Bayesian Inference and Decision

Instructions

1. The student is allowed two, double-sided note sheets
2. Previously proved results should be clearly cited and properly invoked

1. Suppose y_i, for $i = 1, \ldots, n$, is a binary outcome defined by the following representation:

$$y_i = \begin{cases} 0 & z_i < 0 \\ 1 & z_i \geq 0 \end{cases}$$

where $z_i \sim N(x_i'\beta, \lambda^{-1})$ for the $p \times 1$ vectors of covariates x_i and coefficients β. Using the joint prior $\pi(\beta, \lambda) \propto \lambda^{-1}$, answer the following.

(a) Determine the likelihood and the posterior.

Note. $y_i \sim \text{Bernouli}(P_i)$ st $P_i = P(Z_i \geq 0) = \int_{-\infty}^{0} f(z_i \mid x_i, \beta, \lambda^{-1}) dz_i$

$\vec{Z} \sim MVN(X\vec{\beta}, \lambda^{-1} I_{n \times n})$ $\pi(\vec{\beta}, \lambda) = \lambda^{-1}$

Let $\theta = \lambda^{-1}$

$$\mathcal{L}(\vec{y} \mid \vec{p}) = \prod_{i=1}^{n} P_i$$

$$P(\vec{P}, \vec{Z}, \vec{\beta}, \theta) \propto \left[\prod_{i=1}^{n} P_i\right] \cdot \left[\theta^{-n/2} \exp\left\{\frac{-1}{2\theta}(\vec{Z} - X\vec{\beta})^T(\vec{Z} - X\vec{\beta})\right\}\right] \cdot \theta$$

$$= \left[\prod_{i=1}^{n} P_i\right] \cdot \left[\theta^{-n/2+1} \exp\left\{\frac{-1}{2\theta}(\vec{Z} - X\vec{\beta})^T(Z - X\vec{\beta})\right\}\right]$$

$$= \left[\prod_{i=1}^{n} P_i\right] \cdot \left[\lambda^{\frac{n}{2}-1} \exp\left\{\frac{-\lambda}{2}(\vec{Z} - X\vec{\beta})^T(Z - X\vec{\beta})\right\}\right]$$

(b) Find the full conditionals for the binary model and determine the best sampling strategy.

$$P(\lambda \mid \vec{\beta}, \vec{z}, \vec{p}) \propto \lambda^{\frac{n}{2}-1} \exp\left\{ \frac{-\lambda}{2} (\vec{z} - X\vec{\beta})^T (\vec{z} - X\vec{\beta}) \right\}$$

$$\Rightarrow \lambda \mid \vec{\beta}, \vec{z}, \vec{p} \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2} (\vec{z} - X\vec{\beta})^T (\vec{z} - X\vec{\beta}) \right)$$

$$P(\vec{\beta} \mid \lambda, \vec{z}, \vec{p}) \propto \exp\left\{ \frac{-\lambda}{2} (\vec{z} - X\vec{\beta})^T (\vec{z} - X\vec{\beta}) \right\}$$

$$= \exp\left\{ \frac{-1}{2\theta} (\vec{z} - X\vec{\beta})^T (\vec{z} - X\vec{\beta}) \right\}$$

$$\Rightarrow \vec{\beta} \mid \lambda, \vec{z}, \vec{p} \sim MVN\left((X^T X)^{-1} X^T \vec{z}, \frac{1}{\lambda} (X^T X)^{-1} \right)$$ (From board example derived in class)

$$P(\vec{p} \mid \lambda, \beta, \vec{z}) = \prod_{i=1}^{n} \int_{-\infty}^{0} f(z_i \mid x_i, \vec{\beta}, \lambda)\, dz_i$$

$\vec{p}$'s conditional distribution is simply an integral over a Normal distribution.

I would draw samples from Gamma and MVN for λ and $\vec{\beta}$ respectively, compute $\vec{z}$ to derive an emperical distribution of estimates for $\vec{p}$

(c) Now suppose that y_i is an ordinal outcome with three levels defined by the following latent variable representation:

$$y_i = \begin{cases} 0 & c_0 < w_i < c_1 \\ 1 & c_1 \le w_i < c_2 \\ 2 & c_2 \le w_i < c_3 \end{cases}$$

where $w_i \sim N(x_i'\phi, \eta^{-1})$, $c_0 = -\infty$, and $c_3 = \infty$. Place the ridge prior on ϕ and let $\pi(\eta) \propto \eta^{-1}$. Derive the likelihood and find the posterior.

Following from previous, I will ony proceed w/ deriving for w_i, ϕ, η as the same stratgy for y_i and $\vec{p}$ holds. In other words, the notation is more cumbersome, but the strategy is the same.

Let $\theta = \eta^{-1}$ and $\vec{\phi} \sim MVN(\vec{\mu}, \sigma I_{n\times n})$

$$\mathcal{L}(\vec{w} \mid \phi, \eta) \propto \left[\theta^{-n/2} \exp\left\{\frac{-1}{2\theta}(\vec{w} - X\vec{\phi})^T(\vec{w} - X\vec{\phi})\right\}\right]$$

$$P(\vec{\phi}, \eta \mid \vec{w}, \vec{x}) \propto \theta^{-\frac{n}{2}+1} \exp\left\{\frac{-1}{2\theta}(\vec{w} - X\vec{\phi})^T(\vec{w} - X\vec{\phi})\right\} \exp\left\{\frac{-1}{2\sigma}(\vec{\phi} - \vec{\mu})^T(\vec{\phi} - \vec{\mu})\right\}$$

$$= \eta^{\frac{n}{2}-1} \exp\left\{\frac{-\eta}{2}(\vec{w} - X\vec{\phi})^T(\vec{w} - X\vec{\phi}) + \frac{-1}{2\sigma}(\vec{\phi} - \vec{\mu})^T(\vec{\phi} - \vec{\mu})\right\}$$

(d) Find the full conditionals for the ordinal model and determine a sampling strategy.

$$P(\eta \mid \vec{\phi}, \vec{w}, X) \propto \eta^{n/2 - 1} \exp\left\{ \frac{-\eta}{2} (\vec{w} - X\vec{\phi})^T (\vec{w} - X\vec{\phi}) \right\}$$

$$\eta \sim \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}(\vec{w} - X\vec{\phi})^T(\vec{w} - X\vec{\phi})\right)$$

$$P(\vec{\phi} \mid \eta, \vec{w}, X) \propto \exp\left\{ \frac{-1}{2\sigma} (\vec{w} - X\vec{\phi})^T (\vec{w} - X\vec{\phi}) + \frac{-1}{2\sigma} (\vec{\phi} - \vec{\mu})^T (\vec{\phi} - \vec{\mu}) \right\}$$

$$= \exp\{$$

For time sake: Multiply out above. Drop unnecessary params. Complete the square and you'll find

$$\vec{\phi} \sim MVN(A, B)$$

Strategy is as before

Sample $\vec{\phi}$ and η from their conditional distributions, For each sample, calculate the latent variable w_i and get an emperical distribution of $\vec{P}_i$, the categorical parameter for each y_i

2. Let $\boldsymbol{\theta}_i \sim Multinom(\boldsymbol{\alpha}_i)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, J, J > 2$, total categories.

(a) Find the conjugate prior for $\boldsymbol{\alpha}_i$.

If $\theta_i \sim \text{Multinom}(\alpha_i)$, we know from lecture Notes that the Dirichlet prior is conjugate. Specifically, the posterior with a mutinomial likelihood and Dirichlet prior is itself Dirichelt. See below for Derivation.

$\vec{\theta}_i \sim \text{Multinomial}(\vec{\alpha}_i)$ $\quad \vec{\alpha}_i \sim \text{Dir}(\vec{\beta})$

$$\mathcal{L}(\vec{\theta}_i \mid \vec{\alpha}_i) \propto \prod_{j=1}^{J} \alpha_{ij}^{\theta_{ij}} \qquad \pi(\vec{\alpha}_i) \propto \prod_{j=1}^{J} \alpha_{ij}^{\beta_j - 1}$$

$$\Rightarrow P(\vec{\alpha}_i \mid \vec{\theta}_i, \vec{\beta}) \propto \left[\prod_{j=1}^{J} \alpha_{ij}^{\theta_{ij}}\right] \cdot \left[\prod_{j=1}^{J} \alpha_{ij}^{\beta_j - 1}\right]$$

$$= \prod_{j=1}^{J} \alpha_{ij}^{\theta_{ij} + \beta_j - 1}$$

which indicates that $\vec{\alpha}_i \sim \text{Dir}(\vec{\theta}_i + \vec{\beta})$

Note: $\mathcal{L}(\Theta \mid \alpha) = \prod_{i=1}^{n} \mathcal{L}(\vec{\theta}_i \mid \vec{\alpha}_i)$ and

$P(\alpha \mid \Theta, \vec{\beta}) = \prod_{i=1}^{n} P(\vec{\alpha}_i \mid \vec{\theta}_i, \vec{\beta})$ does not change the distribution of each $\vec{\alpha}_i$

(b) Define a transformation of $\boldsymbol{\alpha}_i$, call it γ_i, such that $\gamma_i \overset{iid}{\sim} MVN\,(\boldsymbol{\mu}, \sigma^2 I_{J\times J})$ is a valid prior; where $\boldsymbol{\mu}$ is a $J \times 1$ vector of population means and $I_{J\times J}$ is the $J \times J$ identity matrix. Determine the resulting posterior for this model.

Let $F(\vec{\alpha}_i) = \vec{\gamma}_i$ S.T. $\vec{\gamma}_i \overset{iid}{\sim} MVN(\vec{\mu}, \sigma^2 I_{J\times J})$

Then $\vec{\theta}_i \sim \text{Multinom}(F^{-1}(\vec{\gamma}_i))$

$$\mathcal{L}(\vec{\theta}_i \mid \vec{\alpha}_i) \propto \prod_{j=1}^{J} F^{-1}(\gamma_{ij})^{\theta_{ij}}$$

Note that since $\vec{\gamma}_i \overset{iid}{\sim}$ we can just say $\vec{\gamma}$

$$\Rightarrow \mathcal{L}(\vec{\theta}_i \mid \vec{\gamma}) \propto \prod_{j=1}^{J} F^{-1}(\gamma_j)^{\theta_{ij}}$$

$$* \; P(\vec{\gamma} \mid \vec{\theta}_i) \propto \prod_{j=1}^{J} \left[F^{-1}(\gamma_j)^{\theta_{ij}} \exp\left\{ \frac{-1}{2\sigma^2} (\gamma_j - \mu_j)^2 \right\} \right]$$

Without knowing the form of F (and thus F^{-1}) this is an unknown form

(c) Develop a strategy for drawing samples from the posterior in (b).

I would use a Metropolis Hastings Sampler to sample $\vec{\gamma}$. Noting that the posterior is of the exponential family, I'd start with either an Normal or Gamma candidate distribution. The Normal looks like what's in the exp{ }, but the Gamma has an exponent on the outside parameter a-la $F^{-1}(\gamma_{ij})^{\theta_{ij}}$.

You'd want to tune the sampler so you accept between 30% to 70% of draws. Of course, one should burn-in iterations before convergence. Thinning may be necessary if each draw still seems autocorrelated.

(d) Explain how to make the prior on γ_i informative as well as how to make it weakly informative.

The prior on $\vec{\gamma}_i$ is MVN. To tune how informative it is, one can tune σ^2. A larger σ^2 would diffuse the information around any particular μ_j. A smaller σ^2 would make it more informative