CS340 Machine learning
Bayesian statistics 3

Outline

- Conjugate analysis of μ and σ^2
- Bayesian model selection
- Summarizing the posterior

Unknown mean and precision

- The likelihood function is

$$
\begin{aligned}
p(D|\mu,\lambda) &= \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
&= \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp\left(-\frac{\lambda}{2}\left[n(\mu-\overline{x})^2+\sum_{i=1}^{n}(x_i-\overline{x})^2\right]\right)
\end{aligned}
$$

- The natural conjugate prior is normal gamma

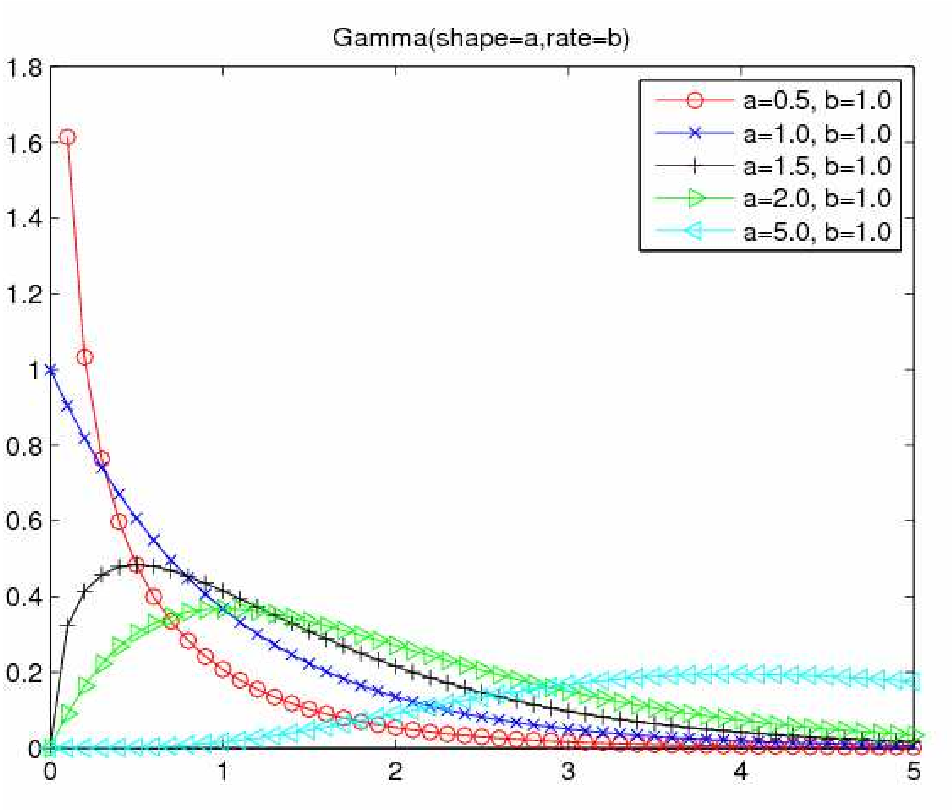
$$
\begin{aligned}
p(\mu,\lambda) &= NG(\mu,\lambda|\mu_0,\kappa_0,\alpha_0,\beta_0) \\
&\stackrel{\text{def}}{=} \mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})Ga(\lambda|\alpha_0,\text{rate}=\beta_0) \\
&= \frac{1}{Z_{NG}}\lambda^{\alpha_0-\frac{1}{2}} \exp\left(-\frac{\lambda}{2}\left[\kappa_0(\mu-\mu_0)^2+2\beta_0\right]\right)
\end{aligned}
$$

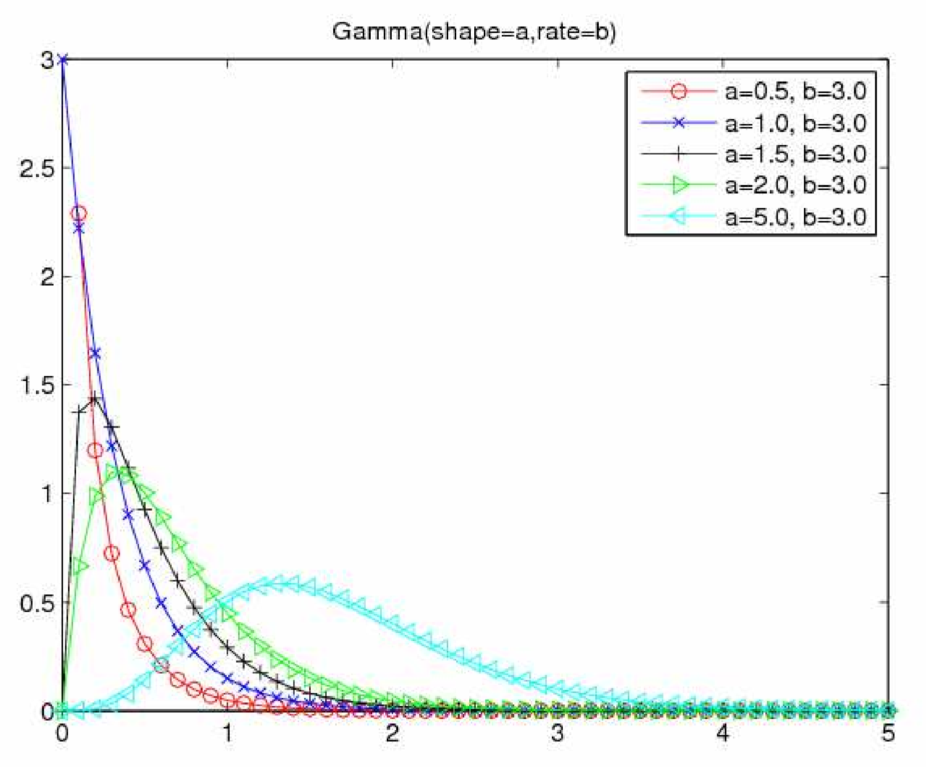
Gamma distribution

- Used for positive reals

$$Ga(x|\text{shape}=a,\text{rate}=b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}, \quad x,a,b>0$$ Bishop

$$Ga(x|\text{shape}=\alpha,\text{scale}=\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}$$ Matlab





Posterior is also NG

- Just update the hyper-parameters

$$
\begin{aligned}
p(\mu, \lambda | D) &= NG(\mu, \lambda | \mu_n, \kappa_n, \alpha_n, \beta_n) \\
\mu_n &= \frac{\kappa_0 \mu_0 + n\overline{x}}{\kappa_0 + n} \\
\kappa_n &= \kappa_0 + n \\
\alpha_n &= \alpha_0 + n/2 \\
\beta_n &= \beta_0 + \tfrac{1}{2} \sum_{i=1}^{n} (x_i - \overline{x})^2 + \frac{\kappa_0 n (\overline{x} - \mu_0)^2}{2(\kappa_0 + n)}
\end{aligned}
$$

Derivation of this result not on exam

Posterior marginals

- Variance

$$p(\lambda|D) \quad = \quad Ga(\lambda|\alpha_n, \beta_n)$$

- Mean

$$p(\mu|D) \quad = \quad T_{2\alpha_n}(\mu|\mu_n, \frac{\beta_n}{\alpha_n \kappa_n})$$

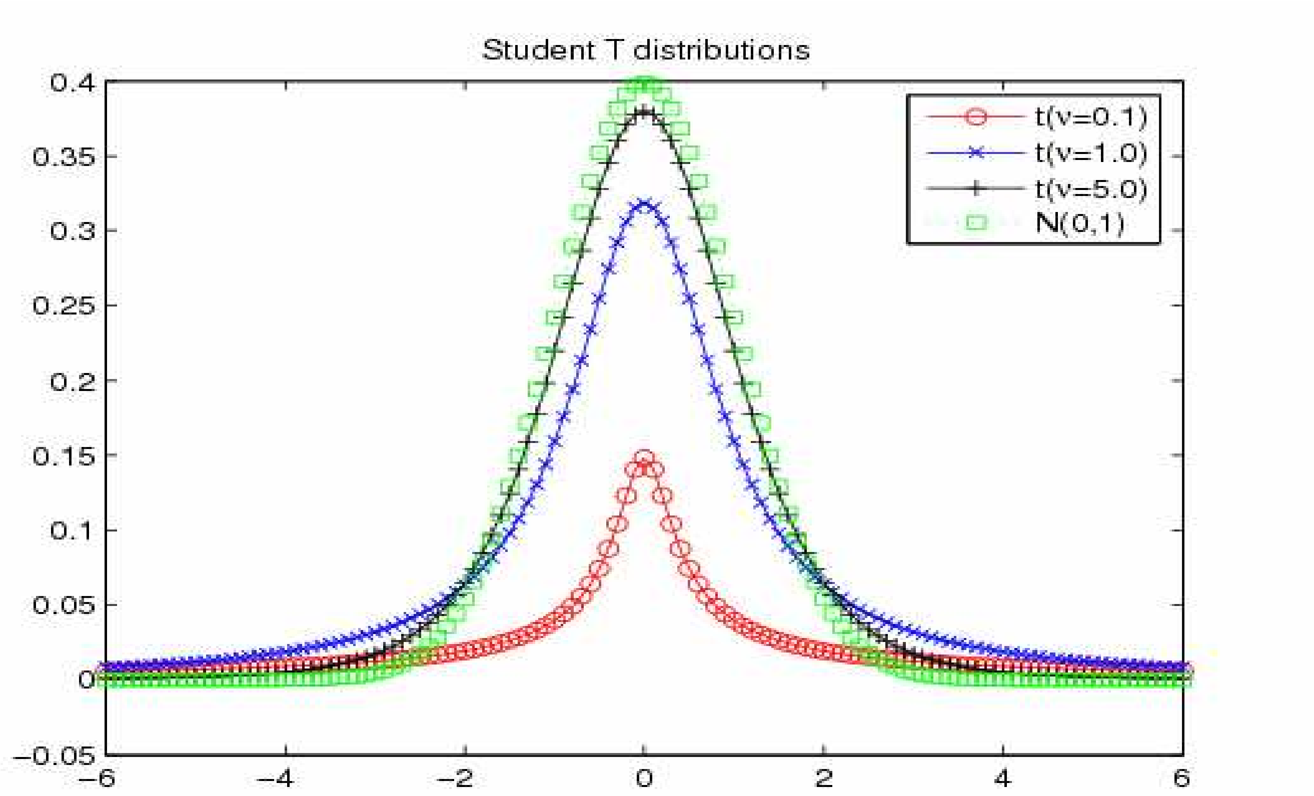
Student t distribution

Derivation of this result not on exam

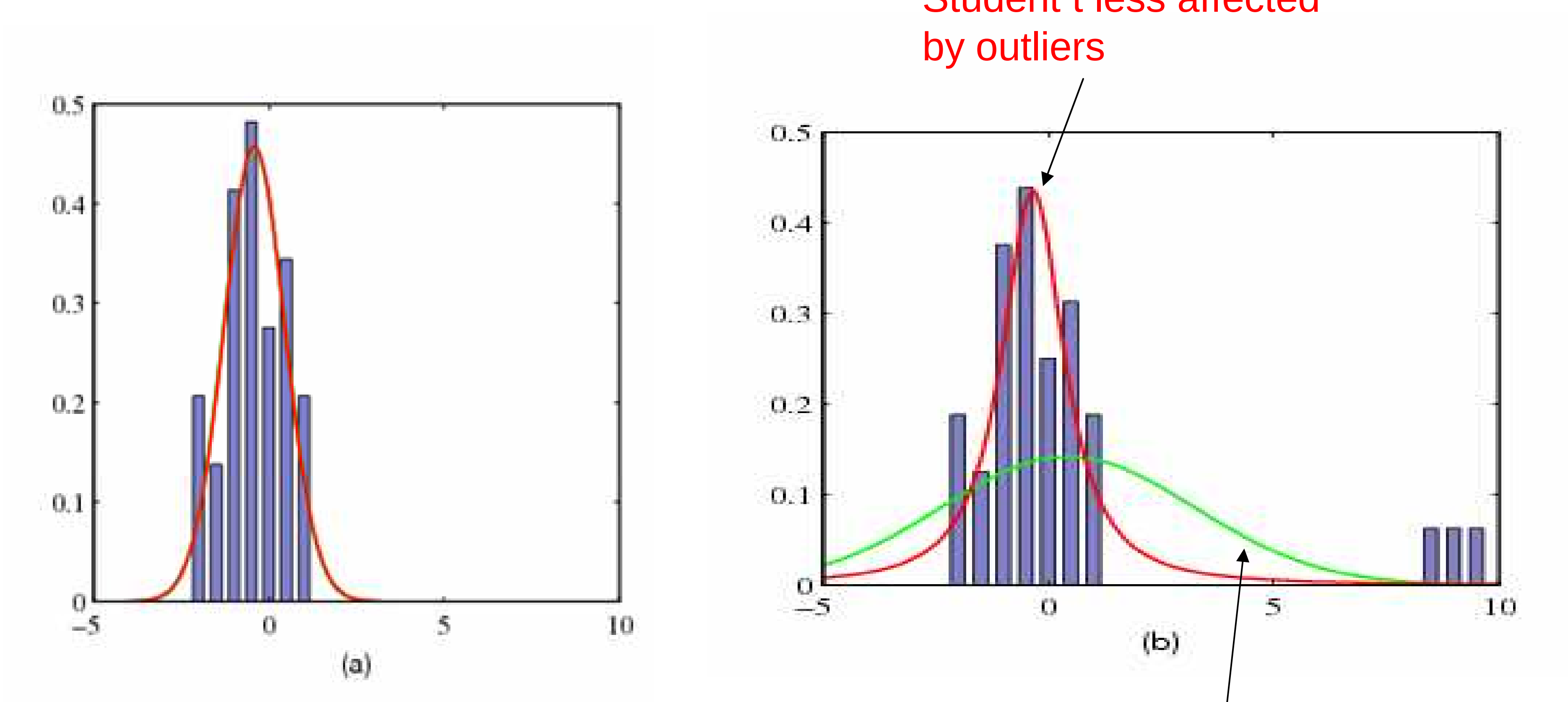
Student t distribution

- Approaches Gaussian as $\nu \to \infty$

$$t_\nu(x|\mu,\sigma^2) \quad \propto \quad \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)}$$



Robustness of t distribution



Bishop 2.16

Posterior predictive distribution

- Also a t distribution (fatter tails than Gaussian due to uncertainty in λ)

$$p(x|D) = t_{2\alpha_n}(x|\mu_n, \frac{\beta_n(\kappa_n + 1)}{\alpha_n \kappa_n})$$

Derivation of this result not on exam

Uninformative prior

- It can be shown (see handout) that an uninformative prior has the form

$$p(\mu, \lambda) \propto \frac{1}{\lambda}$$

- This can be emulated using the following hyper-parameters

$$\begin{aligned} \kappa_0 &= 0 \\ a_0 &= -\frac{1}{2} \\ b_0 &= 0 \end{aligned}$$

- This prior is improper (does not integrate to 1), but the posterior is proper if n $\geq$ 1

Derivation of this result not on exam

Outline

- Conjugate analysis of μ and σ^2
- Bayesian model selection
- Summarizing the posterior

Bayesian model selection

- Suppose we have K possible models, each with parameters θ_i. The posterior over models is defined using the marginal likelihood ("evidence") p(D|M=i), which is the normalizing constant from the posterior over parameters

$$
\begin{aligned}
p(M=i|D) &= \frac{p(M=i)p(D|M=i)}{p(D)} \\
p(D|M=i) &= \int p(D|\theta, M=i)p(\theta|M=i)d\theta \\
p(\theta|D, M=i) &= \frac{p(D|\theta, M=i)p(\theta|M=i)}{p(D|M=i)}
\end{aligned}
$$

Bayes factors

- To compare two models, use posterior odds

$$O_{ij} = \frac{p(M_i|D)}{p(M_j|D)} = \frac{p(D|M_i)p(M_i)}{p(D|M_j)p(M_j)}$$

Bayes factor Prior odds

- The Bayes factor BF(i,j) is a Bayesian version of a likelihood ratio test, that can be used to compare models of different complexity

Marginal likelihood for Beta-Bernoulli

- Since we know $p(\theta|D) = Be(\alpha_1', \alpha_0')$

$$
\begin{aligned}
p(\theta|D) &= \frac{p(\theta)p(D|\theta)}{p(D)} \\
&= \frac{1}{p(D)}\left[\frac{1}{B(\alpha_1,\alpha_0)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}\right]\left[\theta^{N_1}(1-\theta)^{N_0}\right] \\
&= \frac{\theta^{\alpha_1'-1}(1-\theta)^{\alpha_0'-1}}{B(\alpha_1',\alpha_0')}
\end{aligned}
$$

- Hence the marginal likelihood is a ratio of normalizing constants

$$
p(D) = \int p(D|\theta)p(\theta)d\theta = \frac{B(\alpha_1',\alpha_0')}{B(\alpha_1,\alpha_0)}
$$

Example: is the Eurocoin biased?

- Suppose we toss a coin N=250 times and observe N_1=141 heads and N_0=109 tails.
- Consider two hypotheses: H_0 that θ=0.5 and H_1 that θ ≠ 0.5. Actually, we can let H_1 be p(θ) = U(0,1), since p(θ=0.5|H_1) = 0 (pdf).
- For H_0, marginal likelihood is

$$p(D|H_0) = 0.5^N$$

- For H_1, marginal likelihood is

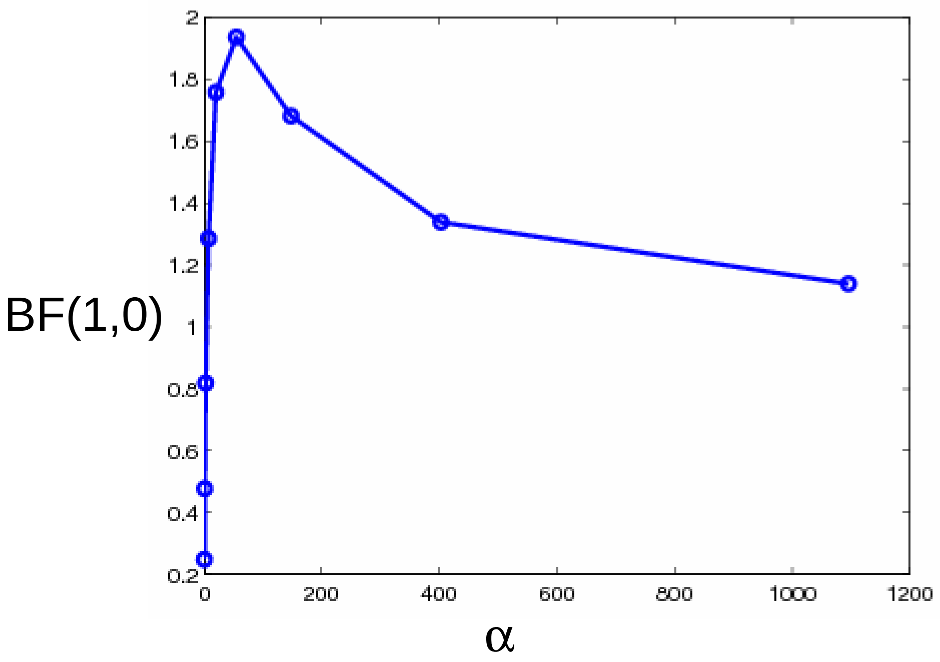
$$P(D|H_1) = \int_0^1 P(D|\theta, H_1) P(\theta|H_1) d\theta = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)}$$

- Hence the Bayes factor is

$$BF(1,0) = \frac{P(D|H_1)}{P(D|H_0)} = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)} \frac{1}{0.5^N}$$

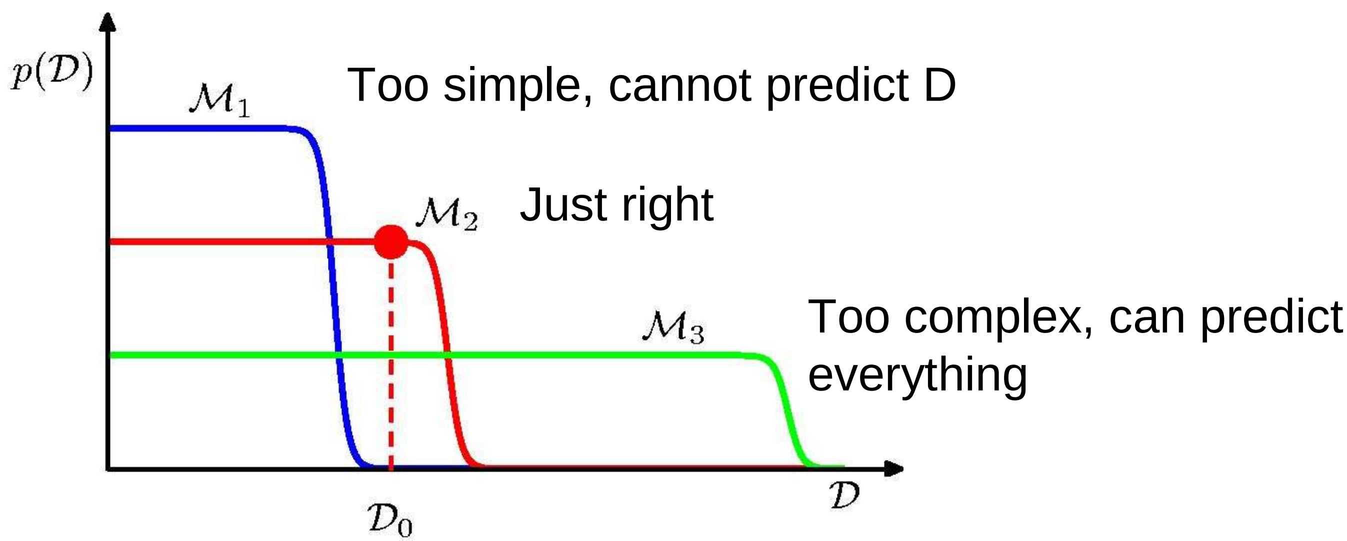
Bayes factor vs prior strength

- Let $\alpha_1=\alpha_0$ range from 0 to 1000.
- The largest BF in favor of H1 (biased coin) is only 2.0, which is very weak evidence of bias.



Bayesian Occam's razor

- The use of the *marginal* likelihood p(D|M) automatically penalizes overly complex models, since they spread their probability mass very widely (predict that everything is possible), so the probability of the actual data is small.



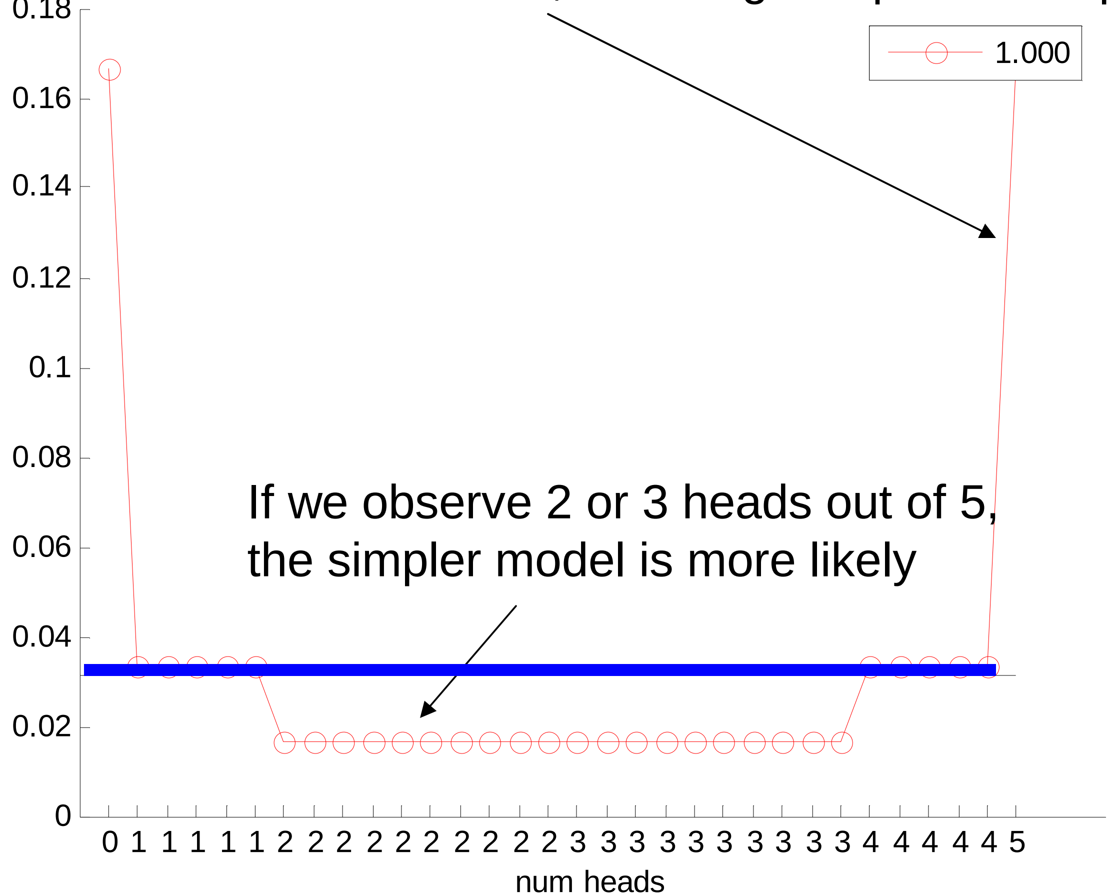
Bishop 3.13

Bayesian Occam's razor for biased coin

Blue line = $p(D|H_0) = 0.5^N$

Red curve = $p(D|H_1) = \int p(D|\theta)\, \mathrm{Beta}(\theta|1,1)\, d\theta$

If we have already observed 4 heads, it is much more likely to observe a 5th head than a tail, since θ gets updated sequentially.



Bayesian Information Criterion (BIC)

- If we make a Gaussian approximation to p(θ|D) (Laplace approximation), and approximate |H| $\approx$ N^d, the log marginal likelihood becomes

$$\log p(D) \approx \log p(D|\theta_{ML}) - \frac{1}{2} d \log N$$

- Here d is the dimension/ number of free parameters.
- AIC (Akaike Info criterion) is defined as

$$\log p(D) \approx \log p(D|\theta_{ML}) - d$$

- Can use penalized log-likelihood for model selection instead of cross-validation.

Outline

- Conjugate analysis of μ and σ^2
- Bayesian model selection
- Summarizing the posterior

Summarizing the posterior

- If p(θ|D) is too complex to plot, we can compute various summary statistics, such as posterior mean, mode and median

$$
\begin{aligned}
\hat{\theta}_{mean} &= E[\theta|\mathcal{D}] \\
\hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|\mathcal{D}) \\
\hat{\theta}_{median} &= t : p(\theta > t|\mathcal{D}) = 0.5
\end{aligned}
$$

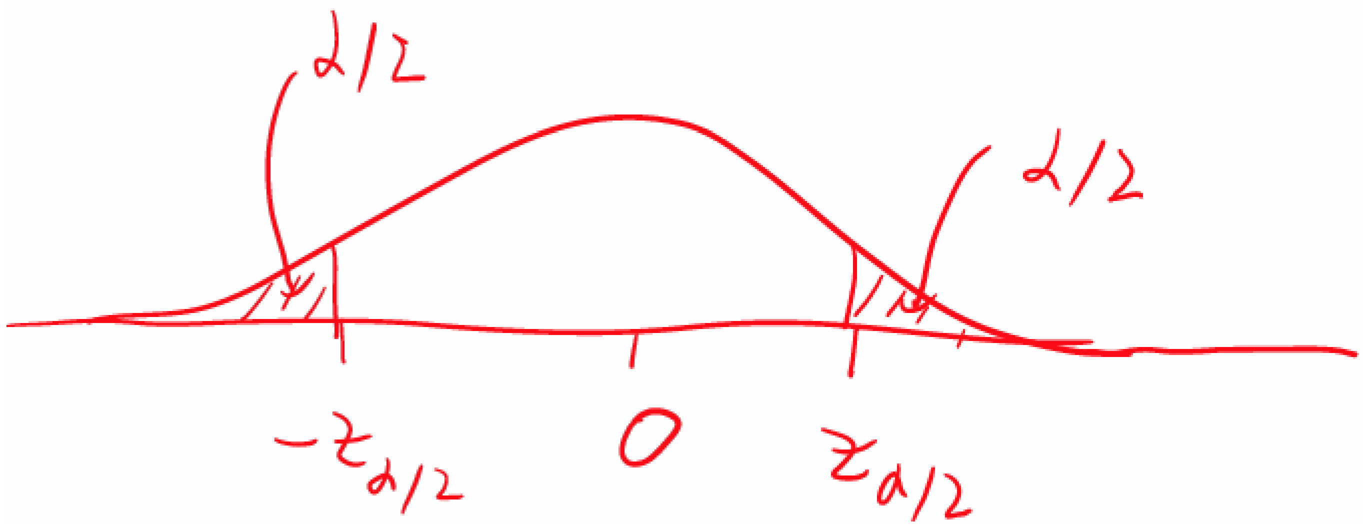
Bayesian credible intervals

- We can represent our uncertainty using a posterior credible interval

$$p(\ell \leq \theta \leq u | D) \geq 1 - \alpha$$

- We set

$$\ell = F^{-1}(\alpha/2), u = F^{-1}(1 - \alpha/2)$$



Example

- We see 47 heads out of 100 trials.
- Using a Beta(1,1) prior, what is the 95% credible interval for probability of heads?

```
S = 47; N = 100; a = S+1; b  = (N-S)+1; alpha = 0.05;
l = betainv(alpha/2, a, b);
u = betainv(1-alpha/2, a, b);
CI = [l,u]
 0.3749    0.5673
```

Posterior sampling

- If θ is high-dimensional, it is hard to visualize $p(\theta|D)$.
- A common strategy is to draw typical values $\theta^s \sim p(\theta|D)$ and analyze the resulting samples.
- Eg we can generate fake data $p(x^s|\theta^s)$ to see if it looks like the real data (a simple kind of posterior predictive check of model adequacy).
- See handout for some examples.