Introduction to Bayesian Statistics

Shin-ichi Mayekawa

mayekawa@hum.titech.ac.jp
www.ms.hum.titech.ac.jp

Outline

1. Introduction to Bayes Theorem
2. Binomial Model and Poisson Model
3. Normal Model
4. Normal Regression Model
5. Bayesian Numerical Calculations (Markov Chain Monte Carlo)
6. Application of MCMC
7. Model Selection
8. miscellaneous topics

Gibbs Sampler: general form

3 : Gibbs Sampler (p blocks)

1. Initialize $X_1^{(0)} = x_1^{(0)}$, $X_2^{(0)} = x_2^{(0)}$, ... ,$X_p^{(0)} = x_p^{(0)}$
2. For $t = 0, 1, \ldots$ repeat the following steps:

(a) Draw $X_1^{(t+1)} = x_1^{(t+1)}$ according to
$f_{X_1}(X_1^{(t+1)}|X_2^{(t)} = x_2^{(t)}, X_3^{(t)} = x_3^{(t)}, \ldots, X_p^{(t)} = x_p^{(t)})$

(b) Draw $X_2^{(t+1)} = x_2^{(t+1)}$ according to
$f_{X_2}(X_2^{(t+1)}|X_1^{(t+1)} = x_1^{(t+1)}, X_3^{(t)} = x_3^{(t)}, \ldots, X_p^{(t)} = x_p^{(t)})$

(c) Draw $X_3^{(t+1)} = x_3^{(t+1)}$ according to
$f_{X_3}(X_3^{(t+1)}|X_1^{(t+1)} = x_1^{(t+1)}, X_2^{(t+1)} = x_2^{(t+1)}$
$, X_4^{(t)} = x_4^{(t)}, \ldots, X_p^{(t)} = x_p^{(t)})$

$\vdots$ $\vdots$

(d) Draw $X_j^{(t+1)} = x_j^{(t+1)}$ according to
$f_{X_j}(X_j^{(t+1)}|X_1^{(t+1)} = x_1^{(t+1)}, \ldots, X_{j-1}^{(t+1)} = x_{j-1}^{(t+1)}$
$, X_{j+1}^{(t)} = x_{j+1}^{(t)}, \ldots, X_p^{(t)} = x_p^{(t)})$

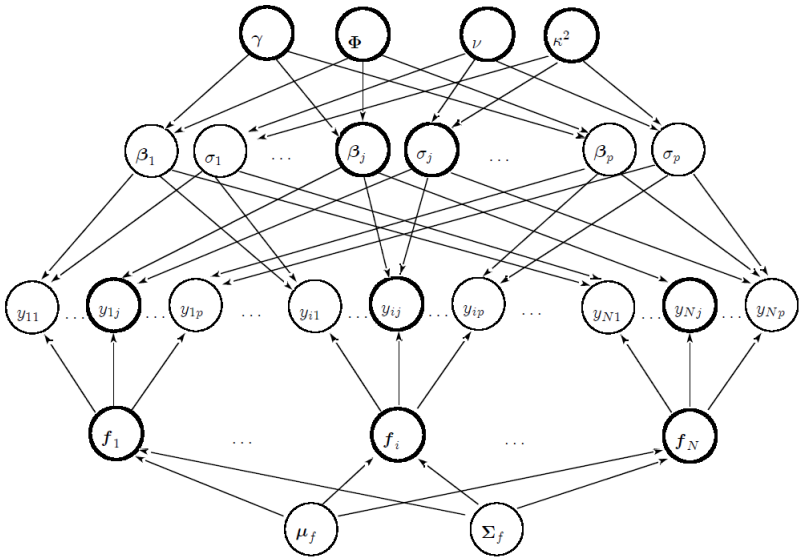
$\vdots$ $\vdots$

(e) Draw $X_p^{(t+1)} = x_p^{(t+1)}$ according to
$f_{X_p}(X_p^{(t+1)}|X_1^{(t+1)} = x_1^{(t+1)}, X_2^{(t+1)} = x_2^{(t+1)}, \ldots$
$, X_{p-1}^{(t+1)} = x_{p-1}^{(t+1)})$

How to find the Full Conditional Dist.

- $f(x_j \mid x_1, x_2, \ldots x_{j-1}, x_{j+1}, x_{j+2}, \ldots, x_p)$
- Regard the joint pdf as the pdf of X_j given the rest.
- Use DAG : Directional Acyclic Graph

DAG: Directional Acyclic Graph



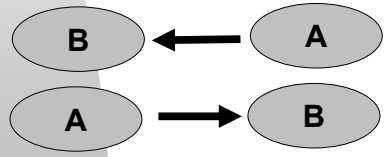
DAG

- How to draw a DAG.
  - circle → random variable
  - arrow → dependance
- Example:
  - B depends on A.

No arrow indicates (conditional) independence.

$$Pr(A,B) = Pr(B|A)Pr(A)$$

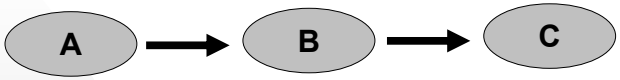
DAG indicates the decomposition of joint probability.



DAG

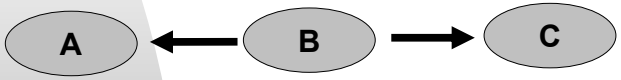
- B depends on A and C depends on B.

$$Pr(A,B,C) = Pr(C|B)\,Pr(B|A)\,Pr(A)$$



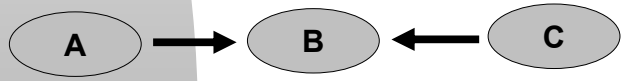
- A depends on B and C depends on B.

$$Pr(A,B,C) = Pr(C|B)\,Pr(A|B)\,Pr(B)$$



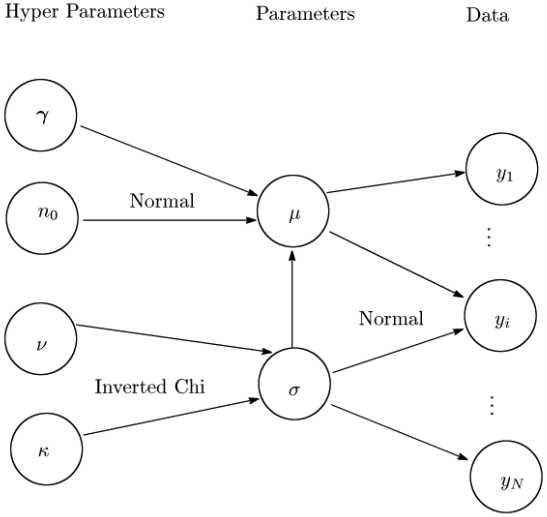
- B depends on both A and C.

$$Pr(A,B,C) = Pr(B|A,C)Pr(A)Pr(C)$$



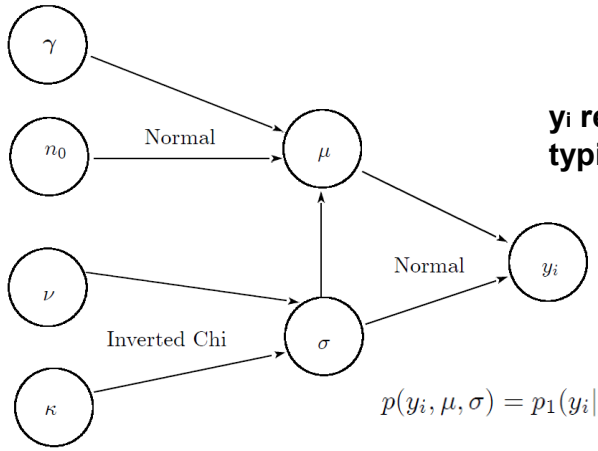
DAG: Normal with Nat. Conj. Prior

$$\begin{aligned} y_i \mid \mu, \sigma &\sim N(\mu, \sigma^2) \\ \mu \mid \sigma &\sim N(\gamma, \frac{1}{n_0}\sigma^2) \\ \sigma &\sim \chi^{-1}(\nu, \kappa^2) \end{aligned}$$



DAG: Normal with Nat. Conj. Prior

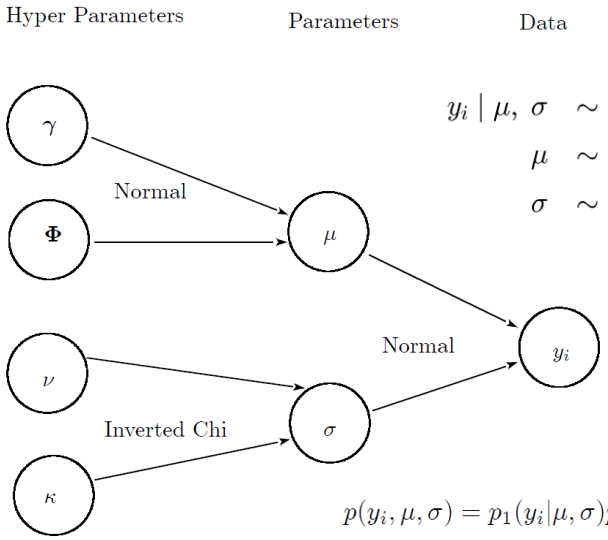
$$
\begin{aligned}
y_i \mid \mu, \sigma &\sim N(\mu, \sigma^2) \\
\mu \mid \sigma &\sim N(\gamma, \frac{1}{n_0}\sigma^2) \\
\sigma &\sim \chi^{-1}(\nu, \kappa^2)
\end{aligned}
$$



y_i represents typical observation.

$$p(y_i, \mu, \sigma) = p_1(y_i|\mu, \sigma)p_2(\mu|\sigma, \gamma, n_0)p_3(\sigma|\nu, \kappa)$$

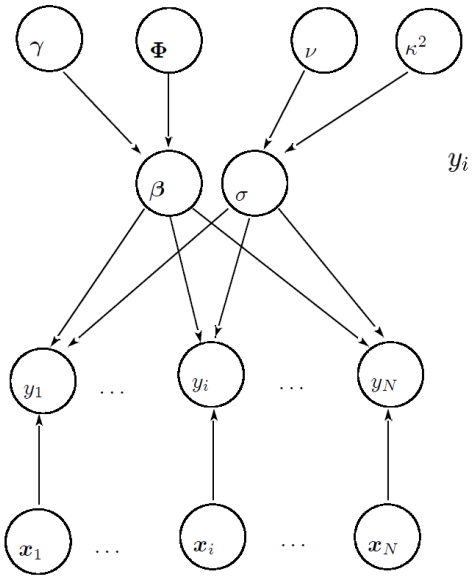
DAG: Normal with Semi Nat. Conj. Prior



$$
\begin{aligned}
y_i \mid \mu, \sigma &\sim N(y_i|\mu, \sigma^2) \\
\mu &\sim N(\gamma|\phi^2) \\
\sigma &\sim \chi^{-1}(\nu, \kappa^2)
\end{aligned}
$$

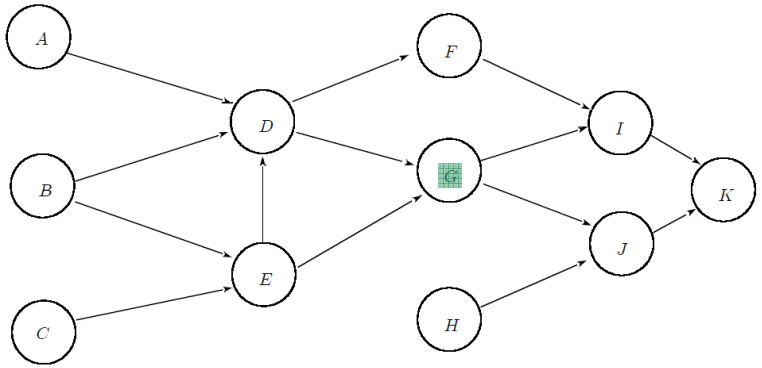
$$p(y_i, \mu, \sigma) = p_1(y_i|\mu, \sigma)p_2(\mu|\gamma, \Phi)p_3(\sigma|\nu, \kappa)$$

DAG: Regression Model



$$
\begin{aligned}
y_i \mid \boldsymbol{\beta}, \sigma, \boldsymbol{x}_i &\sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2) \\
\boldsymbol{\beta} &\sim N(\boldsymbol{\gamma}, \boldsymbol{\Phi}) \\
\sigma &\sim \chi^{-1}(\nu, \kappa^2)
\end{aligned}
$$

DAG can be used to find the full conditional distribution of a variable.

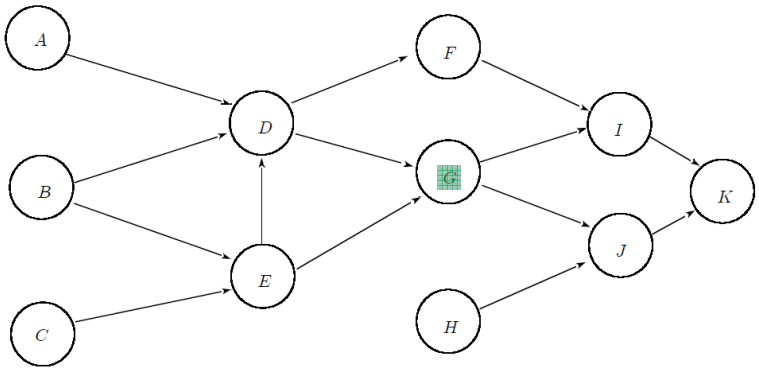


$$
\begin{aligned}
p(A, B, C, D, E, F, G, H, I, J, K) &= p(K|I, J)p(J|G, H)p(I|F, G)p(H)p(G|D, E)p(F|D) \\
&\times p(D|A, B, E)p(E|B, C)p(A)p(B)p(C)
\end{aligned}
$$

What is the full conditional distribution of G given the rest?

$$p(G|A, B, C, D, E, F, H, I, J, K) \propto p(J|G, H)p(I|F, G)p(G|D, E) \propto p(G|D, E, F, H, I, J)$$

DAG can be used to find the full conditional distribution of a variable.



What is the full conditional distribution of G given the rest?

$$p(G|A,B,C,D,E,F,H,I,J,K) = p(G|D,E,F,H,I,J)$$

Depends on: Parents(D,E), Children(I,J), and Spouses(F,H)
Spouse = Other Parents of the Children
Markov Blanket

Markov Blanket

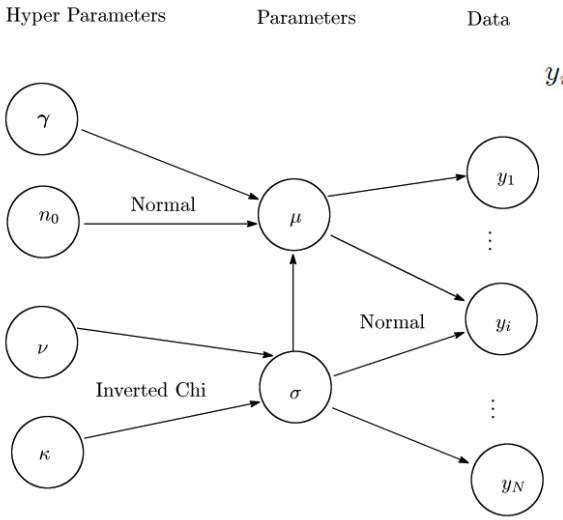
- Given the Markov Blanket of a variable, say, G, (D,E,F,H,I,J), G and the remaining variables, (A,B,C,K), are independent.

$$p(A,B,C,K,G|D,E,F,H,I,J) = p(G|D,E,F,H,I,J)\, p(A,B,C,K|D,E,F,H,I,J)$$

- Therefore,

$$\begin{aligned} p(G|A,B,C,D,E,F,H,I,J,K) &= \frac{p(A,B,C,K,G|D,E,F,H,I,J)}{p(A,B,C,K|D,E,F,H,I,J)} \\ &= p(G|D,E,F,H,I,J) \end{aligned}$$

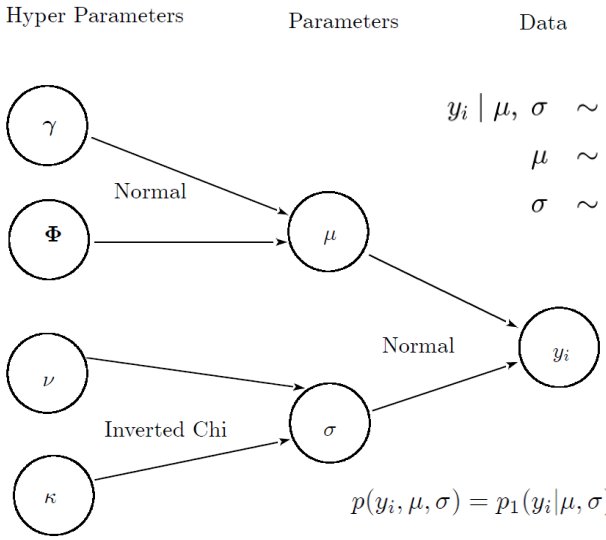
DAG: Normal with Nat. Conj. Prior



$$\begin{aligned} y_i \mid \mu,\sigma &\sim N(\mu,\sigma^2) \\ \mu \mid \sigma &\sim N(\gamma,\frac{1}{n_0}\sigma^2) \\ \sigma &\sim \chi^{-1}(\nu,\kappa^2) \end{aligned}$$

$$p(y_i,\mu,\sigma) = p_1(y_i|\mu,\sigma)p_2(\mu|\sigma,\gamma,n_0)p_3(\sigma|\nu,\kappa)$$

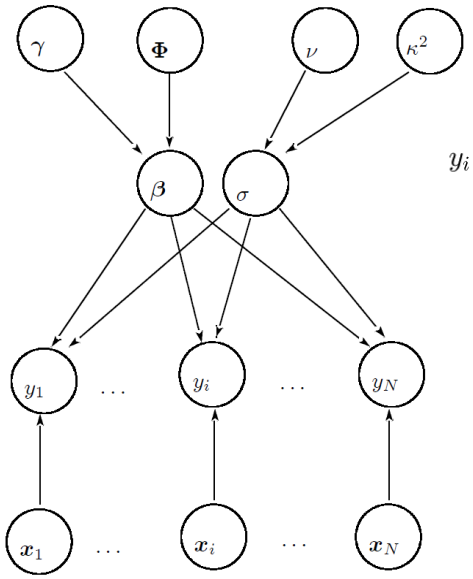
DAG: Normal with Semi Nat. Conj. Prior



$$\begin{aligned} y_i \mid \mu,\ \sigma &\sim N(y_i|\mu,\sigma^2) \\ \mu &\sim N(\gamma|\phi^2) \\ \sigma &\sim \chi^{-1}(\nu,\kappa^2) \end{aligned}$$

$$p(y_i,\mu,\sigma) = p_1(y_i|\mu,\sigma)p_2(\mu|\gamma,\ \Phi)p_3(\sigma|\nu,\kappa)$$

DAG: Regression Model



$$
\begin{aligned}
y_i \mid \boldsymbol{\beta}, \sigma, \boldsymbol{x}_i &\sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2) \\
\boldsymbol{\beta} &\sim N(\boldsymbol{\gamma}, \boldsymbol{\Phi}) \\
\sigma &\sim \chi^{-1}(\nu, \kappa^2)
\end{aligned}
$$

Random Coefficient Model

- Suppose that
  - data (y_k,X_k), k=1,2,…,m, are collected in m classes of the same school.
  - regress y_k on X_k for each class as

$$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{E}_k,\ \boldsymbol{E}_k \sim N(\boldsymbol{0}, \sigma_k^2\boldsymbol{I}_{N_k}),\ k = 1, 2, \ldots, m$$

- Assume that the regression coefficients, β_k, are similar and express this similarity by assuming the same (prior) distributions for the regression coefficients.

$$\boldsymbol{\beta}_k \sim N(\boldsymbol{\gamma}, \boldsymbol{\Phi}),\ k = 1, 2, \ldots, m$$

Random Coefficient Model

- m group regression
  borrowing strength from the neighbors

  Hierarchical Linear Model : HLM
  Multilevel Model
  Slope as Outcome Model

1st stage

$$\boldsymbol{y}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{E}_k,\ \boldsymbol{E}_k \sim N(\boldsymbol{0}, \sigma_k^2\boldsymbol{I}_{N_k}),\ k = 1, 2, \ldots, m$$

$$\boldsymbol{y}_k \sim N(\boldsymbol{X}_k\boldsymbol{\beta}_k, \sigma_k^2\boldsymbol{I}_{N_k}),\ k = 1, 2, \ldots, m$$

2nd stage

$$
\begin{aligned}
\boldsymbol{\beta}_k &\sim N(\boldsymbol{\gamma}, \boldsymbol{\Phi}),\ k = 1, 2, \ldots, m \\
\sigma_k &\sim \chi^{-1}(\nu, \kappa^2),\ k = 1, 2, \ldots, m
\end{aligned}
$$

$$
\begin{aligned}
h(\boldsymbol{\beta}, \boldsymbol{\sigma} \mid \boldsymbol{y}) &\propto f(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}) h(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{\Phi}) h(\boldsymbol{\sigma} \mid \nu, \kappa) \\
&= \prod_{k=1}^{m} f(\boldsymbol{y}_k \mid \boldsymbol{\beta}_k, \boldsymbol{\sigma}_k) h(\boldsymbol{\beta}_k \mid \boldsymbol{\gamma}, \boldsymbol{\Phi}) h(\boldsymbol{\sigma}_k \mid \nu, \kappa)
\end{aligned}
$$

3rd stage

$$
\begin{aligned}
\boldsymbol{\gamma} &\sim N(\boldsymbol{\zeta}, \boldsymbol{\Psi}) \\
\boldsymbol{\Phi} &\sim W^{-1}(\eta, \mathcal{T})
\end{aligned}
$$

Random Coefficient Model

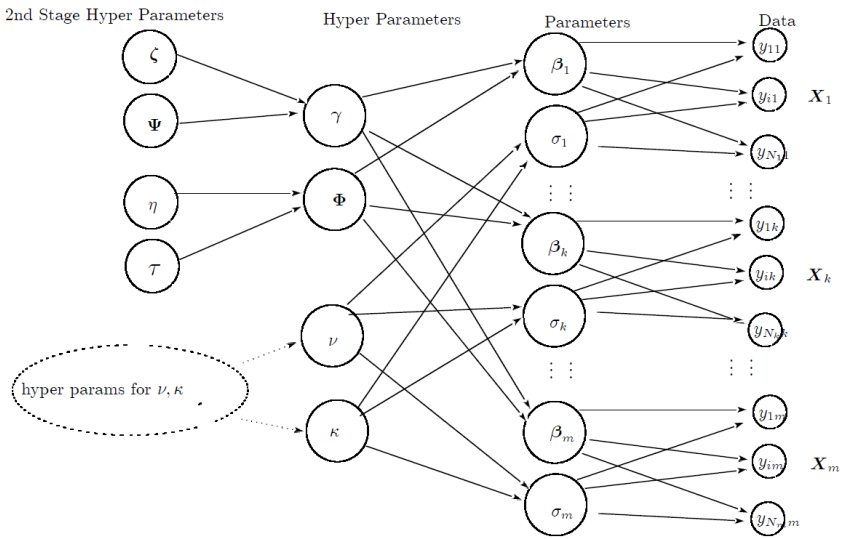
- marginally:

$$\boldsymbol{y}_k \sim N\left(\boldsymbol{X}_k\boldsymbol{\gamma},\ \boldsymbol{X}_k\boldsymbol{\Phi}\boldsymbol{X}_k' + \sigma_k^2\boldsymbol{I}\right)$$

$$
\begin{aligned}
\boldsymbol{\gamma} &\sim N(\boldsymbol{\zeta}, \boldsymbol{\Psi}) \\
\boldsymbol{\Phi} &\sim W^{-1}(\eta, \mathcal{T}) \\
\sigma_k &\sim \chi^{-1}(\nu, \kappa^2),\ k = 1, 2, \ldots, m
\end{aligned}
$$

Difficult to generate RV!

Random Coefficient Model



Finite Mixture Models

- If we know the membership:

$$d_{iq} = \begin{cases} 1 & \textbf{if the i-th obs. belongs class q.} \\ 0 & \textbf{otherwise.} \end{cases}$$

$$f(y_i|d_{iq}=1) = f(y_i|\boldsymbol{\xi}_q) \qquad \boldsymbol{\xi}_q \sim h(\boldsymbol{\xi}_q \mid \boldsymbol{\gamma})$$

- Distribution of the membership

$$h(\boldsymbol{d}_i) \propto \prod_{q=1}^{Q} \rho_q^{d_{iq}} \qquad h(\boldsymbol{\rho}) \propto \prod_{q=1}^{Q} \rho_q^{\alpha_q - 1}$$

- Since we do not know the membership, marginally:

$$\begin{aligned} f(y_i|\boldsymbol{\xi}) &= \sum_{q=1}^{Q} f(y_i, \boldsymbol{d}_i|\boldsymbol{\xi}) \\ &= \sum_{q=1}^{Q} \prod_{q=1}^{Q} \left(f(y_i|\boldsymbol{\xi}_q)\ \rho_q\right)^{d_{iq}} \\ &= \sum_{q=1}^{Q} f(y_i|\boldsymbol{\xi}_q)\ \rho_q \end{aligned}$$

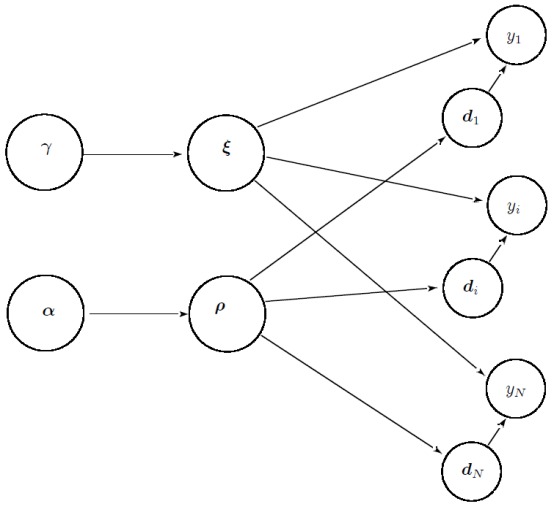
d_{iq} is a latent variable.

Finite Mixture Models

- Univariate Normal Mixtures

$$Y_i \mid d_{iq} \sim N(\mu_q, \sigma_q^2)$$

Finite Mixture Models



t-distributed errors in regression

Errors are distributed as t distribution.

$$\begin{aligned} y_i &= \sum_{j=1}^{q} x_{ij}\beta_j + e_i \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i, \qquad e_i \sim t(\nu, 0, \kappa^2),\ i = 1, 2, \ldots, N,\ iid \qquad var(e_i) = \frac{\nu}{\nu - 2}\kappa^2 \end{aligned}$$

$$y_i \sim t\left(\nu,\ \boldsymbol{x}_i'\boldsymbol{\beta},\ \kappa^2\right),\ i = 1, 2, \ldots, N,\ iid$$

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\beta}, \nu, \kappa) &\propto \prod_{i=1}^{N} \frac{\Gamma(\frac{1}{2}(\nu+1))}{\Gamma(\frac{\nu}{2})}\nu^{-\frac{1}{2}}\kappa^{-1} \left(1 + \frac{1}{\nu\kappa^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right)^{-\frac{1}{2}(\nu+1)} \\ &= \left(\frac{\Gamma(\frac{1}{2}(\nu+1))}{\Gamma(\frac{\nu}{2})}\nu^{-\frac{1}{2}}\kappa^{-1}\right)^{N} \prod_{i=1}^{N}\left(1 + \frac{1}{\nu\kappa^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right)^{-\frac{1}{2}(\nu+1)} \end{aligned}$$

No conjugate priors!

t-distributed errors in regression

Alternative formulation of t-errors

If

$$\begin{aligned} y_i|\sigma_i &\sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma_i^2) \\ \sigma_i &\sim \chi^{-1}(\nu, \nu\kappa^2),\ iid \end{aligned}$$

or

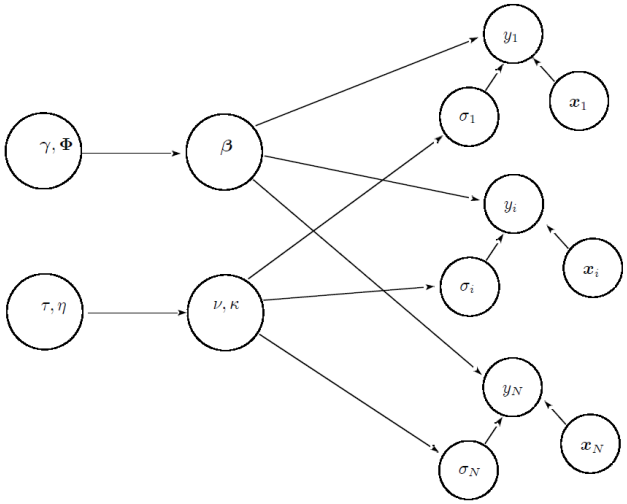
$$\begin{aligned} y_i|\sigma_i, \kappa^2 &\sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma_i^2\kappa^2) \\ \sigma_i &\sim \chi^{-1}(\nu, \nu),\ iid \end{aligned}$$

then , marginally:

$$y_i \sim t(\nu, \boldsymbol{x}_i'\boldsymbol{\beta}, \kappa^2)$$

σ_i is a latent variable.

t-distributed error in regression



Probit Model

- categorical criterion variable: y

$$f(y|\boldsymbol{p}(\boldsymbol{x})) = p(\boldsymbol{x})^y(1 - p(\boldsymbol{x}))^{1-y}$$

$$\Pr(Y = 1|\boldsymbol{x}) = p(\boldsymbol{x})$$

- model for the probability:

$$z = \boldsymbol{x}'\boldsymbol{\beta} + e, \quad e \sim N(0, \sigma^2) \qquad z \sim N(\boldsymbol{x}'\boldsymbol{\beta}, \sigma^2)$$

$$\begin{aligned} p(\boldsymbol{x}) = \Pr(Z > 0) &= \Pr\left(\frac{Z - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma} > -\frac{\boldsymbol{x}'\boldsymbol{\beta}}{\sigma}\right) \\ &= 1 - \Phi(-\frac{\boldsymbol{x}'\boldsymbol{\beta}}{\sigma}) = \Phi(\frac{\boldsymbol{x}'\boldsymbol{\beta}}{\sigma}) \end{aligned}$$

Probit Model

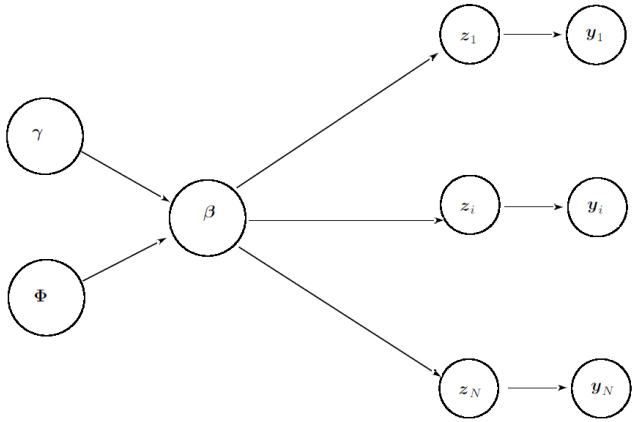
- likelihood

$$f(\boldsymbol{y}|\boldsymbol{\beta},\sigma) = \prod_{i=1}^{N} p(\boldsymbol{x}_i)^{y_i}(1-p(\boldsymbol{x}_i))^{1-y_i}$$
$$= \prod_{i=1}^{N} \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})^{y_i}(1-\Phi(\boldsymbol{x}_i'\boldsymbol{\beta}))^{1-y_i}$$

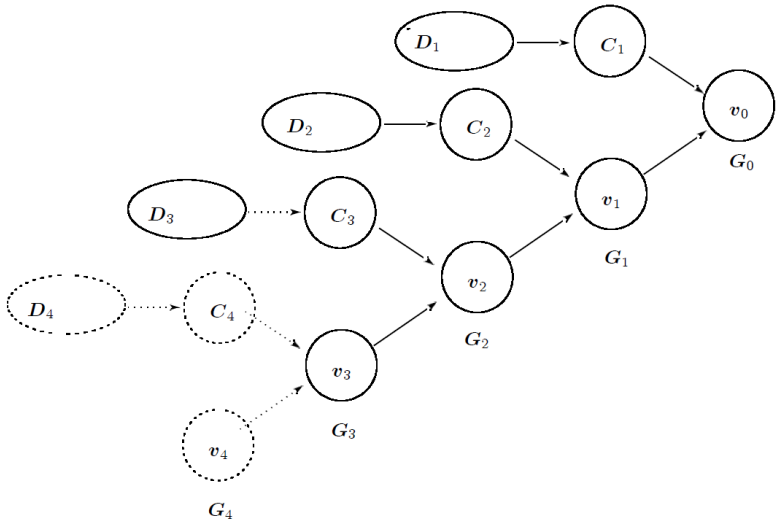
No conjugate priors!

Treat z as the latent variable.

Probit Model



Hierarchical Model



Leftovers

- model comparison
  - M1: y=Xβ+e, e~N(0,σ2)
    M2: y=Xβ+e, e~t(0,ν,c2)
  - posterior odds of models, Bayes Factor, etc
- What is Markov Chain?
  - Sequence of random numbers, $X_1, X_2, \ldots$, in which $f(X_i \mid X_1, X_2, \ldots, X_{i-1}) = f(X_i \mid X_{i-1})$.
- What do we do when the full-conditional distribution cannot be identified and therefore we are unable to generate random numbers from it?
  - Use Metropolis-Hastings method which is another form of MCMC algorithm.

Leftovers

- multivariate analysis
  - multivariate regression, analysis of variance
- latent variable models
  - factor analysis
  - item response theory
- comparison with the classical statistics
- how to assess prior distribution
- Bayesian decision theory
- etc etc

Exam Topics

- Outline the general procedure for the Bayesian inference using the following keywords.
  - keywords:
    model, parameters, model distribution,
    subjective probability, prior distribution, likelihood,
    posterior distribution, posterior mean,
    posterior standard deviation, natural conjugate prior,
    semi-natural conjugate prior, Monte Carlo method,
    Markov Chain Monte Carlo, gibbs sampler, DAG,
    etc etc…
- What do you think about the Bayesian statistics?
- Due on Aug. 16, in pdf format, by e-mail