Bayesian Book Notes

632 Bayesian Statistics

Common Distributions

1. Binomial: $X \sim B(n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \quad \text{for} \quad \theta \in [0, 1]$
    1. Expectation: $n\theta$
    2. Variance: $n\theta(1-\theta\ )$
    3. Mode: $\lfloor \theta n + 1 \rfloor$
    4. $p(y|n, \theta)$= `dbinom(y,n,theta)`
    5. if X and Y are both binomial, then $X + Y \sim B(n_1 + n_2, \theta)$
    6. When n=1, Binomial becomes Bernoulli
2. Beta: $X \sim \beta(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \quad \text{for} \quad \theta \in [0, 1]$; $\Gamma(a) = (a-1)!$
    1. Expectation: $\frac{a}{a+b}$
    2. Variance: $\frac{ab}{(a+b+1)(a+b)^2} = \frac{E(\theta)E(1-\theta)}{a+b+1}$
    3. Mode: $\frac{a-1}{(a-1)+(b-1)}$
    4. $p(y|a, b)$= `dbeta(y,a,b)`
    5. A multivariate Beta distribution is the dirichlet distribution
3. Poisson: $X \sim \text{Poi}(\theta) = \frac{\theta^y e^{-\theta}}{y!}$
    1. Expectation: θ
    2. Variance: θ
    3. Mode: $\lfloor \theta \rfloor$
    4. $p(y|\theta)$= `dpois(y,theta)`
    5. If X and Y are both poisson, then $X + Y \sim \text{Poi}(\theta_X + \theta_Y)$
    6. If sample mean and sample variance are very different, poisson may not be appropriate
    7. If variance is larger than sample mean negative binomial might be a better fit
4. Gamma: $X \sim \Gamma(\theta|a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}$
    1. Expectation: $\frac{a}{b}$
    2. Variance: $\frac{a}{b^2}$
    3. Mode: $\frac{a-1}{b} \quad \text{for} \quad a \geq 1$; 0 otherwise
    4. $p(y|a, b)$= `dgamma(y,a,b)`

    5. If X and Y are both gamma with same b and different a, $X+Y \sim \Gamma(a_X+a_Y, b)$ and $\frac{X}{X+Y} \sim \beta(a_X, a_Y)$
    6. If X is normal, then $X^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2\sigma^2})$
    7. A chi-square distribution with v degrees of freedom is the same as $\Gamma(\frac{v}{2}, \frac{1}{2})$
5. Inverse Gamma: $X \sim \Gamma^{-1}(\theta|a,b) = \frac{b^a}{\Gamma(a)}\theta^{-a-1}e^{-b\theta}$
    1. Expectation: $\frac{b}{a-1}$ if $a \geq 1$, infinity otherwise
    2. Variance: $\frac{b^2}{(a-1)^2(a-2)}$ if $a \geq 2$, infinity otherwise
    3. Mode: $\frac{b}{a+1}$
    4. Remark: Inverse Gamma is NOT the same as gamma but with 1/x instead of x, its replaced by a factor of x^{-2} as a result of the jacobian in the change of variables formula:
        1. Univariate: $f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{d}{dy}g^{-1}(y)\right|$
        2. Multivariate: $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(h(\mathbf{y})) \cdot \left|\det\left(\frac{\partial h}{\partial \mathbf{y}}\right)\right|$
6. Normal: $X \sim N(\theta, \sigma^2)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right)$
    1. Expectation: θ
    2. Variance: σ^2
    3. Mode: θ
    4. $p(y|\theta,\sigma^2)$= `dnorm(x,theta,sigma)`
    5. Remember that R parameterizes in terms of standard deviation and not variance
    6. If X and Y are normal independent, then $aX+bY+c \sim N(a\theta_X+b\theta_Y+c, a^2\sigma_X^2+b^2\sigma_Y^2)$
    7. Normal sampling model can be used even if data is not normally distributed since most statistical models that require normality generally provide good estimates of mean and variance
7. Multivariate Normal: $X \sim \Phi(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\theta})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta})\right)$
    1. Expectation: $\boldsymbol{\theta}$
    2. Variance: $\boldsymbol{\Sigma}$
    3. Mode: $\boldsymbol{\theta}$
    4. Simulating an MVN can be achieved by a linear transformation of a vector of i.i.d. standard normal random variables.
    5. The following code generates a matrix where the rows are IID samples:
        1. `Z<-matrix(rnorm(n*p),nrow=n,ncol=p)`
        2. `X<-t(t(Z%*%chol(Sigma))+c(theta))`
8. Wishart: $\mathbf{X} \sim \mathcal{W}_p(\nu, \boldsymbol{\Sigma}) = \frac{1}{2^{\nu p/2}|\boldsymbol{\Sigma}|^{\nu/2}\Gamma_p\left(\frac{\nu}{2}\right)}|\mathbf{S}|^{(\nu-p-1)/2}\exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})\right)$
    1. Expectation: $\nu\boldsymbol{\Sigma}$
    2. Variance: For two elements, $\nu\left(\Sigma_{ik}\Sigma_{jl}+\Sigma_{il}\Sigma_{jk}\right)$
    3. Mode: $(\nu-p-1)\,\boldsymbol{\Sigma}, \quad \text{for } \nu > p+1$

   4. The Wishart distribution is the distribution of the sample covariance matrix of multivariate normal samples.

9. Inverse Wishart: $\mathbf{\Sigma} \sim \mathcal{W}_p^{-1}(\nu, \mathbf{\Psi}) = \frac{|\mathbf{\Psi}|^{\nu/2}}{2^{\nu p/2}\Gamma_p(\frac{\nu}{2})} |\mathbf{\Sigma}|^{-(\nu+p+1)/2} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{\Psi}\mathbf{\Sigma}^{-1})\right)$
   1. Expectation: $\frac{\mathbf{\Psi}}{\nu-p-1}, \quad \text{for } \nu > p+1$
   2. Mode: $\frac{\mathbf{\Psi}}{\nu+p+1}, \quad \text{for } \nu > p+1$
   3. The Inverse Wishart distribution is often used as a prior for covariance matrices in Bayesian statistics.

Introduction

1. Bayes doesn't tell use what we should believe in but how they will change after seeing new information
2. Typically use θ for parameter of estimate and y for data, then:
   1. $p(\theta)$ is the prior distribution
   2. $p(y|\theta)$ is the sampling model
   3. $p(\theta|y)$ is the posterior distribution (our update belief
   4. $p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \quad \text{where} \quad p(y) = \int p(y|\theta')p(\theta')d\theta')$
   5. Posterior is proportional to likelihood $*$ Prior: $p(\theta|y) \propto L(\theta)\pi(\theta)$
3. EX: estimating the probability of a rare event
   1. Suppose we want to estimate the prevalence of a rare disease in a city. We take a random sample of 20 invididuals. We can state the following:
      1. $\Theta \in [0,1]$: The person is either infected or not
      2. $Y \in [0,1,\ldots,20]$: Individual observations
      3. This is a binomial event with n=20 and p=θ: $y|\theta \sim \text{Bin}(20,\theta)$
   2. Other studies indicate that prevalence is from 5% to 20%, but it averages around 10%. We choose a beta prior, say 20 people, and 2 are infected (to reflect the 10% average)
      1. Symbolically, $\theta \sim \beta(2,20)$
   3. Suppose in our sample, no one was infected (Y=0). We then make our posterior:
      1. $\theta|y \sim \beta(a+y, b+n-y) = \beta(2+0, 20+20-0) = \beta(2,40)$

One Parameter Models

The Binomial Model

1. Suppose we are testing happiness, so our outcome is 1 if the person is happy and 0 otherwise, and we sample 129 individuals.

   1. The probability of any outcome $p(y|\theta) = \theta^y(1-\theta)^{129-y}$
   2. We found 118 people being generally happy and 11 being not happy
      1. so now the probabiloty of an outcome is $p(y|\theta) = \theta^118(1-\theta)^{11}$
   3. How do we find $p(\theta|y)$?
      1. $p(\theta|y) = p(y|\theta)p(y) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{(y+1)-1}(1-\theta)^{(n-y+1)-1}$
      2. $p(\theta|y) = p(y|\theta)p(y) = \frac{\Gamma(129+2)}{\Gamma(118+1)\Gamma(129-118+1)}\theta^{(118+1)-1}(1-\theta)^{(129-118+1)-1} = \beta(119,12)$
         1. Remark: $\Gamma(x+1) = x!$
2. **If $Y \sim Bin(n,\theta)$ and $\theta \sim \beta(1,1)$ (uniform), then the posterior $\theta|y \sim \beta(1+y, 1+n-y)$**
   1. a beta prior distribution and a binomial sampling model lead to a beta posterior distribution. This is called **conjugacy**
   2. Expectation: $\frac{a+y}{a+b+n}$ = $\frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n}$ = $\frac{a+b}{a+b+n}(prior \quad expectation) + \frac{n}{a+b+n}(data \quad average)$
   3. Mode: $\frac{a+y-1}{a+b+n-2}$
   4. Variance: $\frac{E(\theta|y)E(1-\theta|y)}{a+b+n+1}$
   5. Predictive Distribution: $\Pr(\tilde{Y}=0 \mid y_1,\dots,y_n) = 1-\mathbb{E}[\theta \mid y_1,\dots,y_n] = \frac{b+\sum_{i=1}^n(1-y_i)}{a+b+n}.$
      1. The predictive distribution doesn't depend on unknown parameters, and depends on observed data
3. Confidence Intervals (Frequentist) vs Confidence Regions (Bayes)
   1. An interval based on the observed data has 95% bayesian coverage if $\Pr\left(l(y) < \theta < u(y) \mid Y = y\right) = 0.95$
   2. A random interval has 95% frequentist coverage if before the data is gathered $\Pr\left(l(Y) < \theta < u(Y) \mid \theta\right) = 0.95$

The Poisson Model

1. The poisson family has a mean-variance relationship because if one Poisson distribution has a larger mean than another, it will have a larger variance as well.
2. The joint pdf is $\Pr(Y_1 = y_1, \dots, Y_n = y_n \mid \theta) = \prod_{i=1}^n \frac{1}{y_i!}\theta^{y_i}e^{-\theta} = c(y_1,\dots,y_n)\theta^{\sum y_i}e^{-n\theta}.$
3. Same case as IID binary case, the sum of the data contains all information and is a sufficient statistic with posterior following $Poisson(n,\theta)$
4. **Conjugate prior: prior in the same family**, ie Binomial and beta, normal and exponential, poisson and gamma
5. **If $Y \sim Poisson(\theta)$ and $\theta \sim \Gamma(a,b)$ (uniform), then the posterior $\theta|y \sim \Gamma(a+\sum Y_i, b+n)$**
6. The posterior expectation: $\mathbb{E}[\theta \mid y_1,\dots,y_n] = \frac{a+\sum y_i}{b+n} = \frac{b}{b+n}\cdot\frac{a}{b} + \frac{n}{b+n}\cdot\frac{\sum y_i}{n}$
7. The predictive distribution: $\frac{(b+n)^{a+\sum y_i}}{\Gamma(\tilde{y}+1)\Gamma(a+\sum y_i)}\int_0^\infty \theta^{a+\sum y_i+\tilde{y}-1}e^{-(b+n+1)\theta}\,d\theta$
8. Birthrate example

    1. Suppose we are sampling two poisson variables:
        1. Less than bachelors: n=111, sum=217, mean=1.95
        2. bachelors or higher: n=44, sum=66, mean=1.5
    2. Assuming we choose a prior of $\Gamma(2,1)$, the posteriors are $\Gamma(2+217,1+111)=\Gamma(219,112)$ and $\Gamma(2+66,1+44)=\Gamma(68,45)$
9. Note that distringuishing between comparing thetas and Ys are important, because **strong evidence of a difference between two populations does not mean that the difference itself is large.**
10. Exponential families and conjugate priors
    1. Binomial and poisson are all members of single parameter exponential family, which exponential takes the general form $p(y|\phi)=h(y)c(\phi)e^{\phi t(y)}$ where ϕ is the unknown parameter and t(y) is the sufficient statistics
    2. Binomial Model
        1. **the exponential representation of a binomial model is a single paramter binary** $p(y|\theta)=e^{\phi y}(1+e^{\phi})^{-1}$**, where** $\phi=log(\theta(1-\theta))$ **is the log-odds.**
        2. **The conjugate prior for** ϕ **is** $(1+e^{\phi})^{-n_0}e^{n_0 t_0 \phi}$ **where n0 is the prior sample size and t0 is the prior guess**
        3. **we can get a weakly informative prior by setting** $n_0=t_0=1$ **which results in a beta(0.5,0.5) prior, called a Jeffrey Prior**
    3. Poisson Model
        1. The exponential representation is where $t(y)=y$, $\phi=log\theta$, $c(\phi)=exp(e^{-\phi})$
        2. **The conjugate prior is** $p(\phi \mid n_0,t_0)=\exp(n_0e^{-\phi})\,e^{n_0t_0y}$
        3. **A weakly informative prior sets t0 to the prior expectation of Y and n0=1 which results in a gamma(t0+sum y, 1+n)**

Monte Carlo

1. We usually want to summarize all aspects of posterior distributions. obtaining exact values can be difficult or impossible, so we can generate random samples and average outcomes
2. Letting S be the number of loops we do for MC (since it is based on LLN):
    1. Monte carlo sample mean: $\bar{\theta}=\frac{1}{S}\sum_{s=1}^{S}\theta^{(s)}$
    2. Monte carlo Variance: $\hat{\sigma}^2=\frac{1}{S-1}\sum(\theta^{(s)}-\bar{\theta})^2$
    3. Monte Carlo Standard Error: $\sqrt{\frac{\hat{\sigma}^2}{S}}$
    4. Monte Carlo interval: $\hat{\theta}\pm 2\sqrt{\frac{\hat{\sigma}^2}{S}}$

Normal Model

1. The most common probability model for data analysis is the normal model

2. The normal distribution is symmatric about the mean and mean=mode=median
3. joint sampling desntiy: $p(y|\theta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\sum\left(\frac{y_i-\theta}{\sigma}\right)^2\right\}$
4. Inference for a two parameter model can be broken into two one parameter models
    1. Making inference for mean when variance is known and specify a conjugate distribution
    2. for any conditional prior, the posterior is $p(\theta \mid y_1, \ldots, y_n, \sigma^2) \propto p(\theta \mid \sigma^2) \times \exp\left(-\frac{1}{2\sigma^2}\sum(y_i - \theta)^2\right)$
    3. **if the conditional prior is conjugate, it must include the quadratic term**
    4. **for the conjugate, we refer to the mean μ and standard deviation τ^2**
        1. posterior variance: $\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$
        2. posterior mean: $\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$
5. Now combining the posterior parameters (μ_n, τ_n^2) and the prior parameters (μ_0, τ_0^2):
    1. the posterior precision: $\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$,
    2. the posterior mean: $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$.
6. Inverse variance is referred to as precision, as the quantity of information on an additive scale.
7. **The predictive distribution is $\tilde{Y} \mid \sigma^2, y_1, \ldots, y_n \sim \mathrm{N}(\mu_n, \tau_n^2 + \sigma^2)$**
8. Example: Midge wing lenth
    1. midges are small wings, and scientists say the average length of a wing is 1.9 (so we set $\mu_0 = 1.9$). we also know that wing lengths have to be positive ($\theta > 0$). Knowing that, we can choose τ^2 such that $\mu - 2\tau > 0$ so tau=0.95 via 1.9/2. The observations (1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08), giving $\bar{y} = 1.804$.
    2. Now using the posterior formulae:
        1. Mean: $\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1.11 \times 1.9 + \frac{9}{\sigma^2} \times 1.804}{1.11 + \frac{9}{\sigma^2}}$
        2. Variance: $\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1}{1.11 + \frac{9}{\sigma^2}}$. Assuming $\sigma^2 = s^2 = 0.017$, we have N(1.805,0.002)
9. Joint inference for mean and variance
    1. For the variance, we need a distribution that supports $(0, \infty)$. One such distribution is gamma (like we used in poisson), but that family isn't conjugate for normal but is a conjugate when using $\frac{1}{\sigma^2}$ so we say σ^2 has an Inverse-Gamma distribution
        1. $\frac{1}{\sigma^2} \sim \mathrm{Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2\right)$
        2. Under this parameterization, where σ_0^2 is sample variance and ν_0 is sample size:
            1. Expectation: $\sigma_0^2 \cdot \frac{\nu_0/2}{\nu_0/2-1}$
            2. Mode: $= \sigma_0^2 \cdot \frac{\nu_0/2}{\nu_0/2+1}$

            3. Variance decreases
        3. prior sample variance: $\nu_n = \nu_0 + n$
        4. $\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2\right]$; aka “posterior sum of squares equals prior sum of squares plus data sum of squares."
10. Returning to the midge example
    1. The sample mean and variance are 1.804 and 0.0169 (sd=0.130). from these and prior parameters, we can now find:
        1. Posterior Mean: $\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n} = \frac{1.9 + 9 \times 1.804}{1+9} = 1.814$
        2. Posterior Variance: $\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y} - \mu_0)^2\right] = \frac{0.010 + 0.135 + 0.008}{10} = 0.015$
        3. Therefore the mean is normal and precision is gamma (variance is inverse gamma)
            1. $\{\theta \mid y_1, \ldots, y_n, \sigma^2\} \sim \text{normal}(1.814, \sigma^2/10),$
            2. $\left\{\frac{1}{\sigma^2} \middle| y_1, \ldots, y_n\right\} \sim \text{gamma}\left(\frac{10}{2}, \frac{10 \times 0.015}{2}\right).$
11. Monte Carlo Sampling
    1. For population estimation, we need marginal distribution data but we only have it for conditional ones. We sample independent pairs of the mean and variance
    2. The marginal distiribution for theta can be obtained through a closed form: $t(\theta) = \frac{\theta - \mu_n}{\frac{\sigma_n}{\sqrt{\kappa_n}}}$ given $\bar{y}$ and s^2 follows a t-distribution with $\nu_0 + n$ degrees of freedom.
    3. If κ_0 and ν_0 are small, then it follows a t-distribution with n-1 degrees of freedom, so as these values get closer to zero, then $\mu_n = \bar{y}$ and $\sigma_n^2 = \frac{n-1}{n}s^2$
12. Bias, Variance, and MSE
    1. Bias refers to how close the center of mass of the sampling distribution of an estimator is to the true value. An unbiased estimator is an estimator with zero bias, which sounds desirable.
        1. Unbiased: $\mathbb{E}[\hat{\theta}_e \mid \theta = \theta_0] = \theta_0$
        2. Biased: $\mathbb{E}[\hat{\theta}_b \mid \theta = \theta_0] = w\theta_0 + (1-w)\mu_0$ if theta and mu are not equal
    2. To evaluate how close the parameter is to the true value, we use MSE:
        1. $\text{MSE}[\hat{\theta} \mid \theta_0] = \mathbb{E}[(\hat{\theta} - \theta_0)^2 \mid \theta_0]$
        2. $\text{MSE}[\hat{\theta} \mid \theta_0] = \text{Var}[\hat{\theta} \mid \theta_0] + \text{Bias}^2[\hat{\theta} \mid \theta_0]$
13. Prior specification based on expectations
    1. recall for exponential families that $p(y \mid \boldsymbol{\phi}) = h(y)\, c(\boldsymbol{\phi}) \exp\left\{\boldsymbol{\phi}^\top t(y)\right\},$
    2. The normal model is a two dimensional exponential model with
        1. $t(y) = (y, y^2)$
        2. $\boldsymbol{\phi} = \left(\frac{\theta}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$
        3. $c(\boldsymbol{\phi}) = |\phi_2|^{1/2} \exp\left\{\frac{\phi_1^2}{2\phi_2}\right\}$

3. The joint prior for mu and sigma are:
    1. $\theta \mid \sigma^2 \sim \text{normal}(\mu_0, \sigma^2/n_0)$
    2. $\sigma^2 \sim \text{inverse-gamma}\left(\frac{n_0+3}{2}, \frac{(n_0+1)\sigma_0^2}{2}\right)$

Gibbs Sampler

1. Because calculations from multiparameter models is complicated, posterior approximation can be made with the Gibbs sampler, an iterative algorithm that constructs a dependent sequence of parameter values whose distribution converges to the target joint posterior distribution.
2. **A semiconjugate prior distribution**
    1. $\theta \sim \text{normal}(\mu_0, \tau_0^2)$
    2. $\frac{1}{\sigma^2} \sim \text{gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$
    3. true for where tau is proportional to sigma
3. Discrete Approximations
    1. Recall the posterior distribution is the joint distribution divided by likelihood
    2. The joint is $= \text{N}(\theta, \mu_0, \tau_0) \times \text{Gamma}(\tilde{\sigma}^2, \nu_0/2, \nu_0\sigma_0^2/2) \times \prod_{i=1}^n \text{N}(y_i; \theta, 1/\sqrt{\tilde{\sigma}^2})$
    3. Evaluation of this two-parameter posterior distribution at 100 values of each parameter required a grid of size 100×100 = 10000
    4. In general, to construct a similarly fine approximation for a p-dimensional posterior distribution we would need a p-dimensional grid containing 100^P posterior probabilities. This means that discrete approximations will only be feasible for densities having a small number of parameters.
4. Sampling from Conditional Distributions
    1. **If precision and mean are independent, then we can sample sigma directly**. two conditionals can be used to sample a joint distribution, but we need to know the starting point for sigma
5. Gibbs Sampling
    1. finding $p(\theta)$ and $p(\sigma^2)$ are the full conditionals. **Given a sample from an iterated mean and variance, we can generate dependent sequences via Gibbs:**
        1. $\text{sample } \theta^{(s+1)} \sim p(\theta \mid \tilde{\sigma}^{2\,(s)}, y_1, \ldots, y_n)$
        2. $\text{sample } \tilde{\sigma}^{2\,(s+1)} \sim p(\tilde{\sigma}^2 \mid \theta^{(s+1)}, y_1, \ldots, y_n)$
        3. $\text{let } \phi^{(s+1)} = \left\{\theta^{(s+1)}, \tilde{\sigma}^{2\,(s+1)}\right\}$
    2. This uses the identity $ns_n^2(\theta) = \sum_{i=1}^n (y_i - \theta)^2 = (n-1)s^2 + n(\bar{y} - \theta)^2$ since s^2 and $\bar{y}$ do not change across iterations
6. **General Properties of a gibbs sampler**
    1. A vector of dependent sequences is generated, called a markov chain. The sampling distribution approaches the target distribution as more iterative samples are taken,

    resulting in Markov Chain Monte Carlo (MCMC)

    2. Doing bayesian data analysis with monte carlo involves a mix of sampling procedures and probability distributions. The necessary ingredients are:
        1. **Model Specification**: A collection of probability distributions $p(\mathbf{y} \mid \boldsymbol{\phi}) : \boldsymbol{\phi} \in \Phi$ which should represent the sampling distribution of your data for some value of $\boldsymbol{\phi} \in \Phi$.
        2. **Prior Specification**: A probability distribution $p(\boldsymbol{\phi})$, ideally representing someone's prior information about which parameter values are likely to describe the sampling distribution.
        3. **Posterior Summary**: A description of the posterior distribution $p(\boldsymbol{\phi} \mid \mathbf{y})$, done in terms of particular quantities of interest such as posterior means, medians, modes, predictive probabilities, and confidence regions.
    3. **We look at the posterior with MCMC, we are not generating more information using models**

7. Introduction to MCMC Diagnostics
    1. The purpose of MCMC or MC is to get an empirical average that approximates an expected value g(x) under a posterior p(x)
        1. $\frac{1}{S}\sum_{s=1}^{S} g(\boldsymbol{\phi}^{(s)}) \approx \int g(\boldsymbol{\phi})\, p(\boldsymbol{\phi})\, d\boldsymbol{\phi},$
        2. Monte Carlo generates independent samples from the target distribution
        3. Markov Chain Monte Carlo is different because of convergence: $\lim_{s\to\infty} \Pr(\boldsymbol{\phi}^{(s)} \in A) = \int_A p(\boldsymbol{\phi})\, d\boldsymbol{\phi}.$
    2. In some cases, MC will show convergence but MCMC will produce stickiness, called **autocorrelation**, or correlation between values of the chain
        1. Though Gibbs sampler eventually reproduces the distribution, eventually can take a very long time
    3. One thing to check for is **stationarity**, or that samples taken in one part of the chain have a similar distribution to samples taken in other parts.
    4. how quickly the particle moves around the parameter space, which is sometimes called the speed of **mixing**.
        1. An independent monte carlo sampler has perfect mixing: zero autocorrelation and can jump between different points
        2. How does the correlation of MCMC affect posterior approximation?
            1. If ϕ-values are independent MC samples from $p(\phi)$ then the variance of $\bar{\phi} = \frac{\sum \phi^s}{S}$ is $\mathrm{Var}_M C(\bar{\phi}) = \frac{\mathrm{Var}(\phi)}{S} = \frac{\int \phi^2 p(\phi) d\phi - \phi_0^2}{S}$
                1. recall that the square root of this is the MC standard error, a measure of how well we expect the approximation to perform
            2. MCMC variance is MC variance with a term that tells us how samples are correlated. This term is generally positive so we expect MCSE < MCMCSE .

        The higher the autocorrelation in the chain, the larger the MCMC variance and the worse the approximation is.

    3. We can use the lag-t autocorrelation function to estimates the correlation between elements t lags apart: $\operatorname{acf}_t(\boldsymbol{\phi}) = \frac{\frac{1}{S-t}\sum_{s=1}^{S-t}(\phi_s-\bar{\phi})(\phi_{s+t}-\bar{\phi})}{\frac{1}{S-1}\sum_{s=1}^{S}(\phi_s-\bar{\phi})^2}$
        1. The higher the autocorrelation, the more samples we need. we need to find an effective sample size

5. MCMC diagnostics for semi-conjugate normal analysis
    1. Gelman and Rubin, Geweke, Raferty and Lewis

Multivariate Normal Model

1. an MVN is a normal model with a bunch of vectors shoved into it, and we have to estimate means, variances, and covariances
2. recall that if our data is IID univariate normal, then a convenient conjugate prior is also univariate normal. a convenient prior distribution for the multivariate mean is a multivariate normal: $p(\boldsymbol{\theta}|\vec{y}, \Sigma) = \mathrm{MVN}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$
    1. $\boldsymbol{\theta}$ is the multivariate mean vector
    2. $\vec{y}$ is the vector of data
    3. Σ is the VCV matrix
    4. $\mathrm{MVN}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$ is a multivariate normal distribution with mean $\boldsymbol{\mu}_0$ and variance $\boldsymbol{\Lambda}_0$
3. **The prior and the full conditional are both multivariate normal**
    1. posterior precision, or inverse variance, is the sum of the prior precision and the data precision, just as in the univariate normal case.
        1. $\operatorname{Cov}[\boldsymbol{\theta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma] = \Lambda_n = \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)^{-1}$
    2. the posterior expectation is a weighted average of the prior expectation and the sample mean.
        1. $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma] = \mu_n = \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)^{-1}\left(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\mathbf{y}}\right)$
4. The Inverse Wishart Distribution
    1. like how in the univariate case variance has to be positive, in the MVN case, VCV has to be positive definite (no negative values, all values be between -1 and 1)
    2. Think of the wishart distribution as multivariate gamma (if x is a mean zero univariate normal RV, then x^2 is a gamma RV)
    3. **To sample a covariance matrix Σ from Inverse Wishart, follow these steps**
        1. Sample $\mathbf{z}_1, \ldots, \mathbf{z}_{\nu_0} \sim \text{i.i.d. multivariate normal}(\mathbf{0}, \mathbf{S}_0^{-1})$;
        2. Calculate $\mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^{\nu_0} \mathbf{z}_i\mathbf{z}_i^T$;
        3. Set $\Sigma = (\mathbf{Z}^T\mathbf{Z})^{-1}$.
    4. Precision matrix (wishart) as covariance matrix (inverse wishart)
5. **Gibbs Sampling for Mean and Covariance**

    1. The mean is MVN and covariance is Inverse wishart, we can gibbs sample using
        1. Sample $\boldsymbol{\theta}^{(s+1)}$ from its full conditional distribution:
            1. compute $\boldsymbol{\mu}_n$ and $\mathbf{\Lambda}_n$ from $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\mathbf{\Sigma}^{(s)}$;
            2. sample $\boldsymbol{\theta}^{(s+1)} \sim \text{MVN}(\boldsymbol{\mu}_n, \mathbf{\Lambda}_n)$.
        2. Sample $\mathbf{\Sigma}^{(s+1)}$ from its full conditional distribution:
            1. compute $\mathbf{S}_n$ from $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\boldsymbol{\theta}^{(s+1)}$;
            2. sample $\mathbf{\Sigma}^{(s+1)} \sim \text{inverse-Wishart}(\nu_0 + n, \mathbf{S}_n^{-1})$.

6. Complete Case Analysis
    1. Many softwares throw out missing cases or impute with a known mean or value. Throwing away is bad because information loss and the second is statistically incorrect
    2. Assume the data is missing at random, we can get missing values by integrating over the missing data to obtain the marginal probability of the observed data.
    3. Construct a matrix of indicator variables with 1=complete, 0=incomplete. We can use a gibbs sampler with one more step:
        1. $\text{sampling } \boldsymbol{\theta}^{(s+1)} \text{ from } p\left(\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}^{(s)}, \mathbf{\Sigma}^{(s)}\right)$
        2. $\text{sampling } \mathbf{\Sigma}^{(s+1)} \text{ from } p\left(\mathbf{\Sigma} \mid \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{miss}}^{(s)}, \boldsymbol{\theta}^{(s+1)}\right)$
        3. $\text{sampling } \mathbf{Y}_{\text{miss}}^{(s+1)} \text{ from } p\left(\mathbf{Y}_{\text{miss}} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}^{(s+1)}, \mathbf{\Sigma}^{(s+1)}\right)$

Group Comparisons and Hierarchy

1. Comparing two groups
    1. Consider a classical ANOVA model with decision rules
        1. p>0.05:
            1. conclude that $\theta_1 = \theta_2$;
            2. use the estimates $\hat{\theta}_1 = \hat{\theta}2 = \dfrac{\sum y_{i,1} + \sum y_{i,2}}{n_1 + n_2}$.
        2. p<0.05:
            1. conclude that $\theta_1 \neq \theta_2$;
            2. use the estimates $\hat{\theta}_1 = \bar{y}_1$, $\hat{\theta}_2 = \bar{y}_2$.
2. Bayesian ANOVA approach
    1. Define the model for two groups as:
        1. $Y_{i,1} = \mu + \delta + \epsilon_{i,1}$. $Y_{i,2} = \mu - \delta + \epsilon_{i,2}$
        2. $\{\epsilon_{i,j}\} \sim \text{i.i.d. normal}(0, \sigma^2)$
    2. Convenient conjugate priors are
        1. $\{\mu \mid \mathbf{y}_1, \mathbf{y}_2, \delta, \sigma^2\} \sim \text{normal}(\mu_n, \gamma_n^2)$, where
            1. $\mu_n = \gamma_n^2 \times \left[\mu_0/\gamma_0^2 + \sum_{i=1}^{n_1}(y_{i,1} - \delta)/\sigma^2 + \sum_{i=1}^{n_2}(y_{i,2} + \delta)/\sigma^2\right]$
            2. $\gamma_n^2 = \left[1/\gamma_0^2 + (n_1 + n_2)/\sigma^2\right]^{-1}$

        2. $\{\sigma^2 \mid \mathbf{y}_1, \mathbf{y}_2, \mu, \delta\} \sim \text{inverse-gamma}(\nu_n/2, \nu_n\sigma_n^2/2)$, where
            1. $\nu_n = \nu_0 + n_1 + n_2$
            2. $\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + \sum(y_{i,1} - [\mu + \delta])^2 + \sum(y_{i,2} - [\mu - \delta])^2$
        3. $\{\delta \mid \mathbf{y}_1, \mathbf{y}_2, \mu, \sigma^2\} \sim \text{normal}(\delta_n, \tau_n^2)$, where
            1. $\delta_n = \tau_n^2 \times \left[\delta_0/\tau_0^2 + \sum(y_{i,1} - \mu)/\sigma^2 - \sum(y_{i,2} - \mu)/\sigma^2\right]$
            2. $\tau_n^2 = \left[1/\tau_0^2 + (n_1 + n_2)/\sigma^2\right]^{-1}$

3. Comparing multiple groups via hierarchal models
    1. Multilevel data segregates by groups and units within groups
    2. A sequence of random variables is exchangeable if the PDF for our sequence is equal for all permutations. if exchangeability holds for all values of n, then the random variables can be thought of as independent samples from a population described by some fixed but unknown population feature
    3. In particular, the data will be used to estimate the within-and between-group sampling distributions $p(y|\varphi)$ and $p(\varphi|\psi)$, whereas the prior distribution $p(\psi)$ is not estimated from the data.
4. The Hierarchial Normal Model
    1. heterogeneity of means can be described using both within and between group models that are both normal
        1. Within: $\phi_j = \{\theta_j, \sigma^2\}, \quad p(y \mid \phi_j) = \text{normal}(\theta_j, \sigma^2)$
        2. Between: $\psi = \{\mu, \tau^2\}, \quad p(\theta_j \mid \psi) = \text{normal}(\mu, \tau^2)$
    2. by this case $p(\phi|\psi)$ only describes heterogeneity in means and not variances (because constant variance is an assumption)
    3. the full conditionals are the same for the normal model
    4. We need to know three unknown parameters: μ, σ^2, τ^2
        1. $1/\sigma^2 \sim \text{gamma}(\nu_0/2,\ \nu_0\sigma_0^2/2)$
        2. $1/\tau^2 \sim \text{gamma}(\eta_0/2,\ \eta_0\tau_0^2/2)$
        3. $\mu \sim \text{normal}(\mu_0,\ \gamma_0^2)$
    5. Posterior Inference
        1. The Gibbs sampler proceeds by iteratively sampling each parameter from its full conditional distribution. (The normal-inverse gamma thingy)
            1. $\{\mu \mid \theta_1, \ldots, \theta_m, \tau^2\} \sim \text{normal}\left(\frac{m\bar{\theta}/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \left[m/\tau^2 + 1/\gamma_0^2\right]^{-1}\right)$
            2. $\left\{1/\tau^2 \mid \theta_1, \ldots, \theta_m, \mu\right\} \sim \text{gamma}\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum(\theta_j - \mu)^2}{2}\right)$
    6. Assessing Stationarity of MCMC
        1. since iterations are large, plots would likely look jumbled. We can plot every 100 samples instead or do box plots of every 10% of the sample. Afterwards we can check ACF and MCMCSE
    7. Posterior Summaries and shrinkage

1. Expectation of parameter (shrinkage): $\mathbb{E}[\theta_j \mid \mathbf{y}_j, \mu, \tau, \sigma] = \frac{\bar{y}_j n_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}$
    1. groups with low sample sizes get shrunk the most, high sample sizes dont budge
    2. This makes sense: The larger the sample size for a group, the more information we have for that group and the less information we need to “borrow” from the rest of the population.

8. Hierarchial modeling of means and variances
    1. if we assume variance to vary across groups, we need to sample priors for ν_0 and σ_0^2 and obtain the full conditionals
        1. A conjugate class is the gamma distributions: if $p(\sigma_0^2) \sim \Gamma(a, b)$ then $p(\sigma_0^2 \mid \sigma_1^2, \ldots, \sigma_m^2, \nu_0) = \text{dgamma}\left(a + \frac{1}{2} m\nu_0,\ b + \frac{1}{2} \sum_{j=1}^{m} \left(\frac{1}{\sigma_j^2}\right)\right)$
        2. A simple conjugate prior doesn't exist but if we assume variance is a whole number, then we can assume a geometric distribution $p(\nu_0) \propto e^{\alpha \nu_0}$

Linear Regression

1. Variable selection in bayes comes naturally: Any collection of models having different sets of regressors can be compared via their Bayes factors.
    1. When the number of possible regressors is small, this allows us to assign a posterior probability to each regressionmodel.
    2. When the number of regressors is large, the space of models can be explored with a Gibbs sampling algorithm.
2. Our conditional model for target y given covariates x is given by: $\{\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2\} \sim \text{multivariate normal}(\mathbf{X}\boldsymbol{\beta},\ \sigma^2 \mathbf{I})$ with I being the identity matrix
    1. Minimizing the SSR results in the standard LS estimate: $\beta_{OLS} = (X^T X)^{-1} X^T y$ with sampling variance $\text{Var}(\beta_{OLS}) = (X^T X)^{-1} \sigma_e^2$
3. Semiconjugate Priors
    1. Since y is MVN, it implies that an MVN prior distribution for β is conjugate
    2. The semiconjugate for σ^2 is IG: $\{\sigma^2 \mid \mathbf{y}, \mathbf{X}, \boldsymbol{\beta}\} \sim \text{inverse-gamma}\left(\frac{\nu_0 + n}{2},\ \frac{\nu_0 \sigma_0^2 + \text{SSR}(\boldsymbol{\beta})}{2}\right)$
    3. The gibbs algorithm is:
        1. updating $\boldsymbol{\beta}$:
            1. compute $\mathbf{V} = \text{Var}[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \sigma^{2(s)}]$ and $\mathbf{m} = \mathbb{E}[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \sigma^{2(s)}]$
            2. sample $\boldsymbol{\beta}^{(s+1)} \sim \text{multivariate normal}(\mathbf{m}, \mathbf{V})$
        2. updating σ^2:
            1. compute $\text{SSR}(\boldsymbol{\beta}^{(s+1)})$
            2. sample $\sigma^{2(s+1)} \sim \text{inverse-gamma}\left(\frac{\nu_0 + n}{2},\ \frac{\nu_0 \sigma_0^2 + \text{SSR}(\boldsymbol{\beta}^{(s+1)})}{2}\right)$
4. Default and weakly informative Priors

   1. When we increase predictors magrinally, the number of parameters increases quadratically
   2. Unit information prior
      1. A unit information prior is one that contains the same amount of information as that would be contained in only a single observation.
         1. For our OLS estimate, the precision is $\frac{(X^TX)^{-1}}{\sigma^2}$ then the amount of information in one observation should be "one n-th as much" ie $\frac{(X^TX)^{-1}}{n\sigma^2}$
         2. Further suggested is to set $\beta_0 = \beta_{OLS}$ (centers the prior around the estimate)
         3. Similarily a prior for σ^2 should be centered around $\hat{\sigma}^2_{\text{ols}}$ by taking $\nu_0 = 1$ and $\sigma^2_0 = \hat{\sigma}^2_{\text{ols}}$.
   3. Another aspect we need to consider for bayesian estimation is parameters have to be scale invariant
      1. by this principle, the posterior distributions of β and $H\beta$ should be the same. This occurs when our intercept is the zero vector and our covariance matrix is $k(X^TX)^{-1}$.
      2. A popular specification is to relate the k value to error variance such that $k = g\sigma^2$ for some positive g. This is called a g-prior and reproduces the following:
         1. Expectation: $\frac{g}{g+1}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$
         2. Variance: $\frac{g}{g+1}\,\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$
      3. Parameter estimation under the g-prior is simplified as well: It turns out that, under this prior distribution, $p(\sigma^2 \mid \mathbf{y}, \mathbf{X})$ is an inverse-gamma distribution, which means that we can directly sample $(\sigma^2, \boldsymbol{\beta})$ from their posterior distribution by first sampling from $p(\sigma^2 \mid \mathbf{y}, \mathbf{X})$ and then from $p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}, \mathbf{X})$.
      4. The SSR of g-prior converges to the SSR of an OLS estimate as n goes to infinity
      5. Since we can sample from MVN and IG, we don't need a gibbs sampler and can use monte carlo instead
         1. sample $1/\sigma^2 \sim \text{gamma}\left(\frac{\nu_0+n}{2},\ \frac{\nu_0\sigma_0^2+\text{SSR}_g}{2}\right)$
         2. sample $\boldsymbol{\beta} \sim \text{multivariate normal}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\text{ols}},\ \frac{g}{g+1}\sigma^2\left[\mathbf{X}^T\mathbf{X}\right]^{-1}\right)$

5. Model Selection
   1. We can use classical methods like CV, stepwise, etc and the backwards elimination algorithm
      1. Obtain the estimator $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and its t-statistics.
         1. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$,
            1. find the regressor j_{min} having the smallest value of $|t_j|$ and remove column j_{min} from $\mathbf{X}$.
            2. return to step 1.
         2. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

   2. Bayesian model selection
      1. We can define a dummy variable z that controls for one slope at a time and define posterior odds as a product of prior odds x bayes factor
   3. Gibbs Sampling and Model Averaging
      1. Use the following algorithm to average models
         1. Set $\mathbf{z} = \mathbf{z}^{(s)}$
         2. For $j \in \{1, \ldots, p\}$ in random order, replace z_j with a sample from $p(z_j \mid \mathbf{z}_{-j}, \mathbf{y}, \mathbf{X})$
         3. Set $\mathbf{z}^{(s+1)} = \mathbf{z}$
         4. Sample $\sigma^{2(s+1)} \sim p(\sigma^2 \mid \mathbf{z}^{(s+1)}, \mathbf{y}, \mathbf{X})$
         5. Sample $\boldsymbol{\beta}^{(s+1)} \sim p(\boldsymbol{\beta} \mid \mathbf{z}^{(s+1)}, \sigma^{2(s+1)}, \mathbf{y}, \mathbf{X})$

Nonconjugate Priors and Metropolis-Hastings

1. The metropolis Algorithm
   1. Assume we have a generic sampling model $Y \sim p(y|\theta)$ and a prior $p(\theta)$. Recall that integration sucks but we can get by by samling different θ's from $p(\theta|y)$ and obtain monte carlo approximations. We need a large enough collection of theta values who's empirical distribution approximates the posterior
   2. We want to add a new value θ^{s+1} so we choose a number θ^* thats close to θ^2. Should we include θ^* in the set?
      1. $p(\theta^* \mid y) > p(\theta^{(s)} \mid y)$ include in the set
      2. $p(\theta^* \mid y) < p(\theta^{(s)} \mid y)$ exclude in the set
   3. We can compute a value r to do this comparison: $r = \dfrac{p(y \mid \theta^*)\, p(\theta^*)}{p(y \mid \theta^{(s)})\, p(\theta^{(s)})}$ (aka MH Acceptance Ratio)
      1. if R>1
         1. Intuition: Since $\theta^{(s)}$ is already in our set, we should include θ^* as it has a higher probability than $\theta^{(s)}$.
         2. Procedure: Accept θ^* into our set, i.e. set $\theta^{(s+1)} = \theta^*$.
      2. if R<1
         1. Intuition: The relative frequency of θ-values in our set equal to θ^* compared to those equal to $\theta^{(s)}$ should be $p(\theta^* \mid y)/p(\theta^{(s)} \mid y) = r$. This means that for every instance of $\theta^{(s)}$, we should have only a "fraction" of an instance of a θ^* value.
         2. Procedure: Set $\theta^{(s+1)}$ equal to either θ^* or $\theta^{(s)}$, with probability r and $1 - r$ respectively.
      3. This value comes from a symmetric proposal distribution J() usually:
         1. $J(\theta^* \mid \theta^{(s)}) = \text{uniform}(\theta^{(s)} - \delta,\ \theta^{(s)} + \delta)$

         2. $J(\theta^* \mid \theta^{(s)}) = \text{normal}(\theta^{(s)},\ \delta^2)$
         3. The δ value is usually chosen to reflect lag-1 autocorrelation (1/32, 1/2, 2, 32, 64). **The best value tends to occur towards the middle of the set and not at the extremes**
   4. **METROPOLIS ALGORITHM**
      1. Sample $\theta^* \sim J(\theta \mid \theta^{(s)})$
      2. Compute the acceptance ratio
      3. Let $\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability } \min(r,1) \\ \theta^{(s)} & \text{with probability } 1 - \min(r,1) \end{cases}$
         1. Step 3 can be accomplished by sampling $u \sim \text{uniform}(0,1)$ and setting $\theta^{(s+1)} = \theta^*$ if $u < r$, and setting $\theta^{(s+1)} = \theta^{(s)}$ otherwise.
   5. Computing the acceptance ratio directly can be numerically unstable, so we can take the log of it instead. The proposal is accepted if log u < log r where u is a sample from a uniform distribution (0,1)
   6. the algorithm generates a sequence of dependent theta values and is memoryless (t depends only on t-1)
   7. **The standard practice for MCMC via Gibbs or MH**
      1. Run algorithm until some iteration B for which it looks like the Markov chain has achieved stationarity. (burn in period)
      2. Run the algorithm S more times, generating $\{\theta^{(B+1)}, \ldots, \theta^{(B+S)}\}$.
      3. Discard $\{\theta^{(1)}, \ldots, \theta^{(B)}\}$ and use the empirical distribution of $\{\theta^{(B+1)}, \ldots, \theta^{(B+S)}\}$ to approximate $p(\theta \mid y)$.
2. Gibbs sampler and Metropolis algorithms are special cases of the metropolis hastings algorithm
3. **Consider a simple example: we want to sample $p_0(u,v)$ from two random variables U and V (think the normal model where u=mean, v=variance).**
   1. **GIBBS** proceeds by iteratively sampling U and V from their conditionals.
      1. Given $x^{(s)} = (u^{(s)}, v^{(s)})$ we can sample $x^{(s+1)}$ by
         1. update U: sample $u^{(s+1)} \sim p_0(u \mid v^{(s)})$
         2. update V: sample $v^{(s+1)} \sim p_0(v \mid u^{(s+1)})$ (these two steps are interchageable)
   2. **METROPOLIS** proposes changes to X=(U,V) and then accepts or rejects those changes based on p_0 (or we could propose then A/R). **J is symmetric proposal distribution, one for each of U and V**
      1. update U:
         1. sample $u^* \sim J_u(u \mid u^{(s)})$
         2. compute $r = \dfrac{p_0(u^*, v^{(s)})}{p_0(u^{(s)}, v^{(s)})}$

         3. set $u^{(s+1)}$ to u^* or $u^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1-r)$ respectively
      2. update V:
         1. sample $v^* \sim J_v(v \mid v^{(s)})$
         2. compute $r = \dfrac{p_0(u^{(s+1)}, v^*)}{p_0(u^{(s+1)}, v^{(s)})}$
         3. set $v^{(s+1)}$ to v^* or $v^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1-r)$ respectively
   3. **METROPOLIS-HASTINGS** generalizes both of these approaches by allowing arbitrary proposal distributions. The proposal distributions can be symmetric around the current values, full conditional distributions, or something else entirely. **J is not required to be symmetric. Only requirement is it does not depend on U or V**.
      1. update U:
         1. sample $u^* \sim J_u(u \mid u^{(s)}, v^{(s)})$
         2. compute the acceptance ratio $r = \dfrac{p_0(u^*, v^{(s)})}{p_0(u^{(s)}, v^{(s)})} \times \dfrac{J_u(u^{(s)} \mid u^*, v^{(s)})}{J_u(u^* \mid u^{(s)}, v^{(s)})}$
         3. set $u^{(s+1)}$ to u^* or $u^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1-r)$
      2. update V:
         1. sample $v^* \sim J_v(v \mid u^{(s+1)}, v^{(s)})$
         2. compute the acceptance ratio $r = \dfrac{p_0(u^{(s+1)}, v^*)}{p_0(u^{(s+1)}, v^{(s)})} \times \dfrac{J_v(v^{(s)} \mid u^{(s+1)}, v^*)}{J_v(v^* \mid u^{(s+1)}, v^{(s)})}$
         3. set $v^{(s+1)}$ to v^* or $v^{(s)}$ with probability $\min(1, r)$ and $\max(0, 1-r)$
4. Differences recap
   1. The Metropolis-Hastings algorithm looks a lot like the Metropolis algorithm, except that the acceptance ratio contains an extra factor, (a correction)
      1. "**Correction factor**:" If a value u^* is much more likely to be proposed than the current value $u^{(s)}$, then we must down-weight the probability of accepting u^* accordingly, otherwise the value u^* will be overrepresented in our sequence.
   2. That the Gibbs sampler is a type of Metropolis-Hastings algorithm is almost as easy to see. In the Gibbs sampler the proposal distribution for U is the full conditional distribution of U given V = v. **if we propose a value from the full conditional distribution the acceptance probability is 1, and the algorithm is equivalent to the Gibbs sampler.**
5. Markov Chain Terminology
   1. Reducibility
      1. A Markov chain is *irreducible* if **you can get from any state to any other state**, eventually (possibly over several steps).
      2. A *reducible* chain has **some states you can never reach** from others.
   2. Periodicity

   1. A state is *aperiodic* if you can return to it **at irregular times**, not just at fixed intervals
   2. A state is *periodic* if you can only return to it **every d steps** for some *d > 1*

3. Recurrency
   1. A state is *recurrent* if **you are guaranteed to return to it eventually** (with probability 1)
   2. If there's a **chance you never return** to a state after leaving it, it's *transient*.
4. **Ergodicity**: A markov chain reaches stationarity if it is irreducible, aperiodic, and recurrent

Generalized Linear Mixed Effects

1. Hierarchal (linear mixed effects, LME) Regression
   1. Using the between group sampling model ($\beta_j = \theta + \gamma_j$) and plugging it into the within group model: $y_{i,j} = \theta^T x_{i,j} + \gamma_j^T x_{i,j} + \epsilon_{i,j}$
      1. The theta matrix is referred to fixed effects and gammas are random effects that are IIDMVN mean 0. Recall the priors are
         1. $\boldsymbol{\theta} \sim \text{multivariate normal}(\boldsymbol{\mu}_0,\ \boldsymbol{\Lambda}_0)$
         2. $\boldsymbol{\Sigma} \sim \text{inverse-Wishart}(\eta_0,\ \mathbf{S}_0^{-1})$
         3. $\sigma^2 \sim \text{inverse-gamma}\left(\frac{\nu_0}{2},\ \frac{\nu_0\sigma_0^2}{2}\right)$
2. Algorithms for the posterior
   1. We can use Gibbs to obtain estimates for (θ, Σ) and Metropolis for β :
      1. Sample $\boldsymbol{\beta}_j^* \sim \text{multivariate normal}(\boldsymbol{\beta}_j^{(s)},\ V_j^{(s)})$
      2. Compute the acceptance ratio $r = \dfrac{p(\mathbf{y}_j \mid \mathbf{X}_j, \boldsymbol{\beta}_j^*)\, p(\boldsymbol{\beta}_j^* \mid \boldsymbol{\theta}^{(s)},\ \boldsymbol{\Sigma}^{(s)})}{p(\mathbf{y}_j \mid \mathbf{X}_j, \boldsymbol{\beta}_j^{(s)})\, p(\boldsymbol{\beta}_j^{(s)} \mid \boldsymbol{\theta}^{(s)},\ \boldsymbol{\Sigma}^{(s)})}$
      3. Sample $u \sim \text{uniform}(0, 1)$.
      4. Set $\boldsymbol{\beta}_j^{(s+1)} = \boldsymbol{\beta}_j^*$ if $u < r$, and $\boldsymbol{\beta}_j^{(s+1)} = \boldsymbol{\beta}_j^{(s)}$ otherwise.
   2. We can simultaneously estimate by a MH algorithm:
      1. Sample $\boldsymbol{\theta}^{(s+1)}$ from its full conditional distribution.
      2. Sample $\boldsymbol{\Sigma}^{(s+1)}$ from its full conditional distribution.
      3. For each $j \in \{1, \ldots, m\}$:
         1. Propose a new value $\boldsymbol{\beta}_j^*$.
         2. Set $\boldsymbol{\beta}_j^{(s+1)}$ equal to $\boldsymbol{\beta}_j^*$ or $\boldsymbol{\beta}_j^{(s)}$ with the appropriate probability.

Latent Variable Methods

1. Ordered Probit and Rank Likelihood

   1. Imagine a situation where you're modeling college students mental states. You get ordinal variables (Happy, Mild, Soggy, Depressed). We can think of these values as ranked from best to worst. This is an ordered probit regression
      1. $\epsilon_1, \ldots, \epsilon_n \sim \text{i.i.d. normal}(0, 1)$
      2. $Z_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i + \epsilon_i$
      3. $Y_i = g(Z_i)$
   2. We refer to Z as a latent variable, and we have to estimate the g-function and the beta vector. Variance is equal to 1 because level variables are typically discrete
   3. Just like ordinary regression, $\boldsymbol{\beta} \sim \text{multivariate normal}(\mathbf{0},\, n(\mathbf{X}^\top \mathbf{X})^{-1})$
   4. The distribution of Z is constrained normal, $p(z_i \mid \boldsymbol{\beta}, \mathbf{y}, \mathbf{g}) \propto \text{dnorm}(z_i, \boldsymbol{\beta}^\top \mathbf{x}_i, 1) \times \delta_{(a,b)}(z_i)$
      1. to sample a value of X from a normal distribution constrained on (a,b):
         1. sample $u \sim \text{uniform}(\Phi[(a-\mu)/\sigma], \Phi[(b-\mu)/\sigma])$
         2. set $x = \mu + \sigma \Phi^{-1}(u)$
            1. where Φ refers to the CDF of the normal distribution

2. Transformation models and Rank Likelihood
   1. we know simple default priors exist for β but this is not the case for the g-function. estimating g is complicated because it depends on b. we can work around becaus if Z was directly observed, then we don't need to estmate g(z). Since g is nondecreasing, we know $Z_1 > Z_2$ so if we observe new data, Z must be in the set $R(\mathbf{y}) = \{\mathbf{z} \in \mathbb{R}^n : z_{i_1} < z_{i_2} \text{ if } y_{i_1} < y_{i_2}\}$.
   2. Knowing this we can obtain the rank likelihood $L(\boldsymbol{\beta}) = \Pr(\mathbf{Z} \in R(\mathbf{y}) \mid \boldsymbol{\beta}) = \int_{R(\mathbf{y})} \prod_{i=1}^{n} \phi(z_i - \mathbf{x}_i^\top \boldsymbol{\beta})\, dz$
   3. The full conditional of Z is thus $p(z_i \mid \boldsymbol{\beta}, \mathbf{Z} \in R(\mathbf{y}), \mathbf{z}_{-i}) \propto \text{dnorm}(z_i, \boldsymbol{\beta}^\top \mathbf{x}_i, 1) \times \delta_{(a,b)}(z_i)$.
      1. Here, $\delta_{(a,b)}(z_i)$ is an indicator function that enforces $z_i \in (a, b)$, which is the truncation interval implied by $\mathbf{z}_{-i}$ and the ranking constraint $R(\mathbf{y})$.
3. The Gaussian Copula Model
   1. When we have numeric ordinal variables, (years of education etc), define:
      1. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \sim \text{i.i.d. multivariate normal}(\mathbf{0}, \boldsymbol{\Psi})$
      2. The observed data are generated as $Y_{i,j} = g_j(Z_{i,j})$,
      3. where $g_j(\cdot)$ is a potentially non-linear, possibly monotonic transformation (e.g., ranking, thresholding).
      4. Ψ is the correlation matrix with diagonal entries equal 1
   2. A model having separate parameters for the univariate marginal distributions and the multivariate dependencies is generally called a copula model.
      1. The term “copula” refers to the method of “coupling” a model for multivariate dependence (such as the multivariate normal distribution) to a model for the marginal distributions of the data.

3. Parameter prior follow:
    1. $\Sigma \sim \text{inverse-Wishart}(\nu_0, \mathbf{S}_0^{-1})$
    2. $\Psi = h(\Sigma) = \left\{\sigma_{i,j} / = \sqrt{\sigma_i^2 \sigma_j^2}\right\}$.
    3. $\mathbf{Z}_1, \ldots, \mathbf{Z}_n \sim \text{i.i.d. multivariate normal}(\mathbf{0}, \Psi)$
    4. $Y_{i,j} = g_j(Z_{i,j})$