ST4234 Cheatsheet

1. Miscellaneous

Beta Distribution: $X \sim Beta(\alpha, \beta)$

- $\pi(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
- $E[X] = \frac{\alpha}{\alpha+\beta}$, $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Normal approx: when both α and β are large

Gamma Distribution: $X \sim Gamma(\alpha, \frac{1}{\beta})$

- α is the *shape*, $\frac{1}{\beta}$ is the *rate*
- $\pi(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
- $E[X] = \frac{\alpha}{\beta}$, $Var(X) = \frac{\alpha}{\beta^2}$
- $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx = (\alpha-1)!$, $\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \alpha$, $\Gamma(\frac{1}{2}+n) = \frac{(2n)!}{4^n n!}\sqrt{\pi}$
- Normal approx: when α is large

Poisson Distribution: $X \sim Po(\lambda)$

- $\pi(x) = \frac{e^{-\lambda}\lambda^x}{x!}$
- $E[X] = \lambda$, $Var(X) = \lambda$

Exponential Distribution: $X \sim Exp(\frac{1}{\lambda})$

- $\pi(x) = \lambda e^{-\lambda x}$ where $x > 0$
- $E[X] = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$

Normal Distribution: $X \sim N(\mu, \frac{1}{\tau})$

- $\pi(x) = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(x-\mu)^2}$

T-Distribution (Standard): $X \sim t_v$

- $\pi(x) = \frac{1}{B(v/2,1/2)} \frac{1}{\sqrt{v}} (1+\frac{x^2}{v})^{-\frac{v+1}{2}}$
- $E[X] = 0$ for $v > 1$ else ∞
- $Var(X) = \frac{v}{v-2}$ for $v > 2$ else ∞ for $1 < v \le 2$ else undefined

T-Distribution (General): $X \sim t_v(m, \frac{1}{c})$

- $\pi(x) = \frac{1}{B(v/2,1/2)} \sqrt{\frac{c}{v}} (1 + c\frac{(x-m)^2}{v})^{-\frac{v+1}{2}}$
- $E[X] = m$ for $v > 1$ else ∞
- $Var(X) = \frac{1}{c} \cdot \frac{v}{v-2}$ for $v > 2$ else ...

Dirichlet: $\mathbf{X} \sim Dirichlet(\alpha_1, \alpha_2, \alpha_3)$

- Generalisation of *beta* distribution
- $\pi(x_1, x_2, x_3) = \frac{\Gamma(\alpha_1+\alpha_2+\alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} x_3^{\alpha_3-1}$ where $\sum_{i=1}^N x_i = 1$

Pareto: $X \sim Pareto(m, a)$ $(m > 0, a > 0)$

- $\pi(x) = \frac{am^a}{x^{a+1}}$ where $x > m$
- $E[X] = \frac{am}{a-1}$ for $a > 1$ else ∞, mode at m
- $Var(X) = \frac{am^2}{(a-1)^2(a-2)}$ for $a > 2$ else ∞
- $F(x) = 1 - (\frac{m}{x})^a$ for $x > m$

2. Introduction

- Prior: $\pi(\theta)$
- Likelihood: $f(x|\theta)$
- Posterior: $\pi(\theta|x)$
- Posterior $\propto$ Prior $\times$ Likelihood, i.e. $\pi(\theta|x) \propto \pi(\theta) \times f(x|\theta)$

Bayes' Theorem

Bayes' Theorem: $P(A_i|B) \propto P(A_i) \times P(B|A_i)$

- Norm constant is marginal density $P(B)$

Bayes' Theorem for Several RVs: $\pi(y|x_1, \dots, x_k) \propto \pi(y, x_1, \dots, x_k)$

- Sequential updating: $\pi(y|x_1, \dots, x_k) \propto \pi(y|x_1, \dots, x_{k-1}) \cdot \pi(x_k|y, x_1, \dots, x_{k-1})$

3. Bayesian Inference

Bayesian Approach

- Prior: assume θ has prior density $\pi(\theta)$
- Posterior: $\pi(\theta|\mathbf{x}) \propto \pi(\theta) \times \prod_{i=1}^n f(x_i|\theta)$
- Posterior mean: $E[\theta|\mathbf{x}]$ — estimates θ
- Posterior variance: $Var(\theta|\mathbf{x})$
- (1-α) credible set/highest density region (HDR): there is $(1-\alpha)$ chance that interval contains true parameter θ

Useful Identity

- $r(\mu-a)^2 + t(\mu-b)^2 = (r+t)(\mu-\bar{m})^2 + \frac{(a-b)^2}{t^{-1}+r^{-1}}$
- $\bar{m} = \frac{r}{t+r}a + \frac{t}{t+r}b$ — weighted average of a and b

Normal Pop: Known Variance $\tau = r$, Unknown μ

Conjugate family: Normal for mean μ

- Prior: $\mu \sim N(m, \frac{1}{t})$
- Observations: n iid from $X \sim N(\mu, \frac{1}{\tau} = \frac{1}{r})$
- Likelihood: $L(\mu|\mathbf{x}) \propto e^{-\frac{nr}{2}(\mu-\bar{x})^2}$
- Posterior: $\mu|\mathbf{x} \sim N(m_n, \frac{1}{t_n})$
  - $t_n = t + nr$
  - $m_n = w_n m + (1-w_n)\bar{x}$
  - $w_n = \frac{t}{t+nr}$
- Predictive: $X_{n+1}|\mathbf{X} = \mathbf{x} \sim N(m_n, \frac{1}{t_n} + \frac{1}{r})$

Normal Pop: Known Mean $\mu = h$, Unknown τ

Conjugate family: Gamma for precision τ

- Prior: $\tau \sim Gamma(\alpha, \frac{1}{\beta})$
- Observations: n iid from $X|\tau \sim N(\mu = h, \frac{1}{\tau})$
- Likelihood: $L(\mu|\mathbf{x}) \propto \tau^{\frac{n}{2}} e^{-[\frac{1}{2}\sum_{i=1}^n (x_i-h)^2]\tau}$
- Posterior: $\tau|\mathbf{x} \sim Gamma(\alpha_n, \frac{1}{\beta_n})$
  - $\alpha_n = \alpha + \frac{n}{2}$
  - $\beta_n = \beta + \frac{1}{2}\sum_{i=1}^n (x_i - h)^2$
- Predictive: ???

Normal Pop: Unknown Mean and Variance

Conjugate family: Gamma-Normal for (μ, τ)

- Prior: $(\mu, \tau) \sim Gamma-Normal(\alpha, \frac{1}{\beta}; m, \frac{1}{t})$
  - $\tau \sim Gamma(\alpha, \frac{1}{\beta})$, $\mu|\tau \sim N(m, \frac{1}{\tau t})$
  - $\pi(\mu, \tau) \propto \tau^{\alpha-1} e^{-\beta\tau} \cdot \sqrt{\tau} e^{-\frac{\tau t}{2}(\mu-m)^2}$
- Observations: n iid from $X|(\mu, \tau) \sim N(\mu, \frac{1}{\tau})$
- Likelihood: $L(\mu, \tau|\mathbf{x}) \propto \tau^{\frac{n}{2}} e^{-\frac{n\tau}{2}(\mu-\bar{x})^2 - \frac{\tau}{2}\sum_{i=1}^n (x_i-\bar{x})^2}$
- Posterior: $(\mu, \tau|\mathbf{x}) \sim Gamma-Normal(\alpha_n, \frac{1}{\beta_n}; m_n, \frac{1}{t_n})$
  - $\alpha_n = \alpha + \frac{n}{2}$
  - $\beta_n = \beta + \frac{1}{2}[\sum_{i=1}^n (x_i - \bar{x})^2 + \frac{(m-\bar{x})^2}{1/t + 1/n}]$
  - $m_n = \frac{t}{t+n}m + \frac{n}{t+n}\bar{x}$
  - $t_n = t + n$
- Posterior: $\mu|\mathbf{x} \sim t_{2\alpha_n}\left(m_n, (\frac{\alpha_n t_n}{\beta_n})^{-1}\right)$
  - $\pi(\mu|\mathbf{x}) \propto \left[1 + (\frac{\alpha_n t_n}{\beta_n})\frac{(\mu-m_n)^2}{2\alpha_n}\right]^{-(2\alpha_n+1)/2}$
  - $E[\mu|\mathbf{x}] = m_n$
  - $Var(\mu|\mathbf{x}) = (\frac{\alpha_n t_n}{\beta_n})^{-1} \frac{2\alpha_n}{2\alpha_n - 2} = \frac{\beta_n}{t_n(\alpha_n - 1)}$
- Predictive: $X_{n+1}|\mathbf{X} = \mathbf{x} \sim t_{2\alpha_n}\left(m_n, (\frac{\alpha_n t_n}{\beta_n(1+t_n)})^{-1}\right)$

Bernoulli Pop: Unknown Parameter θ

Conjugate family: Beta for θ

- Prior: $\theta \sim Beta(a, b)$
  - $\pi(\theta) = \theta^{a-1}(1-\theta)^{b-1}$
- Observations: n iid from $X|\theta \sim Ber(\theta)$
- Posterior: $\theta|\mathbf{x} \sim Beta(a_n, b_n)$
  - $a_n = a + n\bar{x}$ — add successes
  - $b_n = b + n - n\bar{x}$ — add failures
  - $n\bar{x} = \sum_{i=1}^n x_i$
- Predictive: $X_{n+1}|\mathbf{x} \sim Ber(\frac{a_n}{a_n+b_n})$
  - $f_{X_{n+1}}(1|\mathbf{X} = \mathbf{x}) = \frac{a_n}{a_n+b_n}$
  - $f_{X_{n+1}}(0|\mathbf{X} = \mathbf{x}) = \frac{b_n}{a_n+b_n}$

Poisson Pop: Unknown Parameter λ

Conjugate family: Gamma for λ

- Prior: $\lambda \sim Gamma(\alpha, \frac{1}{\beta})$
- Observations: n iid from $X|\lambda \sim Po(\lambda)$
- Posterior: $\lambda|\mathbf{X} = \mathbf{x} \sim Gamma(\alpha_n, \frac{1}{\beta_n})$
  - $\alpha_n = \alpha + n\bar{x}$ — add sum of observations
  - $\beta_n = \beta + n$ — add number of observations
  - $n\bar{x} = \sum_{i=1}^n x_i$

Exponential Pop: Unknown Parameter λ

Conjugate family: Gamma for λ

- Prior: $\lambda \sim Gamma(\alpha, \frac{1}{\beta})$
- Observations: n iid from $X|\lambda \sim Exp(\lambda)$
- Posterior: $\lambda|\mathbf{x} \sim Gamma(\alpha_n, \frac{1}{\beta_n})$
  - $\alpha_n = \alpha + n$ — add number of observations
  - $\beta_n = \beta + n\bar{x}$ – add sum of observations
  - $n\bar{x} = \sum_{i=1}^n x_i$
- Predictive: $X_{n+1}|\mathbf{X} = \mathbf{x} + \beta_n \sim Pareto(\beta_n, \alpha_n)$
  - $P(X_{n+1} \le x|\mathbf{X} = \mathbf{x}) = 1 - \left(\frac{\beta_n}{x+\beta_n}\right)^{\alpha_n}$

Uniform Pop: Unknown Parameter θ

Conjugate family: Pareto for θ

- Prior: $\theta \sim Pareto(m, a)$
  - $\pi(\theta) = \frac{am^a}{\theta^{a+1}} \mathbf{I}_{\theta > m}$
- Observations: n iid from $X|\theta \sim U(0, \theta)$
  - $f(x|\theta) = \frac{1}{\theta}\mathbf{I}_{0<x<\theta}$
- Posterior: $\theta|\mathbf{x} \sim Pareto(m_n, a_n)$
  - $m_n = \max(m, x_{\max})$ — set to max of self/obs
  - $a_n = a + n$ — add number of observations
  - $x_{\max} = \max_{i=1}^n x_i$

Multinomial Pop: Unknown Parameter p

Conjugate family: Dirichlet for $(p_1, \dots, p_k)$

- Prior: $(p_1, \dots, p_k) \sim Dirichlet(\alpha_1, \dots, \alpha_k)$
- Observations: M iid $\mathbf{x}_i$ from $X \sim Multinomial(n; p_1, \dots, p_k)$
- Posterior: $(p_1, \dots, p_k|\mathbf{x}_1, \dots, \mathbf{x}_M) \sim Dirichlet(\alpha_{n1}, \dots, \alpha_{nk})$
  - $\alpha_{nj} = \alpha_j + m_j$ — within each category, add sum of x across it
  - $m_j = \sum_{i=1}^M x_{ij}$

4. Conjugate Prior Distributions

Conjugate family: A class Π of probability distributions forms a *conjugate family* if the posterior density $\pi(\theta|x) \propto \pi(\theta) \cdot f(x|\theta)$ is in Π for all x, whenever the prior density $\pi(\theta)$ is in Π

5. Predictive Distributions

Proposition: Predictive Distribution

- $F_{X_{n+1}}(x|\mathbf{X} = \mathbf{x}) = E[F(x|\theta)|\mathbf{X} = \mathbf{x}]$
- $f_{X_{n+1}}(x|\mathbf{X} = \mathbf{x}) = E[f(x|\theta)|\mathbf{X} = \mathbf{x}] = \int_{\theta\in\Theta} f(x|\theta) \cdot \pi(\theta|\mathbf{X} = \mathbf{x})$

Double expectation formula

Standard form: $E(Y) = E[E(Y|Z)]$

- Inner $E(Y|Z)$ is expectation wrt Y, yields function of Z
- Outer $E[E(Y|Z)]$ is expectation wrt Z

- Conditional form: $E(Y|X) = E[E(Y|Z,X)|X]$
- E.g. $E[x_{n+1}|\mathbf{x}] = E[E[x_{n+1}|\theta]|\mathbf{x}]$

$Var(Y) = E[Var(Y|Z)] + Var(E[Y|Z])$

- E.g. $Var(x_{n+1}|\mathbf{x}) = Var(E[x_{n+1}|\theta]|\mathbf{x}) + E[Var(x_{n+1}|\theta)|\mathbf{x}]$

6. Hypothesis Testing

Test between $\{\theta = \theta_1\}$ and $\{\theta = \theta_2\}$

- $O_n = \frac{P(\theta=\theta_1|\mathbf{x})}{P(\theta=\theta_2|\mathbf{x})}$
- Favour θ_1 if $P(\theta = \theta_1|\mathbf{x}) > 0.5 \Leftrightarrow O_n > 1$
- Favour θ_2 if $P(\theta = \theta_2|\mathbf{x}) > 0.5 \Leftrightarrow O_n < 1$

Test between $\{\theta \in \Theta_1\}$ and $\{\theta \in \Theta_2\}$

Case 1: $\Theta_1 \cup \Theta_2 = \Theta$

- $P(\theta \in \Theta_1|\mathbf{x}) = \int_{\theta\in\Theta_1} \pi(\theta|\mathbf{x})\ d\theta$
- $P(\theta \in \Theta_2|\mathbf{x}) = \int_{\theta\in\Theta_2} \pi(\theta|\mathbf{x})\ d\theta$

Case 2: $\Theta_1 \cup \Theta_2 \neq \Theta$

- Probabilities need to be re-normalised
- $P(\theta \in \Theta_1|\mathbf{x}, \theta \in \Theta_1 \cup \Theta_2) = \frac{P(\theta\in\Theta_1|\mathbf{x})}{P(\theta\in\theta_1|\mathbf{x})+P(\theta\in\theta_2|\mathbf{x})}$
- $P(\theta \in \Theta_2|\mathbf{x}, \theta \in \Theta_1 \cup \Theta_2) = \frac{P(\theta\in\Theta_2|\mathbf{x})}{P(\theta\in\theta_1|\mathbf{x})+P(\theta\in\theta_2|\mathbf{x})}$

Mixture Priors

For some choice of $P(\theta \in \Theta_1)$ and $P(\theta \in \Theta_2)$:

- $P(\theta \in \Theta_1|\mathbf{x}) \propto P(\theta \in \Theta_1) \cdot \int_{\Theta_1} \pi(\theta|\theta \in \Theta_1) \cdot f(\mathbf{x}|\theta)\ d\theta$
- $P(\theta \in \Theta_2|\mathbf{x}) \propto P(\theta \in \Theta_2) \cdot \int_{\Theta_2} \pi(\theta|\theta \in \Theta_2) \cdot f(\mathbf{x}|\theta)\ d\theta$

Test between $\{\theta = \theta_1\}$ and $\{\theta \in \Theta_2\}$

Let $\theta|\theta \in \Theta_2$ be some proper density.

- $\theta = \theta_1$ with probability $P(\theta = \theta_1) = p$
- $\theta \in \Theta_2$ with probability $P(\theta \in \Theta_2) = 1 - p$
- $P(\theta = \theta_1|\mathbf{x}) \propto P(\theta = \theta_1) \cdot f(\mathbf{x}|\theta_1)$
- $P(\theta \in \Theta_2|\mathbf{x}) \propto P(\theta \in \Theta_2) \cdot \int_{\Theta_2} \pi(\theta|\theta \in \Theta_2) \cdot f(\mathbf{x}|\theta)\ d\theta$

Normalisation constant:
$p\pi(\mathbf{x}|\theta_1) + (1-p)\int_{\Theta_2} \pi(\theta|\theta \in \Theta_2)\pi(\mathbf{x}|\theta)\ d\theta$

Nuisance Parameters

- Get likelihood by integrating over nuisance parameters, e.g. $\pi(\mathbf{x}|\theta) = \int \pi(\mathbf{x}|\theta, \lambda) \cdot \pi(\lambda|\theta)\ d\lambda$
- $P(\theta \in \Theta_i|\mathbf{x}) \propto P(\theta \in \Theta_i) \cdot [\int L(\theta, \lambda|\mathbf{x}) \cdot \pi(\lambda|\theta \in \Theta_i)\ d\lambda]$

7. Bayesian Computation

Monte Carlo Integration

$$E[g(\theta)|\mathbf{x}] \approx \frac{1}{M}\sum_{i=1}^{M} g(\theta_i) \text{ as } M \to \infty$$

- Here, θ_i is drawn from posterior $\pi(\theta|\mathbf{x})$
- E.g. $E[\theta|\mathbf{x}] \approx \frac{1}{M}\sum_{i=1}^{M} \theta_i$
- E.g. $E[\theta^k|\mathbf{x}] \approx \frac{1}{M}\sum_{i=1}^{M} \theta_i^k$

Importance Sampling

- Importance density: $h(\theta)$
- Importance weights: $\omega(\theta) = \frac{\pi(\theta)\cdot\pi(\mathbf{x}|\theta)}{h(\theta)}$
- So $\omega(\theta)h(\theta) = \pi(\theta)\pi(\mathbf{x}|\theta)$

Sampling θ_i from $h(\theta)$:

$$E[g(\theta)|\mathbf{x}] \approx \frac{\sum_{i=1}^{M} g(\theta_i) \cdot \omega(\theta_i)}{\sum_{i=1}^{M} \omega(\theta_i)}$$

Note: it's OK to have something proportional to $\omega(\theta)$ instead of the actual, it'll cancel out

8. Markov Chain Monte Carlo

MCMC Approximation

Sampled Markov chain:
$(x_0, y_0), (x_1, y_2), \ldots, (x_M, y_M)$

- Stationary distribution of Markov chain must be identical to desired $\pi(x, y)$
- Then (x_i, y_i) is approximately sampled i.i.d. from $\pi(x, y)$ as $M \to \infty$

Gibbs Sampler

$$E[g(X,Y)] \approx \frac{1}{M}\sum_{i=1}^{M} g(x_{w+i}, y_{w+i})$$

1. Start with some valid initial state (x_0, y_0)
2. For $w + M$ times, from state (x_i, y_i) to next state (x_{i+1}, y_{i+1})
3. Discard the first w members of the chain, keeping M members $(x_{w+1}, y_{w+1}), \ldots, (x_{w+M}, y_{w+M})$ to calculate estimates

Example

1. Let $\mu_0 = 0$, $\tau_0 = 1$
2. Generate μ_{i+1} from $\pi(\mu|\tau_i, \mathbf{x})$; generate τ_{i+1} from $\pi(\tau|\mu_{i+1}, \mathbf{x})$
3. If $i < w + M$, set $i = i + 1$ and go back to (2). Else, stop

9. Other Useful Things

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i$$

Let $(X_1, \ldots, X_k) \sim Multinomial(n, p_1, \ldots, p_k)$.
Find $X_{k-1}|X_1, \ldots, X_{k-2}$.

- $\pi(x_1, \ldots, x_k) = \frac{n!}{x_1!\ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$
- Let $n^* = n - \sum_{i=1}^{k-2} x_i$ and $p^* = 1 - \sum_{i=1}^{k-2} p_i$
- $\pi(x_{k-1}|x_1, \ldots, x_{k-2}) \propto \pi(x_1, \ldots, x_k) \propto \frac{1}{x_{k-1}!(n^*-x_{k-1})!} p_{k-1}^{x_{k-1}} (p^* - p_{k-1})^{n^*-x_{k-1}} \propto \frac{n^*}{x_{k-1}!(n^*-x_{k-1})!} \tilde{p}^{x_{k-1}} (1-\tilde{p})^{n^*-x_{k-1}}$ where $\tilde{p} = p_{k-1}/p^*$
- Then $x_{k-1}|x_1, \ldots, x_{k-2} \sim Bin(n^*, \tilde{p})$

For $Y = \frac{1}{X}$,

$$f_Y(y) = f_X(\frac{1}{y}) \cdot \frac{1}{y^2}$$

For $X \sim Po(\lambda)$,

$$\begin{aligned} E[X] &= \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda}\sum_{x=1}^{\infty} \frac{x\lambda^x}{x!} \\ &= \lambda e^{-\lambda}\sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda}\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned}$$

Sampling from posterior distribution $\pi(\mu|\mathbf{x})$:

$$\begin{aligned} \widehat{Var}(\mu|\mathbf{x}) &= \frac{1}{M-1}\sum_{i=1}^{M}(\mu_i - \hat{\mu})^2 \\ &= \frac{1}{M-1}\left[\sum_{i=1}^{M}\mu_i^2 - M\hat{\mu}^2\right] \end{aligned}$$

Sampling from importance density $h(\mu)$:

$$\begin{aligned} \widehat{Var}(\mu|\mathbf{x}) &= \hat{E}(\mu^2|\mathbf{x}) - \hat{E}^2(\mu|\mathbf{x}) \\ &= \frac{\sum_{i=1}^{M}\mu_i^2 \cdot \omega(\mu_i)}{\sum_{i=1}^{M}\omega(\mu_i)} - \left(\frac{\sum_{i=1}^{M}\mu_i \cdot \omega(\mu_i)}{\sum_{i=1}^{M}\omega(\mu_i)}\right)^2 \end{aligned}$$