CS 340 Lec. 14: Bayesian Statistics

AD

February 2011

A Bayesian Approach

- In a Bayesian approach, the unknown parameter θ is assumed random with an associated prior distribution $p(\theta)$.
- Given data $\{\mathbf{x}^i\}_{i=1}^N$ distributed according to $p(\mathbf{x}_{1:N}|\theta)$, inference about θ is based on the posterior distribution

$$p(\theta|\mathbf{x}_{1:N}) = \frac{p(\mathbf{x}_{1:N}|\theta)\,p(\theta)}{p(\mathbf{x}_{1:N})}.$$

- From this posterior, we can obtain various point estimates of θ.

Bernoulli and Binomial Models

- Assume independent $\{x^i\}$ where $x^i \in \{0,1\}$ $(= \{Tail, Head\})$ with
$$p(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$$
so
$$p(x_{1:N}|\theta) = \theta^{n_1}(1-\theta)^{N-n_1}$$
where $n_1 = \sum_{i=1}^{n} \mathbb{I}(y^i = 1)$ and $\widehat{\theta}_{MLE} = n_1 / N$.
- n_1 is the number of "success" among N trials, it follows a Binomial distribution
$$p(n_1|\theta) = Bin(n_1; \theta, N) = \binom{N}{n_1} \theta^{n_1}(1-\theta)^{N-n_1}$$
- In a Bayesian framework, we set a prior density $p(\theta)$ on $\theta \in [0,1]$.
- If you know nothing about θ a reasonable prior is the uniform density
$$p(\theta) = 1_{[0,1]}(\theta).$$

Conjugate Priors

- For simplicity, we will mostly focus on a special kind of prior which has nice mathematical properties.
- A prior $p(\theta)$ is said to be conjugate to a likelihood $p(x_{1:N}|\theta)$ (equivalently $p(n_1|\theta)$) if the corresponding posterior $p(\theta|x_{1:N}) = p(\theta|n_1)$ has the same functional form as $p(\theta)$.
- This means the prior family is closed under Bayesian updating.
- So we can recursively apply the rule to update our beliefs as data streams in (online learning).

Beta Prior

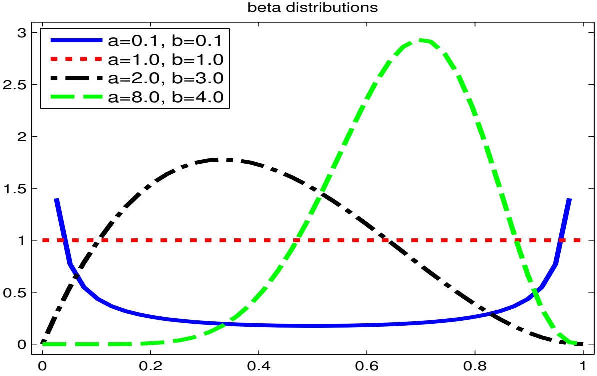
- Let us introduce the class of Beta densities defined for $\alpha, \beta > 0$

$$Beta\left(\theta;\alpha,\beta\right)=\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}1_{[0,1]}\left(\theta\right)$$

  where $\Gamma\left(u\right)=\int_0^\infty t^{u-1}e^{-t}dt$. Note that $\Gamma\left(u\right)=\left(u-1\right)!$ for $u\in\mathbb{N}$.
- Be careful: (α,β) are *fixed* quantities. To distinguish them from θ, we call them *hyperparameters*. For $\alpha=\beta=1$, the Beta density corresponds to the uniform density.
- The Beta prior is such that

$$\mathbb{E}\left(\theta\right)=\frac{\alpha}{\alpha+\beta},\ \mathbb{V}\left(\theta\right)=\frac{\alpha\beta}{\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)}.$$

Beta Prior


beta distributions
a=0.1, b=0.1
a=1.0, b=1.0
a=2.0, b=3.0
a=8.0, b=4.0
0
0.5
1
1.5
2
2.5
3
0
0.2
0.4
0.6
0.8
1

Bayesian Inference with the Binomial-Beta Model

- We obtain

$$\begin{aligned} p(\theta | n_1) &= \frac{p(n_1 | \theta)\, p(\theta)}{p(n_1)} \\ &\propto p(n_1 | \theta)\, p(\theta) \\ &\propto \theta^{n_1} (1-\theta)^{N-n_1} \theta^{\alpha-1} (1-\theta)^{\beta-1} 1_{[0,1]}(\theta) \\ &= \theta^{n_1+\alpha-1} (1-\theta)^{N-n_1+\beta-1} 1_{[0,1]}(\theta) \end{aligned}$$

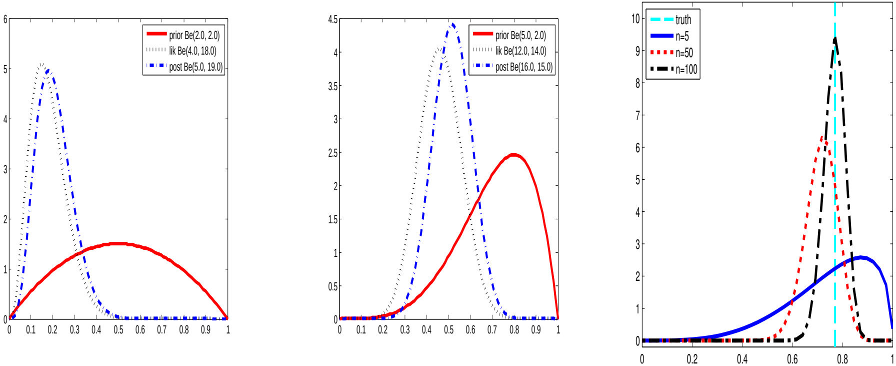
- This implies necessarily that
$p(\theta | x_{1:N}) = Beta(\theta; n_1 + \alpha, N - n_1 + \beta)$.
- The prior on θ can be conveniently reinterpreted as an imaginary initial sample of size $(\alpha + \beta - 2)$ with $\alpha - 1$ observations "1" and $\beta - 1$ observations "0". Provided that $(\alpha + \beta - 2)$ is small with respect to n, the information carried by the data is prominent.

Sequential Bayesian Inference with the Binomial-Beta Model

- Assume we first observe at 'time t' n_1^t '1' among N^t trials where $t = 1, 2, \ldots$
- We have
$$p\left(\theta \middle| n_1^1\right) = Beta\left(\theta; n_1^1 + \alpha, N^1 - n_1^1 + \beta\right)$$
- At time $t > 1$, we use
$$\begin{aligned} p\left(\theta \middle| n_1^1, \ldots, n_1^k\right) &\propto p\left(n_1^k \middle| \theta\right) p\left(\theta \middle| n_1^1, \ldots, n_1^{k-1}\right) \\ &= Beta\left(\theta; \alpha + \sum_{i=1}^{k} n_1^i, \beta + \sum_{i=1}^{k} \left(N^i - n_1^i\right)\right); \end{aligned}$$

i.e. the posterior at time k can be computed using as a prior the posterior at time $k-1$ and the likelihood of the observations at time k.

Bayesian Inference with the Binomial-Beta Model



(left) Updating a Beta(2,2) prior with a Binomial likelihood with $n_1 = 3$, $n_0 = 17$ to yield a Beta(5,19); (center) Updating a Beta(5,2) prior with a Binomial likelihood with $n_1 = 11$, $n_0 = 13$ to yield a Beta(16,15) posterior. (c) Sequentially updating a Beta distribution starting with a Beta(1,1) and converge to a delta function centered on the true value.

Bayesian Inference with the Binomial-Beta Model

- We have

$$\mathbb{E}\left(\theta \middle| n_1\right) = \frac{n_1 + \alpha}{n_1 + \alpha + n_0 + \beta} = \frac{n_1 + \alpha}{N + \alpha + \beta}$$

- The posterior means behave asymptotically like n_1 / n (the 'frequentist' estimator) and converge to θ^*, the 'true' value of θ^*.
- We have

$$\begin{aligned} \mathbb{V}\left(\theta \middle| n_1\right) &= \frac{\left(n_1 + \alpha\right)\left(n_0 + \beta\right)}{\left(n_1 + \alpha + n_0 + \beta\right)^2\left(n_1 + \alpha + n_0 + \beta + 1\right)} \\ &\approx \frac{\widehat{\theta}_{MLE}\left(1 - \widehat{\theta}_{MLE}\right)}{N} \text{ for large } N \end{aligned}$$

- The posterior variance decreases to zero as $n \rightarrow \infty$, at rate n^{-1}: the information you get on θ gets more and more precise.
- For n large enough, the prior is washed out by the data. For a small n, the prior can have a huge impact.

Bayesian Inference with the Bernoulli-Beta Model

- We can compute things like

$$\Pr\left(\theta \in [0.3, 0.7] \middle| n_1\right) = \int_{0.3}^{0.7} p\left(\theta \middle| n_1\right) d\theta$$

- **Be careful**: This has absolutely nothing to do with confidence intervals.
- In classical statistics, and for an univariate problem, the confidence interval at level α is of the form $\left[\widehat{\theta} - z_{\alpha/2}\widehat{\sigma}, \widehat{\theta} + z_{\alpha/2}\widehat{\sigma}\right]$ where $\widehat{\theta}$ is the classical estimator (say MLE) and $\widehat{\sigma}$ is an estimate of its standard deviation.
- In this frequentist perspective, the true value of the parameter is fixed, and the confidence interval is random, having a probability of $(1-\alpha)$ to actually contain this true value (when we repeat the same experiment a great number of times) and it is not possible to interpret $(1-\alpha)$ as the probability that the parameter lies in the confidence interval for the considered experiment.

Bayesian Inference with the Bernoulli-Beta Model

- We can also find the maximum a posterior (MAP)

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max \; p(\theta|\, n_1) \\ &= \arg\max \; \log p(\theta|\, n_1) \\ &= \arg\max \; \log p(n_1|\, \theta) + \log p(\theta) \\ &= \frac{n_1+\alpha-1}{n_1+\alpha-1+n_0+\beta-1} = \frac{n_1+\alpha-1}{N+\alpha+\beta-2}.\end{aligned}$$

- $\widehat{\theta}_{MAP} = \widehat{\theta}_{MLE}$ when $\alpha = \beta = 1$ as then $\log p(\theta)$ is constant over $[0, 1]$.

Prediction: Classical vs Bayesian Approaches

- Assume you have observed n_1 successes among N trials, we want to use these data to come up with the distribution of the outcome of the next trial.
- Using a Maximum Likelihood approach, we would use the plug-in prediction

$$p\left(x=1|\,\widehat{\theta}_{MLE}\right)=\widehat{\theta}_{MLE}=\frac{n_1}{N}$$

  This does not account whatsoever for the uncertainty about $\widehat{\theta}_{MLE}$ (and suffer from Black Swan problem)
- In a Bayesian approach, we will use the predictive distribution

$$\begin{aligned} p\left(x=1|\, n_1\right) &= \int p\left(x=1|\,\theta\right)p\left(\theta|\, n_1\right)d\theta \\ &= \int \theta p\left(\theta|\, n_1\right)d\theta=\frac{n_1+\alpha}{N+\alpha+\beta} \end{aligned}$$

  so even if $n_1=0$ then $p\left(x=1|\, x_{1:N}\right)>0$ and our prediction takes into account the uncertainty about θ.

Prediction: Classical vs Bayesian Approaches

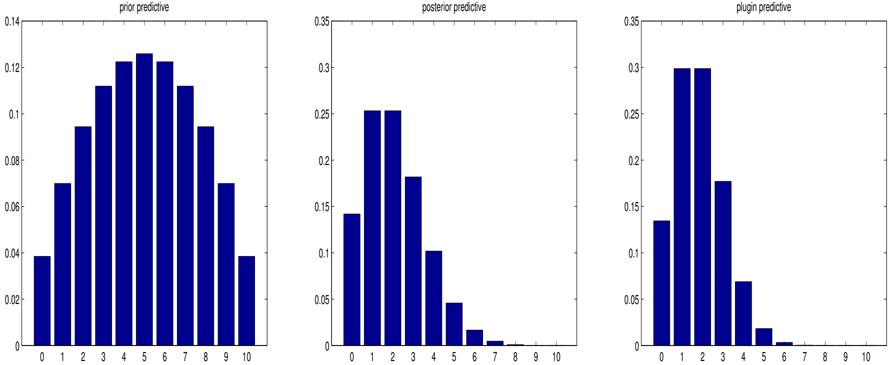
- Suppose now we want to predict the number m_1 of heads in M future trials.
- The standard MLE approach would give use

$$p\left(m_1 \middle| \widehat{\theta}_{MLE}\right) = Bin(m_1; \widehat{\theta}_{MLE}, M) = \begin{pmatrix} M \\ m_1 \end{pmatrix} \widehat{\theta}_{MLE}^{m_1} \left(1 - \widehat{\theta}_{MLE}^{n}\right)^{M-m_1}$$

- The Bayesian approach yields

$$\begin{aligned} & p\left(m_1 | n_1\right) = \int p\left(m_1 | \theta\right) p\left(\theta | n_1\right) d\theta \\ & = \begin{pmatrix} M \\ m_1 \end{pmatrix} \frac{\Gamma(N+\alpha+\beta)}{\Gamma(n_1+\alpha)\Gamma(N-n_1+\beta)} \int \theta^{m_1+n_1-1} \left(1-\theta\right)^{N+M-m_1-n_1-1} d\theta \\ & = \begin{pmatrix} M \\ m_1 \end{pmatrix} \frac{\Gamma(N+\alpha+\beta)}{\Gamma(n_1+\alpha)\Gamma(N-n_1+\beta)} \frac{\Gamma(m_1+n_1+\alpha)\Gamma(N+M-m_1-n_1+\beta)}{\Gamma(N+M+\alpha+\beta)} \end{aligned}$$

Prediction: Classical vs Bayesian Approaches



(left) Prior predictive dist. for a Binomial likelihood with $M = 10$ and a Beta(2,2) prior. (center) Posterior predictive after having seen $n_1 = 3, N = 20$. (right) Plug-in approximation using $\widehat{\theta}_{MLE}$).

From Coins to Dice: Multinomial

- Assume you have independent observations $\left\{\mathbf{x}^i\right\}_{i=1}^M$ such that

$$p\left(\mathbf{x}|\,\theta\right) = \frac{P!}{\prod_{i=1}^{d} x_k!}\prod_{k=1}^{d}\theta_k^{x_k}$$

  for $\theta_k > 0$, $\sum_{k=1}^{d}\theta_k = 1$ and $x_k = 0, 1, 2, ..., P$ with $\sum_k x_k = P$.
- We have seen that

$$\widehat{\theta_k}_{,MLE} = \frac{\sum_{i=1}^{M} x_k^i}{\sum_{i=1}^{M}\sum_{k=1}^{d} x_k^i} = \frac{N_k}{N}$$

- We want now to perform a Bayesian analysis

$$p\left(\theta|\,\mathbf{x}^{1:M}\right) = \frac{p\left(\mathbf{x}^{1:M}\middle|\,\theta\right)p\left(\theta\right)}{p\left(\mathbf{x}^{1:M}\right)}.$$

Dirichlet Prior

- The Dirichlet density is given by

$$\mathrm{Dir}\left(\theta; \left(\alpha_1, \ldots, \alpha_d\right)\right) = \frac{\Gamma\left(\sum_{k=1}^{d} \alpha_k\right)}{\prod_{k=1}^{d} \Gamma\left(\alpha_k\right)} \prod_{k=1}^{d} \theta_k^{\alpha_k - 1}$$

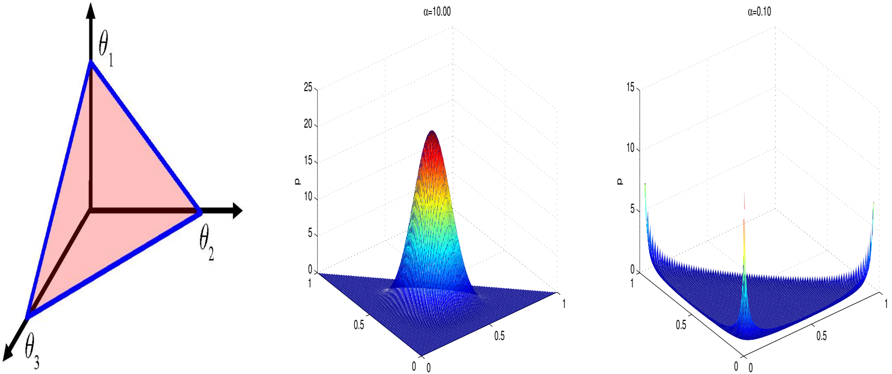
  for $\alpha_k > 0$ and corresponds to a Beta density for $d = 2$. It is defined on $\left\{\theta : \theta_k > 0 \text{ and } \sum_{k=1}^{d} \theta_k = 1\right\}$.

- $\alpha_0 = \sum_{k=1}^{d} \alpha_k$ controls how peaky the distribution is and the α_k controls where the peak is located.

- We have

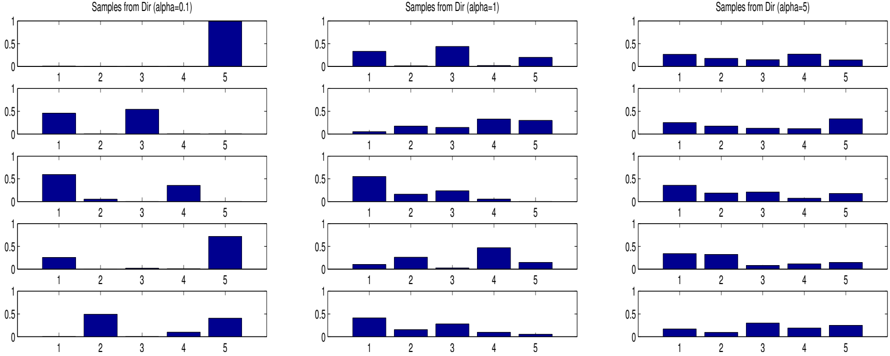
$$\mathbb{E}\left(\theta_k\right) = \frac{\alpha_k}{\alpha_0}, \;\; \mathrm{mode}\left(\theta_k\right) = \frac{\alpha_k - 1}{\alpha_0 - d}, \;\; \mathbb{V}\left(\theta_k\right) = \frac{\alpha_k\left(\alpha_0 - \alpha_k\right)}{\alpha_0^2\left(\alpha_0 + 1\right)}$$

Dirichlet Prior



(left) Support of the Dirichlet density for $d = 3$ (center) Dirichlet density for $\alpha_k = 10$ (right) Dirichlet density for $\alpha_k = 0.1$.

Dirichlet Prior



Samples from a Dirichlet distribution for $d = 5$ when $\alpha_k = \alpha_l$ for $k \neq l$.

Bayesian Inference with the Multinomial-Dirichlet Model

- We obtain

$$
\begin{aligned}
p\left(\theta \middle| \mathbf{x}^{1:M}\right) &= \frac{p\left(\mathbf{x}^{1:M} \middle| \theta\right) p\left(\theta\right)}{p\left(\mathbf{x}^{1:M}\right)} \\
&\propto \prod_{k=1}^{d} \theta_k^{N_k} \prod_{k=1}^{d} \theta_k^{\alpha_k - 1} \\
&\propto \prod_{k=1}^{d} \theta_k^{\alpha_k + N_k - 1}
\end{aligned}
$$

- This implies necessarily that $p\left(\theta \middle| x_{1:M}\right) = \mathrm{Dir}(\theta; \alpha_1 + N_1, \ldots, \alpha_d + N_d)$.

Predictive Distribution with the Multinomial-Dirichlet Model

- We have for a single categorical variable

$$\begin{aligned}
\Pr\left(x = k \middle| \mathbf{x}^{1:M}\right) &= \int \Pr\left(x = k \middle| \theta\right) p\left(\theta \middle| \mathbf{x}^{1:M}\right) d\theta \\
&= \int \theta_k p\left(\theta \middle| \mathbf{x}^{1:M}\right) d\theta \\
&= \int \theta_k p\left(\theta_k \middle| \mathbf{x}^{1:M}\right) d\theta_k \\
&= \frac{\alpha_k + N_k}{\alpha_0 + N}.
\end{aligned}$$

- Once more this avoids the black-swan problem.

Bayesian Naive Bayes for Multinomial Data

- We assume that we have M data $\left(\mathbf{x}^i, y^i\right) \in \mathbb{N}^d \times \{0,1\}^C$ and we use the model

$$p\left(\mathbf{x}, y = c \middle| \theta\right) = \pi_c \, \frac{P!}{\prod_{i=1}^{d} x_k!} \prod_{k=1}^{d} \theta_{k,c}^{x_k}$$

where $\left(\pi_1, ..., \pi_C, \theta_{1,1}, ..., \theta_{d,1}, \cdots, \theta_{1,C}, ..., \theta_{d,C}\right)$ are the unknown parameters.

- If we do MLE, then

$$\widehat{\pi}_{c,MLE} = \frac{M_c}{M}, \; \widehat{\theta}_{k,c,MLE} = \frac{N_{k,c}}{N_c}$$

where M_c =nb. documents class c, $N_{k,c}$ =nb. occurrences word k in class c, $M = \sum_{k=1}^{C} M_c$, $N_c = \sum_{k=1}^{d} N_{k,c}$.

Bayesian Naive Bayes for Multinomial Data

- In a Bayesian context, we can set independent Dirichlet priors

$$\begin{aligned} p(\pi) &= \mathrm{Dir}\left((\pi_1, ..., \pi_C); \beta_1, ..., \beta_C\right), \\ p(\theta_c) &= \mathrm{Dir}\left((\theta_{1,c}, ..., \theta_{d,c}); \alpha_{1,c}, ..., \alpha_{d,c}\right), \ \ c = 1, ..., C \end{aligned}$$

  and obtain

$$\begin{aligned} p\left(\pi_1, ..., \pi_C \middle| D\right) &= \mathrm{Dir}\left((\pi_1, ..., \pi_C); \beta_1 + M_1, ..., \beta_C + M_C\right), \\ p\left(\theta_c \middle| D\right) &= \mathrm{Dir}\left((\theta_{1,c}, ..., \theta_{d,c}); \alpha_{1,c} + N_{1,c}, ..., \alpha_{d,c} + N_{d,c}\right). \end{aligned}$$

- From this posterior, you can compute $\widehat{\pi}_{MAP}$, $\widehat{\theta}_{c,MAP}$ or $\widehat{\pi}_{MMSE} = \mathbb{E}\left(\pi_M \middle| D\right)$, $\widehat{\theta}_{c,MMSE} = \mathbb{E}\left(\theta_c \middle| D\right)$ and use

$$p\left(y = c \middle| \mathbf{x}, \widehat{\pi}, \widehat{\theta}\right) \propto p\left(y = c \middle| \widehat{\pi}\right) p\left(\mathbf{x} \middle| y = c, \widehat{\theta}\right)$$

Bayesian Naive Bayes for Multinomial Data

- A better way to do it is to
- Given a new input $\mathbf{x}$, we compute using

$$p\left(y=c\,|\,\mathbf{x},D\right)\propto p\left(y=c\,|\,D\right)p\left(\mathbf{x}\,|\,y=c,D\right)$$

where

$$\begin{aligned} p\left(y=c\,|\,D\right) &= \int \underbrace{p\left(y=c\,|\,D,\pi\right)}_{=\pi_c} p\left(\pi\,|\,D\right)d\pi = \frac{\beta_c+M_c+1}{\beta_0+M+1}, \\ p\left(\mathbf{x}\,|\,y=c,D\right) &= \int p\left(\mathbf{x}\,|\,D,\theta_c\right)p\left(\theta_c\,|\,D\right)d\theta_c = \ldots \end{aligned}$$

Bayesian Inference for Normal Data

- Assume you have independent data $\left\{x^i\right\}_{i=1}^N$ such that

$$p\left(x|\,\theta\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right)$$

where $\theta=\left(\mu,\sigma^2\right)$.

- We have seen that

$$\widehat{\mu}_{ML}=\frac{1}{N}\sum_{i=1}^N x^i,\ \widehat{\sigma^2}_{ML}=\frac{1}{N}\sum_{i=1}^N\left(x^i-\widehat{\mu}_{ML}\right)^2.$$

Bayesian Inference for Normal Data

- In a Bayesian framework, the conjugate prior is $p\left(\mu,\sigma^2\right)=p\left(\mu|\,\sigma^2\right)p\left(\sigma^2\right)$ where

$$p\left(\mu|\,\sigma^2\right)=\mathcal{N}\left(\mu;\mu_0,\frac{\sigma^2}{\kappa_0}\right)=\frac{1}{\sqrt{2\pi(\sigma^2/\kappa_0)}}\exp\left(-\frac{\kappa_0(\mu-\mu_0)^2}{2\sigma^2}\right)$$
$$p\left(\sigma^2\right)=\mathcal{IG}\left(\sigma^2;\alpha,\beta\right)=\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\sigma^2\right)^{-\alpha-1}\exp\left(-\beta/\sigma^2\right)1_{(0,\infty)}\left(\sigma^2\right).$$

- The posterior is given by $p\left(\theta|\,x^{1:n}\right)=p\left(\sigma^2\middle|\,x^{1:n}\right)p\left(\mu|\,x^{1:n},\sigma^2\right)$ where

$$p\left(\mu|\,x^{1:N},\sigma^2\right)=\mathcal{N}\left(\mu;\frac{\kappa_0\mu_0+N\widehat{\mu}_{ML}}{\kappa_0+N},\frac{\sigma^2}{\kappa_0+N}\right)$$
$$p\left(\sigma^2\middle|\,x_{1:N}\right)=\mathcal{IG}\left(\sigma^2;\alpha+N/2,\beta+\frac{N}{2}\,\widehat{\sigma^2}_{ML}+\frac{N\kappa_0}{2(N+\kappa_0)}\left(\widehat{\mu}_{ML}-\mu_0\right)^2\right)$$

- Once more we see clearly the influence of the prior on the posterior and, as $N\rightarrow\infty$, the posterior concentrates around $\widehat{\mu}_{ML}$ and $\widehat{\sigma^2}_{ML}$.

Bayesian Model Selection

- Suppose we have K different models for the data D; each model being associated to some parameters θ_i.
- Using a Bayesian approach, we can compte

$$p(M = i|D) = \frac{p(M = i)\,p(D|M = i)}{P(D)}$$

where

$$p(D) = \sum_{i=1}^{K} p(M = i)\,p(D|M = i)$$

- The *marginal likelihood* or *evidence* $p(D|M = i)$ is given by

$$p(D|M = i) = \int p(D|\theta_i)\,p(\theta_i)\,d\theta_i$$

which is the normalizing constant of

$$p(\theta_i|D) = \frac{p(\theta_i)\,p(D|\theta_i)}{p(D)}.$$

Bayes Factors

- To compare two models, we can use posterior odds of Bayes factors

$$\underbrace{\frac{p\left(M=i|\,D\right)}{p\left(M=j|\,D\right)}}_{\text{posterior odds}}=\underbrace{\frac{p\left(D|\,M=i\right)}{p\left(D|\,M=j\right)}}_{\text{Bayes factor}}\underbrace{\frac{p\left(M=i\right)}{p\left(M=j\right)}}_{\text{prior odds}}$$

- The Bayes factor is a Bayesian version of a likelihood ratio test, that can be used to compare models of different complexity.
- Bayes factors and posterior odds tell you whether one should prefer $M=i$ to $M=j$: it does NOT tell you whether these models are sensible!

Example: Is the Euro coin biased?

- Suppose we toss a coin $N = 250$ times and observe $n_1 = 141$ heads and $n_0 = 109$ tails:

$$p\left(x^{1:N} \middle| \theta\right) = \theta^{n_1} (1-\theta)^{n_0}$$

- Consider two models/hypotheses: $M_1 =$ coin unbiased, that is $\theta_1 = 0.5$ and $M_2 =$coin biased and $p(\theta_2) = Beta(\theta_2; \alpha_1, \alpha_0)$.
- We have

$$p(D|M_1) = 0.5^{n_1} (1-0.5)^{n_0} = 0.5^N$$

and

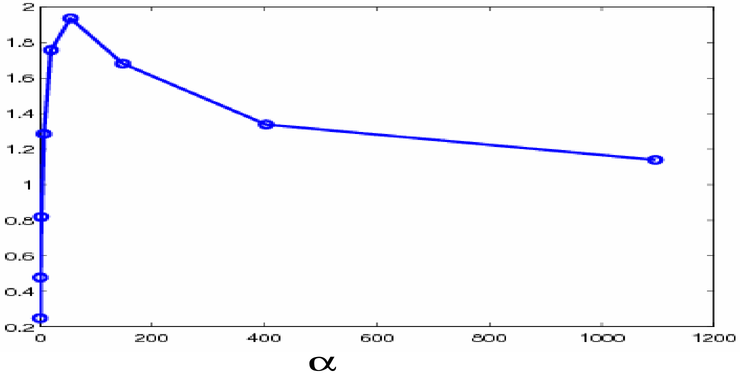
$$p(D|M_2) = \frac{\Gamma(\alpha_0 + n_0)\,\Gamma(\alpha_1 + n_1)}{\Gamma(\alpha_0 + \alpha_1 + N)} \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\,\Gamma(\alpha_1)}$$

so

$$\frac{p(D|M_2)}{p(D|M_1)} = \frac{\Gamma(\alpha_0 + n_0)\,\Gamma(\alpha_1 + n_1)}{\Gamma(\alpha_0 + \alpha_1 + N)} \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\,\Gamma(\alpha_1)} 0.5^{-N}$$

Computation of Bayes Factors

- Let $\alpha = \alpha_0 = \alpha_1$ varying over 0 to 1000.



Bayes factor $p\left(D|M_2\right) / p\left(D|M_1\right)$ as a function of α.

- The largest BF in favor of M_2 (biased coin) is only 2.0, which is very weak evidence of bias.

Bayesian Computation

- For complex Bayesian models, we cannot compute the posterior and marginal likelihood analytically.
- In such cases, analytical (Laplace, variational) and Monte Carlo methods approximations are necessary.
- For example, a crude approximation of the marginal likelihood is provided by the Bayesian Information Criterion

$$\log p\left(D \middle| M_i\right) = \log p\left(D \middle| \theta_i^{MLE}\right) - \frac{d}{2} \log n$$

where n is the number of data and d is the dimension/number of free parameters