What is bayesian probability?

1. Frequentist: Probability is a long run frequency
2. Bayesian: Probability is a degree of belief.
    1. Degree of sureness
3. As opposed to saying the confidence interval is random, the *parameter* is random as well
    1. EX: Say flipping a coin and get heads 7/10 time.
    2. F: 70% chance of heads
    3. B: the true proportion treats a coin flip as Beta(5,5) while we got Beta(7,3), so say that the chance of getting heads is Beta(12,8)=0.6, yielding an expectation and variance.
4. Take the known (prior $p(\theta)$) distribution), and use data to make the unknown (posterior $p(y|\theta)$) distribution.
5. Bayesian Derivation
    1. Posterior is proportional to likelihood * prior
    2. $p(\theta|data) = \theta^{\alpha+n_H-1}(1-\theta)^{\beta+n_T-1} = Beta(\alpha+n_H, \beta+n_T)$
6. Applications
    1. Clinical Trials: we can use clinical trials from drugs to assess performance of new drug
    2. Stock Markets: predicting stock price by using fundamental valuation + drift.
    3. Time series is a variable over time, stochastic processes are different values of the variable at different points in time.
    4. Useful for searching blackboxes after airplane crashes, producing a probability heatmap
    5. Good for limited data, or ones where overparameterization occurs
7. Marginal distributions of (x) have nothing to do with y
8. Different Dichotomous Trials
    1. Bernoulli: Probability of one trial, p=probability of success.
    2. Binomial: Probability of success from n=binomial trials.
    3. Beta: Uncertainty over the probability
    4. Remark: Recall that the gamma function $\Gamma(n) = (n-1)!$
9. Maximum Likelihood Estimate MLE
    1. $logp(\theta|y) = \sum_i \{y_i log(\theta) + (1-y_i) log(1-\theta)\}$
10. Binomial (Single Parameter) Model
    1. Model: $P(Y=y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$
    2. A uniform prior makes a beta posterior, but a beta prior makes another beta posterior

3. Posterior Mean: $E[\theta|y] = \frac{a+y}{a+b+n}$

11. Confidence Regions
    1. An interval from a to b is $(1-\alpha)100\%$ credible for θ if $Pr(a < \theta < b|y) = (1-\alpha)100\%$
    2. Different from Confidence Intervals, the probability that the interval will cover the true value *before* the data are observed
12. High Posterior Density (HPD) Regions
    1. A HPD $(1-\alpha)100\%$ Region is a subset of the parameter space such that:
    2. $\Pr(\theta \in A \mid y) = 1 - \alpha$
    3. $\text{If } \theta_1 \in A \text{ and } \theta_2 \notin A, \text{ then } p(\theta_1 \mid y) > p(\theta_2 \mid y)$
    4. They may not have symmetric tails (skewed dist) and may not be an interval (bimodal dist)
13. Now, given a new data point IID, we want to make a prediction of the possible values through the predictive (posterior) distribution
    1. $p(z=1|y) = \int p(z=1,\theta|y_1,\ldots,y_n), d\theta = E(\theta|y_1,\ldots,y_n) \implies \frac{a+\sum_i y_i}{a+b+n}$
    2. Remark: $\theta = p(z=1|\theta, y_1,\ldots,y_n)$
14. Conjugate Priors
    1. if prior and posterior distributions belong to the same family, they are conjugate priors
        1. Advantage: easy to derive posterior distribution
        2. Disadvantage: expert opinion may not conform to conjugate prior
    2. For the exponential family of distributions
        1. $p(y|\theta) = f(y)g(\theta)exp\{\phi(\theta)u(y)\}$, turning the product into the sum
        2. Normal, Binomial, Poisson, Negative Chi-Square

Model	Parameter	Prior	Posterior
$X \sim \text{Binomial}(n,p)$	$0 \le p \le 1$	$\text{Beta}(a,b)$ $a>0, b>0$	$\text{Beta}(a',b')$ $a' = a + x$ $b' = b + n - x$
$X = (X_1, \cdots, X_n)$ $X_i \overset{\text{IID}}{\sim} \text{Poisson}(\lambda)$	$\lambda > 0$	$\text{Gamma}(a,b)$ $a>0, b>0$	$\text{Gamma}(a',b')$ $a' = a + n\bar{x}$ $b' = b + n$
$X = (X_1, \cdots, X_n)$ $X_i \overset{\text{IID}}{\sim} \text{Exponential}(\lambda)$	$\lambda > 0$	$\text{Gamma}(a,b)$ $a>0, b>0$	$\text{Gamma}(a',b')$ $a' = a + n$ $b' = b + n\bar{x}$
$X = (X_1, \cdots, X_n)$ $X_i \overset{\text{IID}}{\sim} \text{Normal}(\mu, \sigma^2)$ σ^2 known	$-\infty < \mu < \infty$	$\text{Normal}(a, b^2)$ $-\infty < a < \infty$ $b > 0$	$\text{Normal}(a', b'^2)$ $a' = \frac{nb^2\bar{x} + \sigma^2 a}{nb^2 + \sigma^2}$ $b'^2 = \frac{\sigma^2 b^2}{nb^2 + \sigma^2}$
$X = (X_1, \cdots, X_n)$ $X_i \overset{\text{IID}}{\sim} \text{Normal}(\mu, \sigma^2)$	$-\infty < \mu < \infty$ $\sigma^2 > 0$	$\text{N}\chi^{-2}(m,k,r,s)$ $-\infty < m < \infty$ $k>0, r>0, s>0$	$\text{N}\chi^{-2}(m',k',r',s')$ $m' = \frac{km + n\bar{x}}{k+n}$ $k' = k + n$ $r' = r + n$ $s' = \frac{rs + \frac{kn}{k+n}(\bar{x}-m)^2 + (n-1)s_x^2}{r+n}$

15. Poisson Model (birth rate example)
    1. Assume both θ_1 and θ_2 are IID Poisson. Given the (n, $\sum y$, $\bar{y}$) for both are $n_1 =$ (111, 217, 1.95) and $n_2 =$(44,66,1.5)
    2. Recall that a poisson model:
    3. $Pr(Y = y|\theta) = \frac{\theta^y e^{-\theta}}{y!}$ with expected value and variance equal to θ
    4. The likelihood $L(\theta) = e^{-n\theta}\theta^s$ where $s = \sum(Y_i)$
    5. The conjugate prior is Gamma(a,b) and $\pi(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}$
        1. We can treat $\Gamma(a) = (a-1)!$ with $E(\theta) = \frac{a}{b}$ and $Var(\theta) = \frac{a}{b^2}$
    6. The posterior distribution is $p(\theta|y) = e^{-(n+b)\theta}\theta^{s+a-1}$ with $Var(\theta) = \frac{a+s}{(b+n)^2}$
        1. $E(\theta) = w\mu_{prior} + (1-w)MLE(\theta) = \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{s}{n}$
        2. For either n-fixed, b-sufficiently large or vice versa, w approaches 0

Prior Types

1. We need to be very careful when specifying the prior, since Bayes Stats relies so heavily on it. Likelihood favors prior, and if prior is zero, mode is extreme, which is not desirable
2. *Subjective Priors*: priors chosen to reflect expert opinion or personal beliefs
    1. Ways to specify
        1. for a single event, look at A and the complement of A
        2. for continuous quantities, divide the distribution into intervals and assign probabilities for each
            1. if an expert says $\theta \sim Beta(a,b)$ exhibits mean 0.3 and sd 0.1, we can solve for a and b to find that it is Beta(6,14)
            2. if an expert says the sample mean is 25 but is 95% confident it is between 10 and 40, we can infer $N \sim (25, \frac{30}{4})$
3. *Objective Priors*: priors chosen to let data (likelihood) dominate the posterior, based on a sampling model in use
4. *Non-Informative Priors*: choose priors that reflect no information about θ, (prior has minimal impact on the posterior)
    1. *Reference Priors*
        1. Assume a sample was drawn from Normal and that θ is unknown but σ is known
            1. we can assume the prior $\pi(\theta) = 1$ which means all values of theta are equally likely (known as a flat prior0
                1. flat priors don't have to follow a probability distribution, but the posterior in complex models cannot be improper
        2. Constant (flat) priors are not transformation invariant

    2. The remedy for flat priors are known as *Jeffrey Priors*
        1. For single parameter priors
        2. $\pi(\theta) \sim \sqrt{I(\theta)}$ where $I(\theta) = -\mathbb{E}_{y|\theta}\left[\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2}\right]$
        3. these are transformation invariant as well
        4. Fact: A second derivative of a log of a function is equal to the derivative of the function squared: $\frac{d^2 \log f(\theta)}{d\theta^2} = (\frac{d \log f(\theta)}{d\theta})^2$
5. *Diffuse Priors*
    1. A prior with little information about the parameter
    2. choose a conjugate prior with large standard deviation as a diffuse or weakly informative prior
        1. EX: if normally distributed and variance known, a prior could set variance to 1000
6. Prelude to MCMC Methods
    1. Posterior quantities cannot be derived in closed forms.
    2. Available methods to find posterior distributions
        1. Analytical methods: laplace approximation, numerical integration (not covered)
        2. Simulation Methods: based on direct sampling, like rejection/importance sampling
        3. Markov Chains: Metropolis-Hasting or Gibbs Sampling

The Normal Model

1. The normal distribution
    1. PDF: $N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$
    2. Using precision $\tau = \sigma^{-2}$: $N(x \mid \mu, \tau^{-1}) = \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{1}{2}\tau(x-\mu)^2\right)$
2. Common Properties
    1. Mean, median, and mode are all μ.
    2. Symmetric about the mean
    3. 95% probability within $\pm 1.96\sigma$.
    4. if $X \sim N(\mu, \sigma^2)$ and $Y \sim N(m, s^2)$ IID, then $aX + bY \sim N(a\mu + bm, a^2\sigma^2 + b^2 s^2)$
3. Normal-Normal Model
    1. Denoted by $p(\theta) = N(\theta|\mu_0, \tau_0^{-1})$
    2. Posterior Derivation:
        1. Begin with likelihood $\mathcal{N}(x \mid \theta, \ell^{-1}) \propto_\theta \exp\left(\ell x\theta - \frac{1}{2}\ell\theta^2\right)$
            1. We drop the constant term and we will do this a lot when working with normal distribution
        2. Due to symmetry of normal distribution, the prior is the same, replacing $x = \mu_0$ and $\ell = \tau_0$

    3. By bayes rule, the posterior is
       $p(\theta|x_1, x_2, \ldots, x_n) = N(\theta|M, L^{-1}) = N(\theta|\frac{\tau_0\mu_0+\tau\sum x_i}{\tau_0+n\tau}, \tau_0 + n\tau)$
    4. The likelihood is $L(\mu, \tau) \propto \tau^{n/2} \exp\left(-\frac{1}{2}\tau\sum_i(y_i - \mu)^2\right)$
3. Non informative prior ($\pi(\mu, \sigma^2) \propto \tau$)arises when mean and variance are a priori independent and taking the product of the standard non-inf priors.
    1. This is not a conjugate setting (the posterior does not factor into a product of two independent distributions).
    2. Prior is improper but posterior is proper.
    3. This is also the Jeffreys' prior.
4. Conditional posterior distribution:
   $p(\mu \mid \sigma^2, y) \propto L(\mu \mid \sigma^2, y) \cdot \pi(\mu) \propto \exp\left(-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right) \sim \mathcal{N}\left(\bar{y}, \frac{\sigma^2}{n}\right)$
5. Marginal posterior distribution: $\tau|y \sim \Gamma(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$
6. Sampling from the joint posterior distribution
    1. simulate σ^2 from $\Gamma^{-1}(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$
    2. simulate μ from $N(\bar{y}, \frac{\sigma}{n})$
7. Student's t-Distribution
    1. The t-distribution is $t_v = \frac{Z}{\sqrt{\frac{U}{v}}}$, with v degrees of freedom
        1. Z is standard normal
        2. U is Chi-Square with v degrees of freedom
        3. Z is independent of U
    2. Connection to bayesian inference
        1. when marginalizing over variance, the posterior of μ follows a t-distribution rather than a normal distribution since it is more robust to outliers

Multinomial Model

---

1. $\vec{x} = (x_1, \ldots, x_k)$ is a vector of counts where x_j is the number of counts for the j-th category, and $\sum x_j = n$
2. $p(x|\theta) = \prod_{j=1}^{k} \theta_j^{x_j}$ with Sum of $x_j = 1$
3. Conjugate prior is $\theta \sim D(\alpha)$ Dirichlet Distributed:
    1. $\theta = (\theta_1, \theta_2, \ldots, \theta_k) \quad \text{with} \quad \sum_{j=1}^{k} \theta_j = 1$; $\pi(\theta) = \frac{1}{B(\alpha)} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}, \quad \alpha_j > 0$
    2. Where $B(\alpha)$ is the multivariate Beta distribution
    3. Remark: when k=2, it reduces to beta distribution
    4. The parameter α is for counts, so:
        1. if α_j is large, we have strong belief for category j

        2. if $\alpha_j = 1$, we get uniformity
4. Posterior: $\theta|x \sim D(\alpha + x) = \pi(\theta \mid x) \propto \prod_{j=1}^{k} \theta_j^{x_j+\alpha_j-1}$
5. priors are improper if $\alpha_j = 0 \forall j$

Bayes Factors

1. used to text hypothesis and compare models in bayesian framework
2. Suppose we have two models M_1 and M_2 with parameter vectors θ_1 and θ_2.
    1. Bayes Factors: $BF = \frac{\frac{p(M_1 \mid x)}{p(M_2 \mid x)}}{\frac{p(M_1)}{p(M_2)}} = \frac{p(x|M_1)}{p(x|M_2)}$
3. The marginal distribution of x under model M_1 is
   $p(x \mid M_i) = \int f(x \mid \theta_i, M_i)\, \pi_i(\theta_i)\, d\theta_i, \quad i = 1, 2$
4. The prior is how likely are the data based on each model, integrating over the uncertainty in the parameters
5. Marginalization Paradox:
    1. if $\pi_i(\theta_i)$ is improper then $p(x \mid M_1)$ is improper, so BF is undefined
        1. Ok for parameter estimation but not for model comparison
6. Bayes Factor Interpretation:

Jeffreys' - scale of evidence in favor of M_1		
log_{10}BF	Bayes factor	Interpretation
0 - 0.5	$1 \leq BF \leq 3.2$	weak
0.5 - 1.0	$3.2 < BF \leq 10$	substantial
1.0 - 2.0	$10 < BF \leq 100$	strong
> 2	$BF \geq 100$	decisive

Kass & Raftery - scale of evidence in favor of M_1		
2*ln*BF	Bayes factor	Interpretation
0 - 2	$1 \leq BF \leq 3$	weak
2 - 6	$3 < BF \leq 20$	positive
6 - 10	$20 < BF \leq 150$	strong
> 10	$BF \geq 150$	very strong

7. Frequentist vs Bayes Hypothesis Testing
    1. F: One sample z-test for proportions
        1. Null hypothesis: $H_0 : \theta \leq 0.6$ with sample (observed) proportion $\hat{\theta} = 0.81$ yields a z-statistic of 2.18 (p=0.0147). Since p<0.05, we reject the null and conclude strong evidence for $\theta \geq 0.6$
    2. B: Posterior distribution with bayes factors
        1. We choose three priors: Beta(0.5,0.5), Beta(1,1), and Beta(2,2). We find the posterior probability $P(\theta > 0.6|x) \approx 0.93 - 0.96$. The bayes factors are 34, 30, and 24 for each of the betas. There is strong evidence favoring $\theta \geq 0.6$
    3. Conclusion: Frequentist approach rejects the null without strength measure, while bayesian gives strong evidence strength
8. BF is defined ONLY for proper prior distributions and may be sensitive to prior choices
9. Other Model Selection Criteria
    1. Bayes Information Criterion (BIC)
    2. Log-pseudo-marginal Likelihood (LPML)
    3. deviance information Criterion (DIC)

Monte Carlo and Indirect Methods

---

1. Used to perform numerical integration for $E(f(\theta)) = \int f(\theta)p(\theta)d\theta$
    1. If we cannot directly sample from the joint posterior, we cannot use a regular MC algorithm, we instead use:
    2. Importance sampling
        1. suppose $f(x)$ can be approximated from some easily sampleable $g(x)$
        2. $E(h(x)) = \mu = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx = \int h(x)w(x)g(x)dx$
            1. w(x) is the weighting function, and g(x) is the envelope function
        3. Unstandardized method: $\hat{\mu}^*_{IS} = \frac{1}{n}\sum_{i=1}^n h(X_i)w^*(X_i)$, $w^*(X_i) = \frac{f(X_i)}{g(X_i)}$ is the unstandardized weights
        4. Standardized method: $\hat{\mu}_{IS} = \sum_{i=1}^n h(X_i)w(X_i)$, $w(X_i) = \frac{w^*(X_i)}{\sum_{i=1}^n w^*(X_i)}$ is the standardized weight
        5. Effective sample size for importance efficiency (ESS)
            1. ESS indicates that n weighted samples used in IS are worth N(f , g) unweighted i.i.d. samples drawn exactly from f and used in a simple Monte Carlo estimate
            2. Unstandardized: $N(f, g) = \frac{n}{1+\widehat{V}\{w^*(x)\}}$, estimates the effective sample size from the variance of unnormalized importance weights

        3. Standardized: $N(f,g) = \frac{n}{1+\hat{cv}^2\{\tilde{w}(x)\}}$, uses normalized weights, coefficient of variation, and constant factor from the variance estimator
    3. Accept/Reject AR sampling
        1. only requires us to know the functional form of f using a simpler function g
            1. ALGORITHM
            2. given a density f, find a density g and constant M such that $f(x) \leq Mg(x)$
            3. Sample $X \sim g(x)$ (from the proposal density)
            4. Sample $U \sim \text{Uniform}(0,1)$
            5. ACCEPT if $U \leq \frac{f(X)}{Mg(X)}$
                1. If accepted, set $Y = X$ as a sample from $f(x)$
            6. REJECT and go back to step 1
        2. choose a small M for better efficiency
        3. g must be simple (uniform, triangular, normal, any density that can be obtained by inversion)
            1. $\frac{f(X)}{Mg(X)}$ must be bounded
        4. Choose a class of densities for g and then find that density for which M is the smallest. (ex: normal from double exponential)

Markov Chain Monte Carlo

---

1. Markov Chains
    1. Sequence of random variables with the markov property: given the present states, the future and past states are independent. The possible values of X_i form a sample space S called the state space of the chain
    2. $\Pr(X_{n+1} = x_j \mid X_n = x_n, \cdots, X_1 = x_1) = \Pr(X_{n+1} = x_j \mid X_n = x_n)$
    3. Defined by:
        1. State space:
        2. Transition Probability: $p_{ij} = \Pr(X_{n+1} = j \mid X_n = i)$
        3. Initial distribution π_0
    4. Properties
        1. Reducibility: $\Pr(X_n = j \mid X_0 = i) = p_{ij}^{(n)} > 0$
            1. A state j is said to be accessible from a state i if a system started in state i has a non-zero probability of transitioning into state j at some point.
            2. A chain is irreducible if all states communicate (there's a positive probability to visit all states in a finite number of steps)
        2. Periodicity

    1. A state i has period k if any return to state i must occur in multiples of k time steps. If $k = 1$, then the state is said to be aperiodic i.e. returns to state i can occur at irregular times.
3. Recurrence
    1. A state i is said to be transient if, given that we start in state i, there is a non-zero probability that we will never return to i.
    2. A state i is called absorbing if it is impossible to leave this state.
    3. A state i is recurrent if the expected number of visit to i is equal to infinity.
4. Ergodicity: To what is the chain converging?
    1. $\lim_{n\to\infty} \|P^n - \pi\| = 0$
    2. A chain is said to be ergodic if it is aperiodic and positive recurrent. (positive = the chain has an invariant probability measure)
5. Stationary distribution
    1. A markov chain with transition matrix P will have an equilibrium distribution π iff $\pi = \pi P$. It is reversible only if $\pi_j p_{ji} = \pi_i p_{ij}$ (detailed balance condition)
        1. continuous case: $\pi(x)p(x,y) = \pi(y)p(y,x)$
    2. to sample from the limiting distribution π, we run a Markov Chain with transition matrix P satisfying the detailed balance condition until the chain appears to have settle down to equilibrium
6. A Markov Chain Monte Carlo (MCMC) method for simulation from a distribution π is any method producing an ergodic Markov chain (X_n) whose stationary distribution is π
    1. For an arbitrary starting value x_0, a chain X_n is generated using a transition kernel with stationary distribution π, which ensures the convergence in distribution of X_n to a random variable from π.
    2. In order for the Markov chain to converge to the target (stationary or equilibrium) distribution π, it must be irreducible, aperiodic and positive recurrent
7. Metropolis Hastings (MH) Algorithm
    1. Assume the process moves from x to y too often and y to x too rarely
    2. Introduce the probability of movement $\alpha(x,y) < 1$
    3. $\alpha(x,y) = \min\left\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)},\ 1\right\}$.
    4. MH for inference about θ
        1. Start with θ^0 usually from prior
        2. At iteration t, generate a proposal distribution $\theta^* \sim q\left(\theta^* \mid \theta^{(t-1)}\right)$
        3. Take $\theta^{(t)} = \begin{cases} \theta^*, & \text{with probability } \rho\left(\theta^{(t-1)}, \theta^*\right) \\ \theta^{(t-1)}, & \text{with probability } 1-\rho\left(\theta^{(t-1)}, \theta^*\right) \end{cases}$
            1. where $\rho\left(\theta^{(t-1)}, \theta^*\right) = \min\left\{\frac{\pi(\theta^*|x)\ q\left(\theta^{(t-1)}|\theta^*\right)}{\pi\left(\theta^{(t-1)}|x\right)\ q\left(\theta^*|\theta^{(t-1)}\right)},\ 1\right\}$ is the MH Acceptance Ratio

4. update is obtained by sampling on uniform and setting $\theta^t = \theta^*$
5. A chain constructed via the MH algorithm is Markov,
6. If the process is symmetric, it is known as a Metropolis Algorithm
7. MH is dependent of normalizing constants and, in turn, conditioning variables
8. In addition, a good proposal distribution would satisfy the following:
    1. it should be easy to sample from
    2. it should be easy to compute the acceptance ratio
    3. each proposal should go a reasonable distance in the parameter space, otherwise the random walk moves too slowly
    4. the proposals are not rejected too frequently

8. Integral Approximation
    1. Ergodic Theorem: Once the chain has reached convergence to the stationary distribution, π, we can use the realizations of the chain to approximate quantities of interest such as $\mathbb{E}[h(\theta \mid x)] \approx \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)})$
    2. Although the MH algorithm is valid for any q satisfying the mild conditions we have seen before, the choice of the proposal greatly affect the efficiency of the algorithm.
9. Independent MH Algorithm
    1. Considers a proposal q which is independent of the current state. Then, it proceeds as follows:
    2. Given θ^{t-1}
    3. Generate $\theta^* \sim q(\theta^*)$
    4. Take $\theta^{(t)} = \begin{cases} \theta^*, & \text{with probability } \min\left\{\frac{\pi(\theta^*|x)}{\pi(\theta^{(t-1)}|x)} \cdot \frac{q(\theta^{(t-1)})}{q(\theta^*)}, 1\right\} \\ \theta^{(t-1)}, & \text{otherwise} \end{cases}$
    5. although θ is generated independently, the resulting sample is not IID
10. Random Walk Metropolis Hastings (RWMH)
    1. Intuitively, the chain might be more efficient if we consider proposals that take into account the value previously simulated to generate the following value; that is, if we consider a local exploration of the neighborhood of the current value of the Markov Chain
    2. A first choice is to simulate θ^* as $\theta^* = \theta^{(t-1)} + \epsilon$
    3. Now $q(\theta^* \mid \theta^{(t-1)}) = g(\theta^* - \theta^{(t-1)})$
    4. Choices of g: Most common choices are:
        1. Uniform distributions on spheres centered at the origin
        2. Scaled normal distribution
        3. Scaled Student's t distribution

2. Summary so far

    1. The idea of Markov chain simulation is to simulate a random walk in the space of θ, which converges to a stationary distribution
    2. In Markov chain simulations, the samples are drawn sequentially, with the distribution of the sampled draws depending on the last value drawn
    3. The key is that the approximate distributions are improved at each step in the simulation, in the sense of converging to the target distribution
3. Standard practice for MCMC approximation
    1. Run the algorithm until some iteration B for which it looks like the Markov chain has achieved stationarity;
    2. Run the algorithm S more times
    3. discard the "B" simulations and use the empirical distribution of "S" to approximate $p(\theta|y)$
    4. "B": burn in period in which the Markov chain moves from its initial value to a region of the parameter space that has high posterior probability

Gibbs Sampler and MCMC diagnostics

1. Toxicity test: assume data is provided that gives compound doses to animals. the dosage, number of animals, and number of deaths are reported
2. Outcome model: $y|\theta \sim \text{Bin}(n, \theta)$
3. Could model probability of death seperately but this ignores dosage amount since dose amount is a factor
4. a typical dose response model is $\text{logit}(\theta) = \alpha + \beta x$; $\text{logit}(\theta) = \log(\frac{\theta}{1-\theta})$
    1. Likelihood: $p(y_i \mid \cdot) = \left[\text{logit}^{-1}(\alpha + \beta x_i)\right]^{y_i} \left[1 - \text{logit}^{-1}(\alpha + \beta x_i)\right]^{n_i - y_i}$
    2. Posterior: $\pi(\alpha, \beta \mid y, n, x) \propto \pi(\alpha, \beta) \prod_{i=1}^{k} p(y_i \mid \alpha, \beta, n_i, x_i)$
5. We need to take into account dimensionality of parameters. MH algorithm is inefficient if we sample from a multidimensional target distribution
6. Solution: Sequentially sample from univariate conditional distributions, instead of a single multivariate distribution (Gibbs Sampler)
7. Create a markov chain based on conditionals $\pi(\theta|x, \lambda)$ and $\pi(\lambda|x, \theta)$ to draw samples from $\pi(\theta, \lambda|x)$ (stationary distribution)
    1. known as two-stage gibbs sampler or data augmentation algorithm (since this is used to fill missing data)
    2. Order is not important, so it can go 1-2-2-2-1 etc (called random scan gibbs)
8. Convergence
    1. the validity of the MCMC approach relies on the fact that we can confidently assume that the realizations of the MC are indeed samples from the target distribution

    1. Theoretical (probabilistic) perspective: measuring the distance and establishing theoretical bounds (hard or even impossible to calculate)
    2. Statistical perspective: analyzing the properties of the observed output from the chain
2. It must be used with caution because it is based on the empirical properties of a single chain or a set of chains; hence, it cannot ever guarantee convergence is reached with certainty
3. Diagnostic
    1. An MCMC algorithm has converged at iteration T when its output can be safely thought to arise from the true stationary distribution of the Markov chain for all $t > T$.
    2. If the MCMC fails to converge to the target distribution, the resulting estimates will be biased and unreliable.
    3. There has been much effort in developing MCMC convergence diagnostics.
    4. Note: Passing an MCMC diagnostic does not guarantee that a chain is stationary! You can only check for signs of non-convergence.
        1. Statistics
            1. Geweke: tests for the equality of means of the first part (first 10%) and the last part (last 50%) of the markov chain. if the samples are drawn from stationary data, the means are equal
            2. Gelman and Rubin : convergence is diagnosed when chains have forgotten their initial values, and then within class and between class variance is analyzed (ANOVA). values near 1 suggest convergence
            3. Heidelberger and Welch : uses Cramer-von-Mise statistic on the whole chain, then after discarding 10%, 20%, to 50%. If the stationarity test is passed, the number of iterations to keep and the number to discard (burn-in) are reported.
            4. Raftery and Lewis: Values of I larger than 5 indicate strong autocorrelation which may be due to a poor choice of starting value, high posterior correlations or stickiness of the MCMC algorithm. The number of burn-in iterations to be discarded at the beginning of the chain is also calculated
4. Quick checks
    1. A trace plot of the sequence $\theta(t)$'s against t is a first empirical check of convergence.
    2. In cases of strong attraction from a local mode, the chain can behave as if it was simulated from the neighborhood of this mode and appear to have converged, when in fact, it has not.

3. It is common to run a few parallel chains from different starting values until convergence and check that they settle around common values

9. Autocorrelation
    1. How quickly the sampled values move around the parameter space is called the speed of mixing
    2. A Markov chain with a high autocorrelation moves around the parameter space slowly, taking it longer to achieve stationarity
    3. High autocorrelations within chains indicate slow mixing and slow convergence
10. We cannot guarantee the algorithm has converged, only that it has not
11. Solutions
    1. Run longer and thin output
    2. Reparametrize model. Models that are overparametrized lead to high posterior correlations among the parameters and a dramatic slow down of the movement of the MCMC sampler through the parameter space.
    3. “Block” correlated variables together
    4. Integrate out variables
    5. Add auxiliary variables (Slice-sampler; data augmentation)

Multivariate Normal Model

1. Distribution type
    1. Uses multiple measurements for each experimental unit
    2. Y is a random vector (each component is normal distribution) and the mean is now a vector with a covariance matrix
    3. must ensure the covariance matrix is symmetric positive definite
        1. this matrix follows a wishart distribution
            1. $\text{Wishart}(n, \frac{1}{S}) = f(\mathbf{\Phi} \mid \mathbf{S}^{-1}, n) = \frac{|\mathbf{\Phi}|^{(n-k-1)/2}|\mathbf{S}^{-1}|^{-n/2}\exp\left(-\frac{1}{2}\text{tr}(\mathbf{S\Phi})\right)}{2^{nk/2}\pi^{k(k-1)/4}\prod_{j=1}^{k}\Gamma\left(\frac{n+1-j}{2}\right)}$
            2. expectation: $\mathbb{E}[\mathbf{\Phi}] = n\mathbf{S}^{-1}$
        2. the conjugate prior is inverse wishart
            1. $\text{Wishart}(n, \frac{1}{S}) = f(\mathbf{W} \mid n, \mathbf{S}^{-1}) = \frac{|\mathbf{W}|^{-(n+k+1)/2}|\mathbf{S}|^{n/2}\exp\left\{-\frac{1}{2}\text{tr}(\mathbf{SW}^{-1})\right\}}{2^{nk/2}\pi^{k(k-1)/4}\prod_{j=1}^{k}\Gamma\left(\frac{n+1-j}{2}\right)}$
            2. expectation: $\mathbb{E}[\mathbf{W}] = \mathbb{E}[\mathbf{\Phi}^{-1}] = \frac{1}{n-k-1}\mathbf{S}$
    4. Multivariate t-distribution:
    $f(\mathbf{y} \mid \nu, \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{\Gamma\left(\frac{\nu+k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\nu^{k/2}\pi^{k/2}|\mathbf{\Sigma}|^{1/2}}\left[1 + \frac{1}{\nu}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]^{-\frac{\nu+k}{2}}$
        1. Why use?
            1. More robust to outliers

    2. Converges to normal: $\nu \to \infty, f(\mathbf{y}) \to \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
    3. Mean only exists if v>1, covariance only exists when v>2
2. Marginal posterior of mu $\boldsymbol{\mu} \mid \mathbf{Y} \sim \text{Multivariate t}\left(\nu_n - k + 1, \boldsymbol{\mu}_n, \frac{\kappa_n+1}{\kappa_n(\nu_n-k+1)}\mathbf{S}_n\right)$
    1. where $\kappa_n = \kappa_0 + n$, $\nu_n = \nu_0 + n$
    2. Posterior mean: $\boldsymbol{\mu}_n = \frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\mathbf{y}}}{\kappa_0+n}$
    3. Posterior scale matrix: $\mathbf{S}_n = \mathbf{S}_0 + \mathbf{S} + \frac{n\kappa_0}{n+\kappa_0}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)^\top$
3. Marginal posterior of Sigma $\boldsymbol{\Sigma} \mid \mathbf{Y} \sim \mathcal{IW}(\nu_n, \mathbf{S}_n)$
    1. Expectation: $\mathbb{E}[\boldsymbol{\Sigma} \mid \mathbf{Y}] = \frac{\mathbf{S}_n}{\nu_n - k - 1} \quad (\text{if } \nu_n > k+1)$
4. Marginal predictive distribution

   $\mathbf{y}_{n+1} \mid \mathbf{Y} \sim \text{Multivariate t}\left(\nu_n - k + 1, \boldsymbol{\mu}_n, \frac{\kappa_n+1}{\kappa_n(\nu_n-k+1)}\mathbf{S}_n\right)$
    1. Posterior predictive density:

       $f(\mathbf{y}n+1 \mid \mathbf{Y}) = \frac{\Gamma\left(\frac{\nu_n+1-k}{2}\right)}{\Gamma\left(\frac{\nu_n-k+1}{2}\right)(\pi(\nu_n-k+1))^{k/2}|\boldsymbol{\Sigma}\text{pred}|^{1/2}} \cdot \left[1 + \frac{1}{\nu_n-k+1}(\mathbf{y}n+1 - \boldsymbol{\mu}n)^\top \boldsymbol{\Sigma}\text{pred}^{-1}(\right.$

       where $\boldsymbol{\Sigma}_{\text{pred}} = \frac{\kappa_n+1}{\kappa_n(\nu_n-k+1)}\mathbf{S}_n$

2. Multivariate normal with semiconjugate prior has mean that follows multivariate normal and covariance that follows inverse wishart
    1. posterior density does not follow a standard parametric form
    2. gibbs sampling is used to find the posterior samples
3. Example: 22 students are given reading comprehension test before and after receiving a particular instruction instruction method. First column is pre-instructional scores and second column is post instructional scores
    1. Does the instructional method lead to improvements in reading comprehension, on average? If so, by how much?
    2. Can improvements be predicted based on the first test?
    3. We can model the data as a bivariate normal distribution and use bayes theorem to get the posterior distribution. We let the semi conjugate prior for the mean be bivariate normal and for covariance to be inverse wishart
    4. The test is designed to have a mean score of 50: $\mu = 50 = (50, 50)^T$
    5. The true mean is constrained from 0 to 100, implying $\mu \pm 2\sigma = [0, 100]$ so $\sigma^2 = \frac{\mu}{2}^2$ =625
    6. We assume a correlation of 0.5, so the covariance is 0.5x625=312.5
    7. The mean of inverse wishart is $\frac{\Sigma}{v-k-1}$ so setting $v_0 = k + 2 = 4$ for k=2 groups, and this leads to $E(\Sigma) = S_0$, now we can specify tyhe distributions:
        1. $\boldsymbol{\mu} \sim \mathcal{N}\left((50 \quad 50)^\top, \mathbf{S}_0\right), \quad \boldsymbol{\Sigma} \sim \mathcal{IW}\left(4, \mathbf{S}_0^{-1}\right), \quad \mathbf{S}_0 = \begin{bmatrix} 625 & 312.5 \\ 312.5 & 625 \end{bmatrix}$

Group Comparisons and Hierarchal Models

1. Group comparisons
    1. We have $Y_{i1} = \mu + \delta + \epsilon$ and $Y_{i2} = \mu - \delta + \epsilon$ where μ is the group mean, δ is a test difference, and ϵ is for noise.
        1. if $\delta = 0$ then there is no difference between groups
        2. if $\delta = 0$ then there is no difference between groups
        3. if $\delta > (<) 0$ then group 1 has a higher (lower) mean than group 2
    2. We estimate δ and test if it is significantly different from 0. The larger that $|\delta|$ is relative to σ, the stronger the evidence is for group differences
    3. The exercise of specifying a model over several levels is called hierarchical modeling, with each new distribution forming a new level in the hierarchy.
    4. In a hierarchical model, the observations are given distributions conditional on parameters, and the parameters in turn have distributions conditional on additional parameters called hyperparameters.
        1. Non-hierarchical models are usually inappropriate for hierarchical data
            1. with few parameters, they generally cannot fit the data adequately
            2. with many parameters, they tend to “overfit” the data.
    5. Hierarchical models allow information to be shared across groups of observations.
    6. EX: Student test scores across different schools:
        1. Traditional (non-hierarchical) model: Each school is modeled separately.
        2. Hierarchical model: We assume schools share a common mean and variance structure, allowing information to be shared across schools.
2. Hierarchal Normal Model
    1. The hierarchal normal model is used to describe heterogeneity (difference) of means
        1. $Y \sim N(\theta, \sigma^2)$ is the within group model (by default, assume σ^2 is constant)
        2. $\theta \sim N(\mu, \tau^2)$ is the between group model
    2. The unknown quantities are the group means, within group sampling variability σ^2, and the mean and variance of group specific, all obtainable via Gibbs Sampling
3. Recall: One-Way ANOVA
    1. Model $Y_{ij} = \theta_j + \epsilon_{ij} \sim N(0, \sigma^2)$ where the variance is the deviation of a single student from the school mean
    2. θ is every school's mean (fixed effect)
    3. We test means the same vs means not the same
    4. Only tests two extremes: equal or not equal
4. Hierarchal (Random Effects) ANOVA
    1. $Y_{ij} = \mu + \alpha_j + \epsilon_{ij} \sim N(0, \sigma^2); \alpha \sim N(\mu, \tau^2)$

2. instead of just the error following random noise, now the means (θ) are normally distributed $N(\mu, \tau^2)$
3. now μ is a fixed effect, σ^2 is the variance component (within), and τ^2 is the between group variance
4. there are two additional components compared to the m+1 fixed effects model
5. group effects are random draws from the population (α)
6. Appropriate when groups represent a random sample from a larger population

5. Marginal Model
    1. because linear combinations of normals are still normal, we can say the model is $N(\mu, \tau^2 + \sigma^2)$ which implies students within schools are exchangeable and student achievements across different schools are independent given the school effect
    2. Intra-class correlation $\text{Corr}(Y_{ij}, Y_{i'j}) = \frac{\tau^2}{\tau^2+\sigma^2}$
        1. measure of the proportion of total variation that is explained by between group variability
        2. It is 0 when $\tau^2 = 0$ and 1 when $\sigma^2 = 0$
    3. We can use restricted MLE (REML) to find the estimates for the parameters
6. Bayesian Heirarchal Model
    1. Unknown parameters: $\mu, \theta, \tau^2, \sigma^2$
    2. specify priors for all parameters but theta
    3. default prior: $p(\mu, \tau^2, \sigma^2) \propto \frac{1}{\sigma^2}$ since a reference prior would cause an improper posterior
    4. Model: $p(\vec{\theta}, \mu, \sigma^2\tau^2|Y) \propto p(Y|\vec{\theta}, \sigma^2)p(\vec{\theta}|\mu, \tau^2)p(\mu, \tau^2, \sigma^2)$
    5. We cannot obtain the posterior distributions in closed form.
    6. We can use Gibbs sampling and create a Markov chain that generates values from the following full conditional distributions:
        1. $p(\theta_j \mid Y, \theta_{-j}, \mu, \sigma^2, \tau^2) \quad \text{for } j = 1, \ldots, m$
        2. $p(\mu \mid Y, \theta, \sigma^2, \tau^2)$
        3. $p(\sigma^2 \mid Y, \theta, \mu, \tau^2)$
        4. $p(\tau^2 \mid Y, \theta, \mu, \sigma^2)$
    7. Shrinkage Effect: average of empirical mean $\bar{y}$ and μ (pull away depending on n)
        1. The more we have, the more information. less we have, we borrow information
        2. $E[\theta_j \mid \mu, \tau, Y] = \frac{n_j\bar{y}_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}$
    8. Model extension: unequal variances
        1. Priors:
            1. $\theta_j \mid \mu, \tau^2 \sim \mathcal{N}(\mu, \tau^2)$
            2. $\sigma_j^2 \sim \text{Inverse-Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 s_0^2}{2}\right)$
            3. $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$

4. $\tau^2 \sim \text{Inverse-Gamma}\left(\frac{a_0}{2}, \frac{b_0}{2}\right)$
2. Conditionals:
    1. Group means: $\theta_j \mid \mu, \tau^2, \sigma_j^2, y_j \sim \mathcal{N}\left(\frac{\frac{n_j \bar{y}_j}{\sigma_j^2} + \frac{\mu}{\tau^2}}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n_j}{\sigma_j^2} + \frac{1}{\tau^2}}\right)$
    2. Within-group variance: $\sigma_j^2 \mid y_j, \theta_j \sim \text{Inverse-Gamma}\left(\frac{\nu_0 + n_j}{2}, \frac{\nu_0 s_0^2 + \sum_{i=1}^{n_j}(y_{ij} - \theta_j)^2}{2}\right)$
    3. population mean: $\mu \mid \theta, \tau^2 \sim \mathcal{N}\left(\frac{\frac{J\bar{\theta}}{\tau^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{J}{\tau^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{J}{\tau^2} + \frac{1}{\sigma_0^2}}\right)$
    4. population variance:$\tau^2 \mid \theta, \mu \sim \text{Inverse-Gamma}\left(\frac{a_0 + J}{2}, \frac{b_0 + \sum_{j=1}^{J}(\theta_j - \mu)^2}{2}\right)$
3. Shrinkage: $E[\theta_j \mid \mu, \tau^2, \sigma_0^2, Y] = \left(\frac{n_j}{\sigma_j^2}\right)^{-1}\left(\frac{n_j}{\sigma_j^2}\bar{y}_j + \frac{1}{\tau^2}\mu\right)$

9. Mixed Effects Model
    1. $Y_{ij} = \mu + \delta_j + \epsilon_{ij}$
    2. Fixed effects + school level random effects + individual level rendom effects
    3. Parameters are not identifiable
    4. Model is identifiable with the addition of the prior distributions (conditional autoregressive) leading to full conditional, but this model has poor mixing

Linear Models

---

1. 12 people were recruited to assess the affects of a training program, where 6 attended and 6 didn't attend. How does a subject's change in maximal oxygen uptake depend on which program they were assigned to, as well as other factors such as age?
    1. Run a linear regression with a dummy variable for if the person was in the program or not and an interaction term for age and group
2. Linear models encompass ANOVA, ANCOVA, regression, random effects, and mixed effects models
    1. linear in predictors and predictors assumed to be error-free
    2. Assume n>p and the design matrix is linearly independent; rank(X)=p
    3. Using Matrix form OLS, M is the hat matrix and $(\mathbf{Y} - \mathbf{X}^T\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}^T\hat{\boldsymbol{\beta}}) = \mathbf{Y}^T(\mathbf{I} - \mathbf{M})\mathbf{Y}$ are the residuals
3. Semiconjugate Priors
    1. If $\vec{\beta} \sim MN(\beta_0, \Sigma_0) \propto \exp\left\{\beta^\top\left(\Sigma_0^{-1}\beta_0 + \mathbf{X}^\top\mathbf{Y}/\sigma^2\right) - \frac{1}{2}\beta^\top\left(\Sigma_0^{-1} + \mathbf{X}^\top\mathbf{X}/\sigma^2\right)\beta\right\}$
        1. Expectation: $\mathbb{E}(\beta \mid \mathbf{Y}, \sigma^2) = \left(\Sigma_0^{-1} + \mathbf{X}^\top\mathbf{X}/\sigma^2\right)^{-1}\left(\Sigma_0^{-1}\beta_0 + \mathbf{X}^\top\mathbf{Y}/\sigma^2\right)$
        2. Variance: $\text{Var}(\beta \mid \mathbf{Y}, \sigma^2) = \left(\Sigma_0^{-1} + \mathbf{X}^\top\mathbf{X}/\sigma^2\right)^{-1}$
        3. Model variance is Inverse Gamma

4. Noninformative Priors
    1. need to consider absolutely flat priors on β and $\log \sigma^2$: $\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$
    2. These are also Jeffrey Priors (none integrate to values)
    3. Improper priors; turns out the posteriors are still valid and can conduct valid statistical inference on them.
5. Conjugate Priors
    1. $\pi(\beta \mid \sigma^2) \sim \mathcal{N}(\mu_\beta, \sigma^2 V_\beta)$
    2. $\pi(\sigma^2) \sim \mathrm{IG}(a, b), \quad a, b > 0$
    3. where μ_β is a p-dimensional vector and V_β is a $p \times p$ positive definite symmetric matrix. We call this the **Normal-Inverse-Gamma (NIG)** prior and denote it as $\mathrm{NIG}(\mu_\beta, V_\beta, a, b)$, which is defined by the joint probability distribution of the vector β and the scalar σ^2
    4. PDF of multivariate t-density with ν degrees of freedom, mean μ and covariance V
        1. $f(t \mid \nu, \mu, V) = \frac{\Gamma(\frac{\nu+p}{2})\Gamma(\frac{\nu}{2})}{\pi^{p/2}\nu^{p/2}|V|^{1/2}}[1 + \frac{(t-\mu)^T V^{-1}(t-\mu)}{\nu}]^{-(\nu+p)/2}$
        2. The covariance matrix can be diagonal/block-diagonal/unstructured
6. g-Prior
    1. Motivation: parameter estimation should be invariant to changes in the scale of the regressors
    2. It can be obtained by setting, $\beta_0 = 0$,$\Sigma_0 = c(X^T X)^{-1}$, and $\pi(\sigma^2) \propto \sigma^{-2}$
    3. It places a MVN prior on β
    4. the "c" controls thye strength of the prior relative to the data
    5. larger c = more diffuse prior
    6. Zellner's g-Prior: $c = g\sigma^2$ choose g=n
7. Conditional posterior distribution: $p(\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y}) = \mathcal{N}((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}, (\mathbf{X}^\top \mathbf{X})^{-1}\sigma^2)$
8. Marginal posterior: $p(\sigma^2 \mid \mathbf{Y}) \propto \mathrm{IG}\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right)$
    1. where $s^2 = \hat{\sigma}^2 = \frac{1}{n-p}\left(\mathbf{Y} - \mathbf{X}^\top \hat{\boldsymbol{\beta}}\right)^\top \left(\mathbf{Y} - \mathbf{X}^\top \hat{\boldsymbol{\beta}}\right)$
    2. $\hat{\sigma}^2$ can be viewed as $\frac{(n-p)s^2}{\sigma^2} \sim \chi^2$ and $p(\beta|y)$ is a non-centralized t-distribution with n-p degrees of freedom
9. Gibbs Sampler
    1. Draw $\beta_{(t)} \sim \mathcal{N}((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}, (\mathbf{X}^\top \mathbf{X})^{-1}\sigma^2_{(t)})$
    2. Draw $\sigma^2_{(t)} \sim \mathrm{IG}(\frac{n-p}{2}\frac{(n-p)s^2_{(t)}}{2})$
10. Prediction
    1. We have a new covariate matrix and want to predict new vector of values
    2. For each posterior draw of $(\beta_{(t)}, \sigma^2_{(t)})_{t=1}^T$, draw $\mathbf{Y}^{(t)}_{\text{new}}$ from $\mathcal{N}(\mathbf{X}_{\text{new}}\beta_{(t)}, \sigma^2_{(t)}\mathbf{I})$

General Linear Models

1. GLMs encompass linear models as random component+systematic component+linear predictor
2. Generalization
    1. The distribution of y is from exponential distribution.
    2. The relationship between $\mu_i = \mathbb{E}(y_i \mid x_i, \beta)$ and η_i can be more general, $g(\mu_i) = \eta_i = x_i\beta$
    3. $g(\mu_i)$ is called the μ-link function: it links the mean of y_i to the linear predictor η_i.
    4. For example: canonical parameter θ in the exponential family: $p(y \mid \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$
    5. Distributions included in the exponential family are the normal, binomial, gamma, Poisson, beta, multinomial, and inverse Gaussian distributions
    6. Remark: $\theta = \log\left(\frac{p}{1-p}\right) \quad \Rightarrow \quad p = \frac{e^\theta}{1+e^\theta}$ is logit transformation
3. Estimation
    1. Classical estimation
        1. The MLE estimate of β does not have a closed analytic form in general for GLMs.
        2. Have to rely on iterative methods such as Newton-Raphson of Fisher scoring to obtain MLEs
        3. Large sample asymptotics
    2. Bayesian Estimation
        1. Possible priors on β same as that for linear models
            1. Non-informative
            2. Normal (Gaussian) priors
            3. g-prior
        2. Full conditionals for β are not in closed form
        3. Need to rely on approximation or sampling techniques for posterior inference
4. Noninformative Priors for β
    1. The uniform prior, i.e., $\pi(\beta) \propto 1$ is always an attractive noninformative prior for β, since it yields a posterior with parameters that match frequentist point estimates. Thus with $\pi(\beta) \propto 1$, the posterior of β is $\pi(\beta \mid y) \propto L(\beta)$, and in this case the posterior mode of β equals the MLE estimate of β.
    2. Moreover, by the Bayesian Central Limit Theorem, $\beta \mid y \to N(\hat{\beta}, I^{-1}(\hat{\beta}))$ as $n \to \infty$, where $I(\hat{\beta}) = X^T \Delta V \Delta X\big|_{\beta=\hat{\beta}}$ and $\hat{\beta}$ is the MLE of β, $\Delta = \frac{\partial\theta}{\partial\eta}$ and $V = b''(\theta)$.
5. Jeffrey Priors
    1. The Jeffreys prior for β is proportional to the square root of the determinant of the Fisher information matrix: $\pi(\beta) \propto |I(\beta)|^{\frac{1}{2}} = \left|X^T \Delta V \Delta X\right|^{\frac{1}{2}}$ (Δ and V are dependent on

β)
    2. For the Normal, Poisson and Gamma models, Jeffreys prior for β is improper for any link function
6. Bayesian Estimation (Continued)
    1. For GLMs, the posterior distribution of β does not have a closed form.
    2. Unlike linear models, the posterior p(β | y) in a GLM is not analytically tractable.
    3. Gibbs sampling is used to sample from p(β | y)
    4. Therefore, we must use Metropolis-Hastings within Gibbs or other numerical sampling techniques.
    5. For GLMs, Bayesian inference generally requires more sophisticated MCMC due to intractable conditionals

Bayesian Probit Regression Analysis

---

1. Suppose $Y \sim \text{Bernoulli}(p_i)$ where $p_i = H(x^T\beta)$
    1. If H is standard gaussian, it is probit
    2. If H is logistic CDF, it is logit
2. We need a simulation approach to get exact posterior
    1. Introduce N latent variables such that $z = \eta + \epsilon$ where $\epsilon \sim N(0,1)$ and $Z \sim N(x^T\beta, 1)$
    2. This makes y an indicator variable so $p(y = 1|\beta) = p(z - \eta > -\eta|\beta) = \Phi(\eta)$. The marginal distribution of y must be kept the same
    3. This approach connects probit binary regression model on Y to normal linear regression model on Z
    4. The sampling approach allows us to compute posterior of many parameters
    5. The Bayesian residual is continuous so provides more information about outliers
3. Hierarchal Formulation
    1. Data: $\{y_i, x_i\}_1^n$ and $z = (z_1, \ldots, z_n)$
    2. $p(y_i \mid z_i) \sim I(z_i > 0), \delta_1$, Indicator with delta function
    3. $p(z_i \mid \beta) \sim N(x_i, \beta, 1)$
    4. $p(\beta) \sim N(\mu_\beta, V_\beta)$
        1. Can assume other priors on β like non-informative or g-priors.
4. Gibbs sampling for binary data
    1. $\pi(\beta, \mathbf{Z} \mid \mathbf{y}) = C\pi(\beta) \prod_{i=1}^N \{\mathbb{1}(Z_i > 0)\mathbb{1}(y_i = 1) + \mathbb{1}(Z_i \leq 0)\mathbb{1}(y_i = 0)\} \cdot \phi(Z_i; \mathbf{x}_i^\top \beta, 1)$
    2. Gibbs Sampler
        1. $p(z_i \mid y_i = 1, \beta) = \mathbb{I}(z > 0) \cdot \mathcal{N}(\mathbf{x}_i^\top \beta, \sigma^2)$
        2. $p(z_i \mid y_i = 0, \beta) = \mathbb{I}(z < 0) \cdot \mathcal{N}(\mathbf{x}_i^\top \beta, \sigma^2)$

        3. Assume $\sigma^2 = 1$ for probit case
5. Gibbs Sampling for truncated normal $Z \sim N(\mu, \sigma^2) \cdot I(a < z < b)$
    1. Setting $u_1 = \Phi(a;\, \mu, \sigma^2)$ and $u_2 = \Phi(b;\, \mu, \sigma^2)$
    2. Sampling $u \sim \mathcal{U}(u_1, u_2)$
    3. Setting $z = \Phi^{-1}(u;\, \mu, \sigma^2)$ (Inverse Phi is quantile function)
6. The t-link
    1. Generalize probit link by choosing H as t-distribution.
    2. This helps in investigating the sensitivity of fitted probability to the choice of link function
    3. Most popular link function for binary data is logit
    4. Logistic distribution is a member of t family with approximately 9 d
    5. If Z_i follows a t distribution with location $x_i^T\beta$, scale parameter 1, and degrees of freedom ν, it is equivalent to:
        1. $Z_i \mid \lambda_i \sim N(x_i^T\beta, \lambda_i^{-1})$
        2. $\lambda_i \sim \mathrm{Gamma}(\nu/2, 2/\nu)$ with PDF proportional to $\lambda_i^{\nu/2-1}\exp(-\nu\lambda_i/2)$

Bayesian Data Augmentation

---

1. Introduction
    1. Missing data are common! Such as survey data:
        1. miss a page of survey
        2. a doctor forgets to record
        3. patients are hesitant to report
    2. Important in observational and experimental research
2. Types of Missing Data
    1. Missing Completely At Random (MCAR): Missingness does not depend on observed or unobserved data, such as $\mathrm{logit}(p_{iw}) = \theta_0$.
    2. Missing At Random (MAR): Missingness depends only on observed data, such as $\mathrm{logit}(p_{iw}) = \theta_0 + \theta_1 t_i$.
    3. Missing Not At Random (MNAR):
        1. Missing value depends on the hypothetical value (e.g., people with high salaries generally do not want to reveal their incomes in surveys).
        2. Missing value is dependent on some other variable's value (e.g., females generally do not want to reveal their ages — here the missing value in the age variable is impacted by the gender variable).
3. Popular Methods

1. Complete Case Analysis
    1. Missing data are ignored and only complete cases analyzed
    2. Many R packages by default
    3. Advantage: simple
    4. Disadvantage: introduce bias and inefficient
2. Single Imputation
    1. Aim to “fill in” missing values to create single “completed” (imputed) dataset
    2. Mean imputation
    3. last observation carried forward (LOCF)
        1. Widely used in clinical trial settings
        2. Assumption: all unseen measurements = last seen measurement
        3. Computationally simple but makes strong assumptions about missingness mechanism (usually MCAR)
    4. Ad hoc methods are frequently used, but not recommended
    5. Ignores uncertainty about imputed missing values
4. Statistical Methods
    1. Suppose Y_{obs}, Y_{mis}, M are observed data, missing data, and missing indicator, respectively.
    2. Let ϕ be the set of parameters that determine the probability of missing in addition to the observed data: $p(M \mid Y_{obs}, \phi)$.
    3. When $m_i = 1$, the joint likelihood of $(x_i, y_i, m_i = 1)$ is $p(x_i, y_{obs,i}, m_i = 1; \theta, \phi) = p(m_i = 1 \mid x_i, y_{obs,i}; \phi), p(y_{obs,i} \mid x_i; \theta), p(x_i)$
    4. When $m_i = 0$, the joint likelihood of $(x_i, m_i = 0)$ is $p(x_i, m_i = 0; \theta, \phi) = \int p(m_i = 0 \mid x_i, y_{mis,i}; \phi), p(y_{mis,i} \mid x_i; \theta), p(x_i), dy_{mis,i}$
    5. Missingness in case of MCAR is ignorable
    6. Multiple imputation
        1. Multiple imputation (MI) – generate K > 1 imputed values for missing observations from appropriate probability distribution
        2. Impute the missing entries of the incomplete data sets, not once, but K (typically 5-10) times
        3. Imputed values are drawn from a distribution (that can be different for each missing variable)
        4. This step results in K complete data sets
        5. Analyze each of the K completed data sets using standard methods
        6. Combine the K analysis results into a single final result
        7. Simple rules exist for combining the K analyses to produce estimates and confidence intervals that incorporate uncertainty about the missing data

7. Fully model based (e.g. Bayesian)- write down statistical model for full data (including missingness mechanism) and base analysis on this model
    1. based on a well-defined statistical model for the complete data, and explicit assumptions about the missing value mechanism
    2. the subsequent analysis, inferences and conclusions are valid under these assumptions
    3. doesn't mean the assumptions are necessarily true but it does allow the dependence of the conclusions on these assumptions to be investigated

5. Discussion
    1. Missing response data is trivial to handle in the Bayesian framework under the assumption of an ignorable missing data mechanism.
    2. Equivalent to posterior prediction from the model fitted to the observed data
    3. The most appropriate way of modeling the missing data indicator will be problem specific
    4. Some elaboration may be required to accommodate different types of drop-out (e.g. death and recovery)
    5. Some datasets may have informative drop-in (e.g. medics start to monitor patients if they become more unwell)
    6. For longitudinal studies with drop-out, an alternative is to replace the missingness indicator with a variable representing the time to drop-out and model this using (discrete or continuous time) survival techniques