Posterior Predictive Inference

Sudipto Banerjee

sudiptob@biostat.umn.edu

University of Minnesota

Posterior Predictive Checks

- Assume modelling data as $\mathbf{y} = (y_1, ..., y_n) \sim N(\mu, \sigma^2)$
- Set priors on μ and σ^2
- Run WinBUGS and obtain samples: $\theta_t = \{\mu_t, \sigma_t^2\}$, $t = 1, \ldots, M$
- For each sampled data point θ_t, replicate n data points: $y^t_{rep,i} \sim N(\mu_t, \sigma_t^2)$, $t = 1, \ldots, M$ and $i = 1, \ldots, n$.
- For each sampled value, (μ_t, σ_t^2), we obtain M *replicated data set* $\mathbf{y}^t_{rep} = (y^t_{rep,1}, \ldots, y^t_{rep,n})$.
- Does our model represent our data adequately? Choose a discrepancy measure, say

$$T(\mathbf{y}; \theta) = \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2}$$

Compute $T(\mathbf{y},\ \theta_t)$ and the set of of $T(\mathbf{y}^t_{rep},\ \theta_t)$ and obtain "Bayesian p-values":

$$P(T(\mathbf{y}_{rep},\ \theta) > T(\mathbf{y},\ \theta) \,|\, \mathbf{y}) = \frac{1}{M} \sum_{t=1}^{M} 1[T(\mathbf{y}^t_{rep},\ \theta_t) > T(\mathbf{y},\ \theta_t)].$$

Model Comparisons

- Compute the posterior predictive mean and variance for each observation:

$$\mu_{rep,i} = E[Y_{rep,i}|\mathbf{y}] = \frac{1}{M}\sum_{t=1}^{M} y_{rep,i}^{(t)}, \; i = 1, \ldots, n;$$

$$\sigma_{rep,i}^2 = Var[Y_{rep,i}|\mathbf{y}] = \frac{1}{M}\sum_{t=1}^{M} (y_{rep,i}^{t} - \mu_{rep,i})^2.$$

- Goodness of fit (lower values better):

$$G = \sum_{i=1}^{n} (y_i - \mu_{rep,i})^2$$

- Penalize predictive variance (lower values better):

$$P = \sum_{i=1}^{n} \sigma_{rep,i}^2$$

- Model Comparison Metric (lower values better): $D = G + P$