Outline

• Conjugate analysis of $\mu$ and $\sigma^2$
• Bayesian model selection
• Summarizing the posterior
Unknown mean and precision

• The likelihood function is

\[
p(D|\mu, \lambda) = \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp \left( -\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)
\]

\[
= \frac{1}{(2\pi)^{n/2}} \lambda^{n/2} \exp \left( -\frac{\lambda}{2} \left[ n(\mu - \bar{x})^2 + \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \right)
\]

• The natural conjugate prior is normal gamma

\[
p(\mu, \lambda) = NG(\mu, \lambda | \mu_0, \kappa_0, \alpha_0, \beta_0)
\]

\[
def = N(\mu | \mu_0, (\kappa_0 \lambda)^{-1}) Ga(\lambda | \alpha_0, \text{rate} = \beta_0)
\]

\[
= \frac{1}{Z_{NG}} \lambda^{\alpha_0 - \frac{1}{2}} \exp \left( -\frac{\lambda}{2} \left[ \kappa_0 (\mu - \mu_0)^2 + 2 \beta_0 \right] \right)
\]
Gamma distribution

- Used for positive reals

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

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

- Just update the hyper-parameters

\[
p(\mu, \lambda|D) = NG(\mu, \lambda|\mu_n, \kappa_n, \alpha_n, \beta_n)
\]

\[
\mu_n = \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_0 + n}
\]

\[
\kappa_n = \kappa_0 + n
\]

\[
\alpha_n = \alpha_0 + n/2
\]

\[
\beta_n = \beta_0 + \frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2 + \frac{\kappa_0 n(\bar{x} - \mu_0)^2}{2(\kappa_0 + n)}
\]

Derivation of this result not on exam
Posterior marginals

- Variance
  \[ p(\lambda | D) = \text{Ga}(\lambda | \alpha_n, \beta_n) \]

- Mean
  \[ p(\mu | D) = T_{2\alpha_n}(\mu | \mu_n, \frac{\beta_n}{\alpha_n \kappa_n}) \]

Student t distribution

Derivation of this result not on exam
Student t distribution

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

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

Student t less affected by outliers

Gaussian

Bishop 2.16
Posterior predictive distribution

- Also a t distribution (fatter tails than Gaussian due to uncertainty in $\lambda$)

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

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

\[
\begin{align*}
\kappa_0 &= 0 \\
a_0 &= -\frac{1}{2} \\
b_0 &= 0
\end{align*}
\]

• 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 $\mu$ and $\sigma^2$
- Bayesian model selection
- Summarizing the posterior
Bayesian model selection

• Suppose we have K possible models, each with parameters $\theta_i$. 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

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

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

$$p(\theta|D, M = i) = \frac{p(D|\theta, M = i)p(\theta|M = i)}{p(D|M = i)}$$
Bayes factors

• To compare two models, use posterior odds

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

• 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 \( p(\theta|D) = \text{Be}(\alpha_1', \alpha_0') \)

\[
p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} \]

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

- 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 N₁=141 heads and N₀=109 tails.

• Consider two hypotheses: H₀ that θ=0.5 and H₁ that θ ≠ 0.5. Actually, we can let H₁ be p(θ) = U(0,1), since p(θ=0.5|H₁) = 0 (pdf).

• For H₀, marginal likelihood is

\[ p(D|H₀) = 0.5^N \]

• For H₁, marginal likelihood is

\[ P(D|H₁) = \int_0^1 P(D|θ, H₁)P(θ|H₁)dθ = \frac{B(α₁ + N₁, α₀ + N₀)}{B(α₁, α₀)} \]

• Hence the Bayes factor is

\[ BF(1, 0) = \frac{P(D|H₁)}{P(D|H₀)} = \frac{B(α₁ + N₁, α₀ + N₀)}{B(α₁, α₀)} \frac{1}{0.5^N} \]
Bayes factor vs prior strength

• Let $\alpha_1=\alpha_0$ range from 0 to 1000.
• The largest BF in favor of $H_1$ (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.

![Graph showing Bayesian Occam's razor](image)

- Too simple, cannot predict D
- Just right
- Too complex, can predict everything

Bishop 3.13
Bayesian Occam’s razor for biased coin

Blue line = \( p(D|H_0) = 0.5^N \)
Red curve = \( p(D|H_1) = \int p(D|\theta) \text{Beta}(\theta|1,1) \, d\theta \)

If we have already observed 4 heads, it is much more likely to observe a 5\textsuperscript{th} head than a tail, since \( \theta \) gets updated sequentially.

If we observe 2 or 3 heads out of 5, the simpler model is more likely
Bayesian Information Criterion (BIC)

• If we make a Gaussian approximation to $p(\theta|D)$ (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 $\mu$ and $\sigma^2$
• Bayesian model selection
• Summarizing the posterior
Summarizing the posterior

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

\[
\hat{\theta}_{\text{mean}} = E[\theta|D]
\]
\[
\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\theta|D)
\]
\[
\hat{\theta}_{\text{median}} = t : p(\theta > t|D) = 0.5
\]
Bayesian credible intervals

• We can represent our uncertainty using a posterior credible interval

\[ p(\ell \leq \theta \leq u \mid D) \geq 1 - \alpha \]

• 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 $\theta$ 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.