p(\theta)\end{aligned}$$

- By evaluating the first partial derivative w.r.t θ and setting to 0 at $\hat{\theta}_{MAP}$ we can derive

$$\hat{\theta}_{MAP} = \frac{y + \alpha - 1}{n + \beta - 1 + \alpha - 1} \tag{6}$$

- The point estimate for $\hat{\theta}_{MAP}$ shows choices of α and β correspond to having already seen prior data. Can encode strength of prior belief using these parameters.

Coin toss: point estimates for θ

Point estimate for θ: Choosing α, β

- The point estimate for $\hat{\theta}_{MAP}$ shows choices of α and β correspond to having already seen prior data. Can encode strength of prior belief using these parameters.
- Can also choose uninformative prior: Jeffreys' prior. For beta-binomial model, corresponds to $(\alpha, \beta) = (\frac{1}{2}, \frac{1}{2})$.

- The point estimate for $\hat{\theta}_{MAP}$ shows choices of α and β correspond to having already seen prior data. Can encode strength of prior belief using these parameters.
- Can also choose uninformative prior: Jeffreys' prior. For beta-binomial model, corresponds to $(\alpha, \beta) = (\frac{1}{2}, \frac{1}{2})$.
- Deriving an analytic form for the posterior is possible also if the prior is conjugate. We saw last week that for a single Binomial experiment with a conjugate Beta, $p(\theta|y) \sim Beta(\alpha + y - 1, \beta + n - y - 1)$

Outline

1. Hierarchical Bayesian Modelling
   - Coin toss redux: point estimates for θ
   - Hierarchical models
   - Application to clinical study
2. Bayesian Model Selection
   - Introduction
   - Bayes Factors
   - Shortcut for Marginal Likelihood in Conjugate Case

- Putting a prior on the parameter θ was pretty useful

- Putting a prior on the parameter θ was pretty useful
- We ended up with two parameters α and β we could choose to formally encode our knowledge about the random process

- Putting a prior on the parameter θ was pretty useful
- We ended up with two parameters α and β we could choose to formally encode our knowledge about the random process
- Often, though, we want to go one step further: put a prior on the prior, rather than treating α and β as constants

- Putting a prior on the parameter θ was pretty useful
- We ended up with two parameters α and β we could choose to formally encode our knowledge about the random process
- Often, though, we want to go one step further: put a prior on the prior, rather than treating α and β as constants
- Then, θ is a sample from a population distribution

- Example: now we have information available at different "levels" of the observational units

- Example: now we have information available at different "levels" of the observational units
- At each level the observational units must be **exchangeable**

- Example: now we have information available at different "levels" of the observational units
- At each level the observational units must be **exchangeable**
- Informally, a joint probability distribution $p(y_1, ..., y_n)$ is exchangeable if the indices on the y_i can be shuffled without changing the distribution

- Example: now we have information available at different "levels" of the observational units
- At each level the observational units must be **exchangeable**
- Informally, a joint probability distribution $p(y_1, ..., y_n)$ is exchangeable if the indices on the y_i can be shuffled without changing the distribution
- Then, a *Hierarchical Bayesian model* introduces an additional prior distribution **for each level of observational unit**, allowing additional unobserved parameters to explain some dependencies in the model

Example

A clinical trial of a new cancer drug has been designed to compare the five-year survival probability in a population given the new drug to the five-year survival probability in a population under a standard treatment (Gelman et al. [2014]).

- Suppose the two drugs are administered in separate randomized experiments to patients in different cities.

Example

A clinical trial of a new cancer drug has been designed to compare the five-year survival probability in a population given the new drug to the five-year survival probability in a population under a standard treatment (Gelman et al. [2014]).

- Suppose the two drugs are administered in separate randomized experiments to patients in different cities.
- Within each city, the patients can be considered exchangeable

Example

A clinical trial of a new cancer drug has been designed to compare the five-year survival probability in a population given the new drug to the five-year survival probability in a population under a standard treatment (Gelman et al. [2014]).

- Suppose the two drugs are administered in separate randomized experiments to patients in different cities.
- Within each city, the patients can be considered exchangeable
- The results from different hospitals can also be considered exchangeable

Terminology note:

- With hierarchical Bayes, we have one set of parameters θ_i to model the **survival probability** of the patients y_{ij} in hospital i, and another set of parameters ϕ to model **the random process governing the generation of** θ_j

Terminology note:

- With hierarchical Bayes, we have one set of parameters θ_i to model the **survival probability** of the patients y_{ij} in hospital i, and another set of parameters ϕ to model **the random process governing the generation of θ_j**
- Hence, θ_i are themselves given a probabilistic specification in terms of **hyperparameters** ϕ through a **hyperprior** $p(\phi)$

Outline

1. Hierarchical Bayesian Modelling
   - Coin toss redux: point estimates for θ
   - Hierarchical models
   - Application to clinical study
2. Bayesian Model Selection
   - Introduction
   - Bayes Factors
   - Shortcut for Marginal Likelihood in Conjugate Case

Motivating example: Incidence of tumors in rodents

Adapted from Gelman et al. (2014)

Let's develop a Hierarchical model using the beta-binomial Bayesian approach seen so far

Example

- Suppose we have the results of a clinical study of a drug in which rodents were exposed to either a dose of the drug or a control treatment (no dose)

Let's develop a Hierarchical model using the beta-binomial Bayesian approach seen so far

Example

- Suppose we have the results of a clinical study of a drug in which rodents were exposed to either a dose of the drug or a control treatment (no dose)
- 4 out of 14 rodents in the control group developed tumors

Motivating example: Incidence of tumors in rodents

Adapted from Gelman et al. (2014)

Let's develop a Hierarchical model using the beta-binomial Bayesian approach seen so far

Example

- Suppose we have the results of a clinical study of a drug in which rodents were exposed to either a dose of the drug or a control treatment (no dose)
- 4 out of 14 rodents in the control group developed tumors
- We want to estimate θ, the probability that the rodents in the control group developed a tumor given no dose of the drug

Motivating example: Incidence of tumors in rodents

Data

We also have the following data about the incidence of this kind of tumor in the control groups of other studies:

Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:
4/14

Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of* y_j/n_j: *(number of rats with tumors)/(total number of rats).*

Figure: Gelman et al. 2014 p.102

- Including the current experimental results, we have information on 71 random variables $\theta_1, ..., \theta_{71}$

- Including the current experimental results, we have information on 71 random variables $\theta_1, ..., \theta_{71}$
- We can model the current and historical proportions as a random sample from some unknown population distribution: each y_j is independent binomial data, given the sample sizes n_j and experiment-specific θ_j.

- Including the current experimental results, we have information on 71 random variables $\theta_1, ..., \theta_{71}$
- We can model the current and historical proportions as a random sample from some unknown population distribution: each y_j is independent binomial data, given the sample sizes n_j and experiment-specific θ_j.
- Each θ_j is in turn generated by a random process governed by a population distribution that depends on the parameters α and β

Motivating example: Incidence of tumors in rodents

Bayesian analysis: model

This relationship can be depicted as graphically as

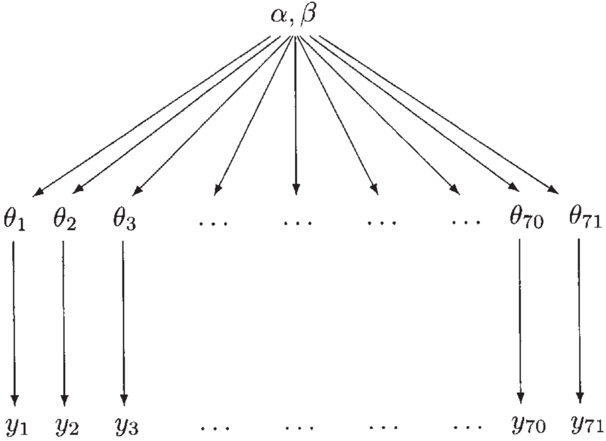

Figure: Hierarchical model (Gelman et al. 2014 p.103)

- Formally, posterior distribution is now of the vector (θ, α, β). The joint prior distribution is

$$p(\theta, \alpha, \beta) = p(\alpha, \beta)p(\theta|\alpha, \beta) \tag{7}$$

and the joint posterior distribution is

$$\begin{aligned} p(\theta, \alpha, \beta|y) &\propto p(\theta, \alpha, \beta)p(y|\theta, \alpha, \beta) \\ &= p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta) \\ &= p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta) \end{aligned} \tag{8}$$

- Since the beta prior is conjugate, we can derive the joint posterior distribution analytically

- Since the beta prior is conjugate, we can derive the joint posterior distribution analytically
- Each y_j is conditionally independent of the hyperparameters α, β given θ_j. Hence, the likelihood function is still

$$\begin{aligned} p(y|\theta,\alpha,\beta) = p(y|\theta) &= p(y_1, y_2, ..., y_J|\theta_1, \theta_2, ..., \theta_J) \\ &= \prod_{j=1}^{J} p(y_j|\theta_j) = \prod_{j=1}^{J} \binom{n_j}{y_j} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j} \end{aligned} \tag{9}$$

Motivating example: Incidence of tumors in rodents

Bayesian analysis: joint posterior density

- Since the beta prior is conjugate, we can derive the joint posterior distribution analytically
- Each y_j is conditionally independent of the hyperparameters α, β given θ_j. Hence, the likelihood function is still

$$\begin{aligned} p(y|\theta,\alpha,\beta) &= p(y|\theta) = p(y_1, y_2, ..., y_J|\theta_1, \theta_2, ..., \theta_J) \\ &= \prod_{j=1}^{J} p(y_j|\theta_j) = \prod_{j=1}^{J} \binom{n_j}{y_j} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j} \end{aligned} \tag{9}$$

- Now we also have a population distribution $p(\theta|\alpha,\beta)$:

$$\begin{aligned} p(\theta|\alpha,\beta) &= p(\theta_1, \theta_2, ..., \theta_J|\alpha,\beta) \\ &= \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \end{aligned} \tag{10}$$

- Then, using equations (8) and (9), the unnormalized joint posterior distribution $p(\theta, \alpha, \beta | y)$ is

$$p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}. \tag{11}$$

- Then, using equations (8) and (9), the unnormalized joint posterior distribution $p(\theta, \alpha, \beta | y)$ is

$$p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} (1 - \theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j} (1 - \theta_j)^{n_j - y_j}. \tag{11}$$

- We can also determine analytically the conditional posterior density of $\theta = (\theta_1, \theta_2, ..., \theta_J)$:

$$p(\theta | \alpha, \beta, y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha + \beta + n_j)}{\Gamma(\alpha + y_j)\Gamma(\beta + n_j - y_j)} \theta_j^{\alpha + y_j - 1} (1 - \theta_j)^{\beta + n_j - y_j - 1}. \tag{12}$$

- Then, using equations (8) and (9), the unnormalized joint posterior distribution $p(\theta, \alpha, \beta|y)$ is

$$p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\prod_{j=1}^{J}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j}. \tag{11}$$

- We can also determine analytically the conditional posterior density of $\theta = (\theta_1, \theta_2, ..., \theta_J)$:

$$p(\theta|\alpha,\beta,y)=\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}\theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}. \tag{12}$$

- Note that equation (11), the conditional posterior, is now **a function of** (α, β). Each θ_j depends on the hyperparameters of the hyperprior $p(\alpha, \beta)$.

To compute the marginal posterior density, observe that if we condition on y, equation (7) is equivalent to

$$p(\alpha, \beta | y) = \frac{p(\theta, \alpha, \beta | y)}{p(\theta | \alpha, \beta, y)} \tag{13}$$

which are equations (10) and (1) on the previous slide. Hence,

$$\begin{aligned} p(\alpha, \beta | y) &= p(\alpha, \beta) \frac{\prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} (1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}}{\prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1} (1-\theta_j)^{\beta+n_j-y_j-1}} \\ &= p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}, \end{aligned} \tag{14}$$

which is computationally tractable, given a prior for (α, β).

Summary: beta-binomial hierarchical model

- We started by wanting to understand the true proportion of rodents in the control group of a clinical study that developed a tumor.

Summary: beta-binomial hierarchical model

- We started by wanting to understand the true proportion of rodents in the control group of a clinical study that developed a tumor.
- By modelling the relationship between different trials hierarchically, we were able to bring our uncertainty about the hyperparameters (α, β) into the model

Summary: beta-binomial hierarchical model

- We started by wanting to understand the true proportion of rodents in the control group of a clinical study that developed a tumor.
- By modelling the relationship between different trials hierarchically, we were able to bring our uncertainty about the hyperparameters (α, β) into the model
- Using analytical methods, we developed a model that, given a suitable population prior and the method of simulating draws from the distribution in order to estimate (α, β).

- In general, if θ_j is the population parameter for an observable x, and ϕ be a hyperprior distribution

$$p(\theta, \phi|x) = \frac{p(x|\theta, \phi)p(\theta, \phi)}{p(x)} = \frac{p(x|\theta)p(\theta|\phi)p(\phi)}{p(x)} \tag{15}$$

- In general, if θ_j is the population parameter for an observable x, and ϕ be a hyperprior distribution

$$p(\theta, \phi|x) = \frac{p(x|\theta, \phi)p(\theta, \phi)}{p(x)} = \frac{p(x|\theta)p(\theta|\phi)p(\phi)}{p(x)} \tag{15}$$

- The models can be extended with more levels by adding hyperpriors and hyperparameter vectors, leading to the factored form:

$$p(\theta, \phi, \psi|x) = \frac{p(x|\theta)p(\theta|\phi)p(\phi|\psi)p(\psi)}{p(x)} \tag{16}$$

Outline

The model selection problem:

Given a set of models (i.e., families of parametric distributions) of different complexity, how should we choose the best one?

- Bayesian approach: compare the posterior over models $H_k \in \mathcal{H}$

$$p(H_k|\mathcal{D}) = \frac{p(\mathcal{D}|H_k)p(H_k)}{\sum_{H' \in \mathcal{H}} p(H', \mathcal{D})} \tag{17}$$

- Bayesian approach: compare the posterior over models $H_k \in \mathcal{H}$

$$p(H_k|\mathcal{D}) = \frac{p(\mathcal{D}|H_k)p(H_k)}{\sum_{H' \in \mathcal{H}} p(H', \mathcal{D})} \tag{17}$$

- then, select MAP model as best

$$\hat{H}_{MAP} = \underset{H' \in \mathcal{H}}{\operatorname{argmax}} \; p(H'|\mathcal{D}). \tag{18}$$

- Bayesian approach: compare the posterior over models $H_k \in \mathcal{H}$

$$p(H_k|\mathcal{D}) = \frac{p(\mathcal{D}|H_k)p(H_k)}{\sum_{H' \in \mathcal{H}} p(H', \mathcal{D})} \tag{17}$$

- then, select MAP model as best

$$\hat{H}_{MAP} = \underset{H' \in \mathcal{H}}{\operatorname{argmax}}\ p(H'|\mathcal{D}). \tag{18}$$

- If we adopt a uniform prior to represent our uncertainty about the choice of models s.t. $p(H_k) \sim \mathcal{U}(0,1) \Rightarrow p(H_k) \propto 1$, then

$$\hat{H}_{MAP} = \underset{H' \in \mathcal{H}}{\operatorname{argmax}}\ p(H'|\mathcal{D}) \Leftrightarrow \underset{H' \in \mathcal{H}}{\operatorname{argmax}}\ p(\mathcal{D}|H') \tag{19}$$

Bayesian Model Selection

Bayesian solution (Adapted from Murphy 2012)

- Bayesian approach: compare the posterior over models $H_k \in \mathcal{H}$

$$p(H_k|\mathcal{D}) = \frac{p(\mathcal{D}|H_k)p(H_k)}{\sum_{H' \in \mathcal{H}} p(H', \mathcal{D})} \tag{17}$$

- then, select MAP model as best

$$\hat{H}_{MAP} = \underset{H' \in \mathcal{H}}{\operatorname{argmax}}\ p(H'|\mathcal{D}). \tag{18}$$

- If we adopt a uniform prior to represent our uncertainty about the choice of models s.t. $p(H_k) \sim \mathcal{U}(0, 1) \Rightarrow p(H_k) \propto 1$, then

$$\hat{H}_{MAP} = \underset{H' \in \mathcal{H}}{\operatorname{argmax}}\ p(H'|\mathcal{D}) \Leftrightarrow \underset{H' \in \mathcal{H}}{\operatorname{argmax}}\ p(\mathcal{D}|H') \tag{19}$$

- and so the problem reduces to choosing the model which maximizes the marginal likelihood (also called the "evidence"):

$$p(\mathcal{D}|H_k) = \int p(\mathcal{D}|\theta_k, H_k)p(\theta_k|H_k)d\theta_k \tag{20}$$

Outline

Bayes Factors

Adapted from Kass and Raftery (1995)

- Bayes Factors are a natural way to compare models using marginal likelihoods

Bayes Factors

Adapted from Kass and Raftery (1995)

- Bayes Factors are a natural way to compare models using marginal likelihoods
- In simplest case, we have two hypotheses $\mathcal{H} = \{H_1, H_2\}$ about the random process which generated $\mathcal{D}$ according to distributions $p(\mathcal{D}|H_1),\ p(\mathcal{D}|H_2)$

- Bayes Factors are a natural way to compare models using marginal likelihoods
- In simplest case, we have two hypotheses $\mathcal{H} = \{H_1, H_2\}$ about the random process which generated $\mathcal{D}$ according to distributions $p(\mathcal{D}|H_1),\ p(\mathcal{D}|H_2)$
- Recall the odds representation of probability: it gives a structure we can use in model selection

$$odds = \frac{proportion\ of\ successes}{proportion\ of\ failures} = \frac{probability}{1 - probability} \tag{21}$$

- Bayes' theorem says

$$p(H_k|\mathcal{D}) = \frac{p(\mathcal{D}|H_k)p(H_k)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D})|H_{h'})p(H_{h'})} \quad (22)$$

- Bayes' theorem says

$$p(H_k|\mathcal{D}) = \frac{p(\mathcal{D}|H_k)p(H_k)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D})|H_{h'})p(H_{h'})} \tag{22}$$

- Since $p(H_1|\mathcal{D}) = 1 - p(H_2|\mathcal{D})$ (in the 2-hypothesis case),

$$\text{odds}(H_1|\mathcal{D}) = \frac{p(H_1|\mathcal{D})}{p(H_2|\mathcal{D})} = \frac{p(\mathcal{D}|H_1)}{p(\mathcal{D}|H_2)} \frac{p(H_1)}{p(H_2)} \tag{23}$$

$$\textit{posterior odds} = \text{Bayes factor } x \textit{ prior odds}$$

Bayes Factors

- Prior odds are transformed into the posterior odds by the ratio of marginal likelihoods. The Bayes factor for model H_1 against H_2 is

$$B_{12} = \frac{p(\mathcal{D}|H_1)}{p(\mathcal{D}|H_2)} \tag{24}$$

Bayes Factors

- Prior odds are transformed into the posterior odds by the ratio of marginal likelihoods. The Bayes factor for model H_1 against H_2 is

$$B_{12} = \frac{p(\mathcal{D}|H_1)}{p(\mathcal{D}|H_2)} \tag{24}$$

- Bayes factor is a summary of the evidence provided by the data in favour of one hypothesis over another

Bayes Factors

- Prior odds are transformed into the posterior odds by the ratio of marginal likelihoods. The Bayes factor for model H_1 against H_2 is

$$B_{12} = \frac{p(\mathcal{D}|H_1)}{p(\mathcal{D}|H_2)} \tag{24}$$

- Bayes factor is a summary of the evidence provided by the data in favour of one hypothesis over another
- Can interpret Bayes factors Jeffreys' scale of evidence:

B_{jk} :	Evidence against H_k :
1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
100 or above	Decisive

Bayes Factors

Coin toss example (Adapted from Arnaud)

- Suppose you toss a coin 6 times and observe 6 heads.

Bayes Factors

Coin toss example (Adapted from Arnaud)

- Suppose you toss a coin 6 times and observe 6 heads.
- If θ is the probability of getting heads, can test $H_1 : \theta = \frac{1}{2}$ against $H_2 : \theta \sim Unif(\frac{1}{2}, 1]$

- Suppose you toss a coin 6 times and observe 6 heads.
- If θ is the probability of getting heads, can test $H_1 : \theta = \frac{1}{2}$ against $H_2 : \theta \sim Unif(\frac{1}{2}, 1]$
- Then, the Bayes factor for fair against biased is

$$B_{12} = \frac{p(\mathcal{D}|H_1)}{p(\mathcal{D}|H_2)} = \frac{\int p(\mathcal{D}|\theta_1, H_1)p(\theta_1|H_1)d\theta_1}{\int p(\mathcal{D}|\theta_2, H_2)p(\theta_2|H_2)d\theta_2}$$

$$= \frac{\frac{1}{2}\int_{\frac{1}{2}}^{1} \theta^x(1-\theta)^{6-x}d\theta}{(\frac{1}{2})^x(1-\frac{1}{2})^{6-x}}$$

$$= \frac{\frac{1}{2}\int_{\frac{1}{2}}^{1} \theta^6 d\theta}{(\frac{1}{2})^6} \approx 4.535.$$

- Suppose we have a random variable $X|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known but μ is unknown.

- Suppose we have a random variable $X|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known but μ is unknown.
- Our two hypotheses are $H_1 : \mu = 0$ vs $H_2 : \mu \sim \mathcal{N}(\xi, \tau^2)$

- Suppose we have a random variable $X|\mu,\sigma^2 \sim \mathcal{N}(\mu,\sigma^2)$ where σ^2 is known but μ is unknown.
- Our two hypotheses are $H_1 : \mu = 0$ vs $H_2 : \mu \sim \mathcal{N}(\xi,\tau^2)$
- Then, the Bayes factor for H_1 against H_2 is

$$\begin{aligned} B_{12} = \frac{p(\mathcal{D}|H_1)}{p(\mathcal{D}|H_2)} &= \frac{\int \mathcal{N}(x|\mu,\sigma^2)\mathcal{N}(\mu|\xi,\tau^2)d\mu}{\int \mathcal{N}(x|\mu,\sigma^2)\delta_0(\mu)d\mu} \\ &= \frac{\int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}\frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{\frac{(\mu-\xi)^2}{2\tau^2}\right\}d\mu}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{x^2}{2\sigma^2}\right\}} \\ &= \frac{\sigma^2}{\sqrt{\sigma^2+\tau^2}}\exp\left\{\frac{\tau^2x^2}{2\sigma^2(\sigma^2+\tau^2)}\right\}. \end{aligned} \tag{25}$$

Bayes Factors

Key points

- Bayes factors allow you to compare models with different parameter spaces: the parameters are marginalized out in the integral

- Bayes factors allow you to compare models with different parameter spaces: the parameters are marginalized out in the integral
- Thus unlike MLE model comparison methods, Bayes factors do not favour more complex models. "Built-in" protection against overfitting

Bayes Factors

Key points

- Bayes factors allow you to compare models with different parameter spaces: the parameters are marginalized out in the integral
- Thus unlike MLE model comparison methods, Bayes factors do not favour more complex models. "Built-in" protection against overfitting
    - Recall AIC is -2 (log (likelihood)) + 2 K, where K is number of parameters in model

- Bayes factors allow you to compare models with different parameter spaces: the parameters are marginalized out in the integral
- Thus unlike MLE model comparison methods, Bayes factors do not favour more complex models. "Built-in" protection against overfitting
  - Recall AIC is -2 (log (likelihood)) + 2 K, where K is number of parameters in model
  - Since based on ML estimate of parameters, which are prone to overfit, AIC is biased towards more complex models and must be adjusted by the parameter K

- Bayes factors allow you to compare models with different parameter spaces: the parameters are marginalized out in the integral
- Thus unlike MLE model comparison methods, Bayes factors do not favour more complex models. "Built-in" protection against overfitting
  - Recall AIC is -2 (log (likelihood)) + 2 K, where K is number of parameters in model
  - Since based on ML estimate of parameters, which are prone to overfit, AIC is biased towards more complex models and must be adjusted by the parameter K
- Bayes factors are sensitive to the prior. In Gaussian examples, as $\tau \to \infty$, $B_{12} \to 0$ regardless of the data x. If prior is vague on a hypothesis, Bayes factor selection will not favour that hypothesis.

Outline

Computing Marginal Likelihood
(Adapted from Murphy 2013)

Suppose we write the prior as

$$p(\theta) = \frac{q(\theta)}{Z_0} \left(= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \right), \tag{26}$$

the likelihood as

$$p(\mathcal{D}|\theta) = \frac{q(\mathcal{D}|\theta)}{Z_\ell} \left(= \frac{\theta^{y}(1-\theta)^{n-y}}{\binom{n}{y}^{-1}} \right), \tag{27}$$

and the posterior as

$$p(\theta|\mathcal{D}) = \frac{q(\theta|\mathcal{D})}{Z_N} \left(= \frac{\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}}{B(\alpha+y,\beta+n-y)} \right). \tag{28}$$

Computing Marginal Likelihood
(Adapted from Murphy 2013)

Then:

$$\begin{aligned} p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\ \Leftrightarrow \frac{q(\theta|\mathcal{D})}{Z_N} &= \frac{q(\mathcal{D}|\theta)q(\theta)}{Z_\ell Z_0 p(\mathcal{D})} \\ \Leftrightarrow p(\mathcal{D}) &= \frac{Z_N}{Z_0 Z_\ell} \left(= \binom{n}{y} \frac{B(\alpha + y, \beta + n - y)}{B(\alpha, \beta)} \right) \end{aligned} \tag{29}$$

The computation reduces to a ratio of normalizing constants in this special case.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin
Bayesian Data Analysis
Chapman Hall/CRC, 2014.

Kevin Murphy
Machine Learning: A Probabilistic Perspective
MIT Press, 2013.

Arnaud Doucet
STAT 535C: Statistical Computing
Course lecture slides (2009)
Accessed 14 January 2016 from
http://www.cs.ubc.ca/~arnaud/stat535.html

References II

Robert E. Kass; Adrian E. Raftery
Bayes Factors
Journal of the American Statistical Association, Vol. 90, No. 430
773-795, 1995.