1 Hierarchical models: data-analysis problems

1.1 Math tests

1.

2. Rewrite the distributions in terms of precision

$$(y_{ij}|\theta_i, \omega) \sim N(\theta_i, (\omega)^{-1})$$

$$(\theta_i|\omega, \lambda) \sim N(\mu, (\omega\lambda)^{-1})$$

Choose prior for parameters

$$\omega \sim \Gamma(\frac{d}{2}, \frac{\eta}{2})$$

$$\lambda \sim \Gamma(\frac{h}{2}, \frac{h}{2})$$

The joint distribution of everything is, suppose n data are grouped into m groups:

$$\text{constant} \times \omega^{n+m}\lambda^m e^{-\frac{\omega\sum_{ij}(y_{ij}-\theta_i)^2}{2}-\frac{\omega\lambda\sum_i(\theta_i-\mu)^2}{2}-\frac{\omega\eta}{2}-\frac{h\lambda}{2}}\omega^{\frac{d}{2}-1}\lambda^{\frac{h}{2}-1}$$

From which we have, suppose each group have g_i elements

$$(\omega|y, \lambda, \theta, \mu) \sim \Gamma(\frac{d+n+m}{2}, \frac{\eta}{2} + \frac{\sum_{ij}(y_{ij}-\theta_i)^2}{2} + \frac{\lambda\sum_i(\theta_i-\mu)^2}{2})$$

$$(\lambda|y, \omega, \theta, \mu) \sim \Gamma(\frac{h+m}{2}, \frac{h}{2} + \frac{\omega\sum_i(\theta_i-\mu)^2}{2})$$

$$(\theta_i|y, \omega, \lambda, \mu) \sim N(\frac{\lambda\mu + \sum_j y_{ij}}{\lambda + g_i}, \frac{1}{\omega(\lambda + g_i)})$$

$$(\mu|y, \omega, \lambda, \theta) \sim N(\frac{\sum_i \theta_i}{m}, \frac{1}{m\omega\lambda})$$

We will update the parameter according to these distribution. For the code, see mathtest.r.

3. See "mathtest\mathtest.r"

1.2 Price elasticity of demand

Let $y = \log V$, $x = (1, \log P, d, d\log P)$. Where $d = 0, 1$ is an indicator whether a particular store display ads or not. Each data point are indexed by the store indices $i = 1, ..., n$ and time index $t = 1, ..., p_i$. Each x and β are four vectors whose vector index will be suppresed. We want to build the model:

$$y_{it} = x_{it}\beta_i + \epsilon_i$$

we have the prior distribution

$$\epsilon_i \sim N(0, \sigma^2)$$

$$\beta_i \sim N(\mu, \Sigma)$$

$$(\mu, \Sigma) \sim NIW(m, \lambda, \psi, \nu)$$

$$\sigma^2 \sim U(0, \infty)$$

from which we have

$$(y_{it}, \beta_i, \Sigma, \mu, \sigma^2) \sim N(X_i\beta_i, \sigma^2 I)N(\mu, \Sigma)NIW(m, \lambda, \Psi, \nu)$$

The conditional distributions are:

$$(\beta|y_{it}, \mu, \Sigma, \sigma^2) \sim N(\mu_i', \Sigma_i')$$

where

$$\mu_i' = \Sigma_i'(\frac{X_i^T Y_i}{\sigma^2} + \Sigma^{-1}\mu)$$

$$\Sigma_i'^{-1} = \frac{X_i^T X_i}{\sigma^2} + \Sigma^{-1}$$

$$(\mu, \Sigma|\beta_i) \sim NIW(m', \lambda', \Psi', \nu')$$

where

$$m' = \frac{\lambda m + \sum \beta_i}{\lambda + n}$$

$$\lambda' = \lambda + n$$

$$\Psi' = \Psi + \sum_i (\beta_i - \bar{\beta})(\beta_i - \bar{\beta})^T + \frac{\lambda n}{\lambda + n}\sum_i (\beta_i - \mu)(\beta_i - \mu)^T$$

$$\nu' = \nu + n$$

$$(\sigma^{-2}|y_{it}, \beta_i) \sim \Gamma(\frac{s}{2} + 1, \sum_i \frac{(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)}{2})$$

s is the total number of observations.

For details, see "cheese\cheese.r"

1.3 A hierachical probit model with data augmentation

Let $i = 1, ..m$ denote quantities belongs to a certain state, $j = 1, ...g_i$ the number of samples within a given state. $n = \sum_i g_i$ the total data points. β is a f dimensional vector indicating f factors.

$$(z_{ij}|\mu_i, \beta) \sim \begin{cases} N_+(\mu_i + x_{ij}^T\beta, \sigma^2) & y_{ij} = 1, \\ N_-(\mu_i + x_{ij}^T\beta, \sigma^2) & y_{ij} = 0 \end{cases}$$

$$(\mu_i|\mu_0, \lambda) \sim N(\mu, \lambda^2)$$

$$(\lambda, \mu_0) \sim NIG(m_0, \frac{d_0}{2}, \frac{\eta_0}{2})$$

$$\beta \sim N(0, \tau^2 I)$$

$$\tau^2, \sigma^2 \sim U(0, \infty)$$

Besides z, we have these updates for gibbs sampler.

$$(\lambda^2, \mu_0 | z_{ij}, \mu_i) \sim NIG(\frac{1}{m}\sum \mu_i, \frac{d+m}{2}, \frac{\eta + \sum_i (\mu_i - \mu_0)^2}{2})$$

$$(\mu_i | \beta, z_{ij}, \lambda^2, \mu, \sigma^2) \sim N(\frac{\sigma^2 \mu + \lambda^2 \sum_j (z_{ij} - x_{ij}^T \beta)}{\sigma^2 + \lambda^2 g_i}, \frac{\lambda^2 \sigma^2}{\sigma^2 + \lambda^2 g_i})$$

$$(\beta | \mu_i, z_{ij}, \lambda^2, \sigma^2, \tau^2) \sim N(\frac{1}{\sigma^2} VX(z_{ij} - \mu_i), V)$$

where

$$V^{-1} = \frac{1}{\tau^2} I + \frac{1}{\sigma^2} X^T X$$

$$(\sigma^{-2} | z_{ij}, \beta, \mu_i) \sim \Gamma(\frac{n}{2}, \frac{\sum_{ij}(z_{ij} - \mu_i - X\beta)^2}{2})$$

$$(\tau^{-2} | \beta) \sim \Gamma(\frac{f}{2}, \frac{\beta^2}{2})$$

2 Gene expression over time

$$y_{ijrt} = f_{it} + h_{ijt} + \epsilon + \tau_{ij}$$

Here $i = 1 \cdots 3$ stand for groups, $j = 1 \cdots n_i$ stands for genes within each group, sometimes will use $k = 1 \cdots 14$ to label all genes without specifying gropus. $r = 1, 2, 3$ stands for replicat of a gene.

$$\epsilon \sim N(0, \sigma_\epsilon^2)$$

$$\tau_k \sim N(0, \sigma_{\tau k}^2)$$

$$f_{it} \sim N(m_f, \sigma_{fi}^2 C)$$

$$h_{kt} \sim N(m_h, \sigma_{hk}^2 C)$$

$$\sigma_{\tau k}^2, \sigma_\epsilon^2, \sigma_{fi}^2, \sigma_{hk}^2 \sim \text{Inverse}$$

C is a 12×12 matrix computed from Mater(5/2) covariance function with $\tau_2 = 0$ and $b = 1$

$$C_{ij} = e^{-\frac{(i-j)^2}{2}}$$

The marginal distribution of all paramters are

$$\prod_k (\sigma_{\tau k}^{-2-3}\sigma_{hk}^{-2-3}) \prod_i \sigma_{fi}^{-2-n_i}\sigma_\epsilon^{-2-n}$$
$$\times e^{-\frac{1}{2}\sum_{kr}(y-f-h)^2(\frac{1}{\sigma_\epsilon}+\frac{1}{\sigma_{\tau k}^2})-\frac{1}{2}\sum_i \frac{1}{\sigma_{fi}^2}(f_i-m_f)^T C^{-1}(f_i-m_f)-\frac{1}{2}\sum_k \frac{1}{\sigma_{hk}^2}(h_k-m_h)^T C^{-1}(h_k-m_h)} \tag{1}$$

From which we have posterior distribution

$$(h_k|\cdots) \sim N(V_{hk}(\frac{\sigma_\epsilon^2+\sigma_{\tau k}^2}{\sigma_\epsilon^2\sigma_{\tau k}^2}\sum_r(y_{kr}-f_k)+\frac{1}{\sigma_{hk}^2}C^{-1}m_h), V_{hk})$$

$$V_{hk}^{-1} = 3\frac{\sigma_\epsilon^2+\sigma_{\tau k}^2}{\sigma_\epsilon^2\sigma_{\tau k}^2}+\frac{1}{\sigma_{hk}^2}C^{-1}$$

$$(m_h|\cdots) \sim N(\bar{h}, \frac{\sigma_{hk}^2}{14}C)$$

$$(f_i|\cdots) \sim N(V_{fi}(\sum_{jr}(\frac{1}{\sigma_\epsilon^2}+\frac{1}{\sigma_{\tau j}^2})(y_{ijr}-h_{ij})+\frac{1}{\sigma_{fi}^2}C^{-1}m_f), V_{fi})$$

$$V_{fi}^{-1} = 3(\frac{n_i}{\sigma_\epsilon^2}+\sum_j \frac{1}{\sigma_{\tau j}^2})+\frac{1}{\sigma_{ni}^2}C^{-1}$$

$$(m_f|\cdots) \sim N(\bar{f}, \frac{\sigma_{fi}^2}{3}C)$$

$$(\sigma_\epsilon^{-2}|\cdots) \sim \Gamma(\frac{n}{2}+2, \frac{\text{RMS of all data}}{2})$$

$$(\sigma_{\tau k}^{-2}|\cdots) \sim \Gamma(\frac{7}{2}, \frac{\text{RMS of gene k}}{2})$$

$$(\sigma_{hk}^{-2}|\cdots) \sim \Gamma(\frac{7}{2}, \frac{(h_k-m_h)^T C^{-1}(h_k-m_h)}{2})$$

$$(\sigma_{fi}^{-2}|\cdots) \sim \Gamma(\frac{n_i}{2}+2, \frac{(f_i-m_f)^T C^{-1}(f_i-m_f)}{2})$$

Given the gnenerated sample of the model, we will predict the evolution of gene on the whole time line by method based on the problem in excercise 3. For an unsampled time point $t^* \neq 1, \cdots, 12$. Its prediction will be given by

$$y(t^*)_{ij} = C_*^T C^{-1}(f_i^{est}+h_{ij}^{est}) \pm 1.96\sqrt{(\sigma_h^{est2}+\sigma_f^{est2})(C_{**}-C_*^T C^{-1}C_*)+\sigma_\epsilon^{est2}+\sigma_{\tau ij}^{est2}}$$

Estimations of f and h are the average from sample. Estimations of σ's are computed from the inverse average. C_* and C_**'s definition is the same as that in excercise 3.