

Statistics 606 Problem Set 4

Due April 24

1. Recall the model for a two-way layout with random interactions from problem set 3. Devise and implement a Gibbs sampling algorithm to estimate the model parameters in a Bayesian sense using posterior means. Use flat (improper) priors on the $\{\alpha_i\}$, $\{\beta_j\}$, σ^2 and τ^2 parameters.

You do not need to implement this algorithm. Just derive and clearly explain the steps of the Gibbs sampler.

Hint: At one point you will need to sample from a density of the form

$$P(Z_1, \dots, Z_m | \sum_j Z_j = 0),$$

where the unconditional density $P(Z_1, \dots, Z_m)$ is normal with mean vector θ and variance $\eta^2 I$, where $\eta^2 > 0$ is a scalar. You can do this by sampling sequentially using the fact that

$$P(Z_1, \dots, Z_m | \sum_j Z_j = 0) = P(Z_1 | \sum_j Z_j = 0) P(Z_2 | Z_1, \sum_j Z_j = 0) \cdots P(Z_m | Z_1, \dots, Z_{m-1}, \sum_j Z_j = 0).$$

To simulate the k^{th} term, note that

$$P(Z_k | Z_1, \dots, Z_{k-1}, \sum_j Z_j = 0) \propto P(\sum_j Z_j = 0 | Z_1, \dots, Z_k) P(Z_k),$$

and

$$P(\sum_j Z_j = 0 | Z_1, \dots, Z_k) = P(\sum_{j=k+1}^m Z_j = -\sum_{j=1}^k Z_j | Z_1, \dots, Z_k).$$

The density of $\sum_{j=k+1}^m Z_j$ is $N(\mu_k, (m-k)\eta^2)$, where $\mu_k = \sum_{j=k+1}^m \theta_k$.

$$P(Z_k | Z_1, \dots, Z_{k-1}, \sum_j Z_j = 0) \propto \exp\left(-\frac{1}{2(m-k)\eta^2}(\sum_{j=1}^k Z_j + \mu_k)^2\right) \exp\left(-\frac{1}{2\eta^2}(Z_k - \theta_k)^2\right).$$

With some algebra you will see that $Z_k | Z_1, \dots, Z_{k-1}, \sum_j Z_j = 0$ is normal with variance

$$\frac{\eta^2}{1 + 1/(m - k)}$$

and mean

$$\frac{\theta_k(m - k) - \sum_{j=1}^{k-1} Z_j - \mu_k}{m - k + 1}.$$

Note: the mean was wrong by a factor of -1 when I originally posted this.

More on this point (added 4/19): Since the Z_i are Gaussian, there is an alternative approach. Consider the multivariate Gaussian random vector

$$(Z_1, \dots, Z_m, \sum_j Z_j)',$$

which has mean

$$\tilde{\theta} = (\theta_1, \dots, \theta_m, \sum_j \theta_j)$$

and variance

$$\tilde{\Sigma} = \left(\begin{array}{ccc|ccc} & & & \sigma^2 & & \\ & & & \sigma^2 & & \\ & & \sigma^2 I_{m \times m} & \sigma^2 & & \\ & & & \sigma^2 & & \\ & & & \sigma^2 & & \\ \hline \sigma^2 & \sigma^2 & \dots & \sigma^2 & & m\sigma^2 \end{array} \right).$$

From the properties of the multivariate normal distribution, it follows that $P(X_1, \dots, X_m | \sum_j X_j = 0)$ is multivariate normal with mean

$$(\theta_1 - \sum_j \theta_j/m, \dots, \theta_m - \sum_j \theta_j/m)'$$

and variance

$$\sigma^2 I_{m \times m} - \sigma^2/m.$$

You can check directly that this is the same distribution as that of the vector

$$(Z_1 - \bar{Z}, \dots, Z_m - \bar{Z})'.$$

Thus it is correct to sample the Z_j 's from $N(\theta, \sigma^2 I)$ and then center them.

2. Suppose we observe a 2×2 contingency table:

$$\begin{array}{cc} N_{11} & N_{12} \\ N_{21} & N_{22} \end{array}$$

from an underlying distribution in which each observation falls into a cell according to the following probabilities

$$\begin{array}{cc} p_{11} & p_{12} \\ p_{21} & p_{22} \end{array}$$

We parameterize the underlying probability distribution in terms of the marginal row probabilities p (row 1) and $1 - p$ (row 2), and the marginal column probabilities q (column 1) and $1 - q$ (column 2). To fully specify the joint distribution we also have a parameter λ for the log-odds ratio

$$\lambda = \log p_{11} + \log p_{22} - \log p_{12} - \log p_{21}.$$

- (a) To construct a map between the parameters (p, q, λ) and the cell probabilities $(p_{11}, p_{12}, p_{21}, p_{22})$, write the cell probabilities in the form

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} pq + \delta & p(1 - q) - \delta \\ (1 - p)q - \delta & (1 - p)(1 - q) + \delta \end{pmatrix}$$

Show that if $0 < p, q < 1$, then for each λ there exists a unique δ such that the corresponding p_{ij} cell probabilities have odds ratio λ . Implement a numerical algorithm for finding δ for specified values of p, q, λ .

- (b) Suppose we put independent uniform priors on the model parameters p, q and λ . Use the Metropolis-Hastings algorithm to generate a sample from the posterior distribution of the parameter estimates given the data.
- (c) Run your procedure from part (ii) on the data

$$\begin{array}{cc} 3 & 1 \\ 1 & 3 \end{array}$$

based on the posterior distribution of the odds ratio, what do you conclude?