Hierarchical Models

Simon Jackman

Stanford University
http://jackman.stanford.edu/BASS

February 11, 2012

Hierarchical Models

- data span multiple groups or time periods
- "context matters"
- group-specific statistical structure to data

$$\begin{aligned}
y_j|\theta_j &\sim p(y_j|\theta_j) \quad && \text{(model for the data in group } j = 1, \ldots, J\text{)} \\
\theta_j|\nu &\sim p(\theta_j|\nu) \quad && \text{(between-group model or ``prior'' for the parameters } \theta_j\text{)} \\
\nu &\sim p(\nu) \quad && \text{(prior for the \textit{hyperparameter} } \nu\text{),}
\end{aligned}$$

Example: one-way analysis of variance

$$
\begin{aligned}
y_{ij}|\alpha_j, \sigma^2 &\sim N(\alpha_j, \sigma^2) \quad (1a)\\
\alpha_j|\mu, \omega^2 &\sim N(\mu, \omega^2). \quad (1b)
\end{aligned}
$$

- i indexes observations; j indexes J groups
- α_j: mean of y in group j
- equation 1b is a model for how α_j varies across groups.
- μ is the mean of the distribution of the group means (the "grand mean")
- variance ω^2, also known as the *between* variance;
- σ^2 is known as the *within* variance for group j; constant across groups here, could relax this and have σ_j^2 (group-wise heteroskedasticity)
- $\omega^2/(\omega^2 + \sigma^2)$ is known as the *intra-class correlation* and is a measure of the "relative similarity" of observations in each group
- Bayesian analysis: need priors for μ, σ^2 and ω^2.

Example: multilevel regression

$$\begin{aligned} y_{ij} &\sim N(\mathbf{x}_{ij}\boldsymbol{\beta}_j, \sigma_j^2) && \text{(2a)} \\ \beta_{jk} &\sim N(\mathbf{z}_j\boldsymbol{\gamma}_k, \omega_k^2) && \text{(2b)} \end{aligned}$$

- i indexes observations; j indexes J groups
- k indexes K covariates; i.e., $\mathbf{x}_{ij}\boldsymbol{\beta}_j = x_{ij1}\beta_{j1} + \ldots + x_{ijk}\beta_{jk}$
- need priors for $\boldsymbol{\gamma}_k$ and ω_k^2, $k = 1, \ldots, K$; priors for σ_j^2, $j = 1, \ldots, J$.

Representation as a Mixed Model

$$\mathbf{y}|\mu, \alpha, \sigma^2, \omega^2 \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Sigma}) \tag{3}$$

- $\mathbf{X}$ is a n-by-k matrix of predictors that pick up fixed effects $\boldsymbol{\beta}$ (a k-by-1 vector of coefficients)
- $\mathbf{Z}$ is a n-by-p matrix of predictors that pick up random effects $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is a p-by-p covariance matrix.
- $\boldsymbol{\Sigma}$ is a n-by-n covariance matrix.
- i.e., the effect of $\mathbf{x}$ in group j is $\beta + b_j$.
- $\text{cov}(b_j, \varepsilon_{ij}|\mu) = 0 \; \forall \; i, j$.

Variance Components Representation

- One-way ANOVA decomposes variation in $\mathbf{y}$ around μ into two components: the "within-group" variance σ^2 and the "between-group" variance ω^2.

$$
\begin{aligned}
\text{var}(\mathbf{y}) &= \text{var}(\mu) + \underbrace{\mathbf{Z}\text{var}(\mathbf{b})\mathbf{Z}'}_{\text{"between variance"}} + \underbrace{\text{var}(\boldsymbol{\varepsilon})}_{\text{"within variance"}} && (4a) \\
&= \omega^2 \mathbf{Z}\mathbf{I}_J\mathbf{Z}' + \sigma^2 \mathbf{I}_n && (4b) \\
&= \omega^2 \mathbf{F} + \sigma^2 \mathbf{I}_n, && (4c)
\end{aligned}
$$

- $\mathbf{F} = \mathbf{Z}\mathbf{Z}'$ is a block diagonal matrix with blocks $\boldsymbol{\iota}_{n_j}\boldsymbol{\iota}'_{n_j}$ (a square n_j-by-n_j matrix of ones), $j = 1, \ldots, J$.

Variance Components Representation

$$\text{var}(\mathbf{y}) = \text{var}(\mu) + \underbrace{\mathbf{Z}\text{var}(\mathbf{b})\mathbf{Z}'}_{\text{“between variance”}} + \underbrace{\text{var}(\boldsymbol{\varepsilon})}_{\text{“within variance”}}$$

- for group j, we have

$$\text{var}(\mathbf{y}_j) = \omega^2 \boldsymbol{\iota}_{n_j}\boldsymbol{\iota}'_{n_j} + \sigma^2 \mathbf{I}_{n_j} = \begin{bmatrix} \sigma^2+\omega^2 & \omega^2 & \dots & \omega^2 \\ \omega^2 & \sigma^2+\omega^2 & \dots & \omega^2 \\ \vdots & \vdots & \ddots & \vdots \\ \omega^2 & \omega^2 & \dots & \sigma^2+\omega^2 \end{bmatrix},$$

- non-zero covariance across observations *within* groups; these observations share the group-specific term $b_j \sim N(0, \omega^2)$.
- within-cluster covariance is explicit
- Contrast classical approaches that treat the “clustered” nature of the data as a nuisance; inference for the fixed effects with “cluster robust” standard errors.

Exchangeable parameters generate hierarchical models

- introduced exchangeability earlier
- extend concept from data to parameters
- generates models for parameters
- may require covariates if not unconditionally exchangeable; i.e., parameters in hierarchical model induce conditional exchangeability

Example: Exchangeability and hierarchical models for polling data

- Suppose we have data from a survey conducted in J counties in the United States. Individual level responses, support for border protection: $y_{ij} = 1$ if respondent i wants more spending on border protection and 0 otherwise.
- No individual-level predictors with which to model the responses, and so exchangeability is a reasonable assumption at the micro-level. Thus, $r_j \sim \text{Binomial}(\theta_j, n_j)$, $r_j = \sum_{i=1}^{n} y_{ij}$
- d_j is distance of county j from US border

$$
\begin{aligned}
r_j &\sim \text{Binomial}(\theta_j, n_j) \\
\log\left(\frac{\theta_j}{1 - \theta_j}\right) &\sim N(\beta_0 + \beta_1 d_j, \omega^2), \\
(\beta_0, \beta_1)' &\sim N(\mathbf{b}, \boldsymbol{\Sigma})
\end{aligned}
$$

Hierarchical models "borrow strength" across units

$$\begin{aligned} y_{ij} | \alpha_j, \sigma_j^2 &\sim N(\alpha_j, \sigma^2) \\ \alpha_j | \mu, \omega^2 &\sim N(\mu, \omega^2). \end{aligned}$$

- inferences for the group-level parameters α_j reflect not just the information about α_j in group j, but, via the hierarchical model, will also draw on relevant information in the other groups.
- information about the α_j flows "up" the hierarchy to inform inferences about the distribution of the α_j across groups
- i.e., data informative for μ and ω^2 too
- data from group j helps shape the posterior density over α_k ($\forall\, k \neq j$) via contribution to inferences for the hyperparameters μ and ω^2.
- "sharing" or "borrowing" information across groups follows from exchangeability/hierarchical models

Hierarchical model as "semi-pooling"

$$\begin{aligned} y_{ij}|\alpha_j, \sigma_j^2 &\sim N(\alpha_j, \sigma^2) \\ \alpha_j|\mu, \omega^2 &\sim N(\mu, \omega^2). \end{aligned}$$

- Compare two other "extreme" models:
- **No pooling:** $y_{ij}|\alpha_j, \sigma^2 \sim N(\alpha_j, \sigma^2)$, dropping the hierarchical component of the model.
- Equivalent to setting $\omega^2 = \infty$ in $\alpha_j \sim N(\mu, \omega^2)$.
- **Complete pooling:** grouping in the data is irrelevant, and we impose the restriction that $\alpha_j = \mu, \forall j$, generating the model $y_{ij} \sim N(\mu, \sigma^2)$.
- Equivalent to setting $\omega^2 = 0$ in $\alpha_j \sim N(\mu, \omega^2)$.
- **Hierarchical model** lies between these two extreme cases

Hierarchical Model as a "Shrinkage" Estimator

$$\begin{aligned} y_{ij}|\alpha_j, \sigma_j^2 &\sim N(\alpha_j, \sigma^2) \\ \alpha_j|\mu, \omega^2 &\sim N(\mu, \omega^2). \end{aligned}$$

- Bayes estimates of α_j are:
  1. shifted away from the group-level mean $\bar{y}_j$ in the direction of μ, the mean of the distribution of the group level parameters.
  2. are more precise than inferences based on an analysis of group j in isolation from the other groups.
- "shrinkage": the α_j are pulled towards the grand mean μ, relative to the distribution of group level parameters we obtain with no pooling.
- Stein's (1955) result: The Bayesian "semi-pooled" or "shrinkage" estimator dominates both the "no pooling" and "complete pooling" estimators with respect to total mean square error.

Theorem (Bayes estimates, one-way ANOVA as hierarchical model)

If $y_{ij} \sim N(\alpha_j, \sigma_j^2)$, $\alpha_j \sim N(\mu, \omega^2)$ *then* $\alpha_j | \mathbf{y}_j, \sigma_j^2, \mu, \omega^2 \sim N(\tilde{\mu}_j, V_j)$ *where*

$$\tilde{\mu}_j = \frac{\mu\omega^{-2} + \bar{y}_j \frac{n_j}{\sigma_j^2}}{\omega^{-2} + \frac{n_j}{\sigma_j^2}} \quad \textit{and} \quad V_j = \left(\omega^{-2} + \frac{n_j}{\sigma_j^2}\right)^{-1}.$$

Equivalently,

$$\begin{aligned} \tilde{\mu}_j &= \lambda_j \mu + (1 - \lambda_j)\bar{y}_j \\ \lambda_j &= \frac{\omega^{-2}}{\omega^{-2} + \frac{n_j}{\sigma_j^2}} = \frac{\frac{\sigma_j^2}{n_j}}{\omega^2 + \frac{\sigma_j^2}{n_j}} = \frac{V(\bar{y}_j)}{V(\bar{y}_j) + \omega^2} \end{aligned}$$

and λ_j *is a measure of how much* α_j *is "shrunk" away from* $\bar{y}_j$ *towards* μ.

$$\tilde{\mu}_j = \lambda_j \mu + (1 - \lambda_j)\bar{y}_j$$

where

$$\lambda_j = \frac{\omega^{-2}}{\omega^{-2} + \frac{n_j}{\sigma_j^2}} = \frac{\frac{\sigma_j^2}{n_j}}{\omega^2 + \frac{\sigma_j^2}{n_j}} = \frac{V(\bar{y}_j)}{V(\bar{y}_j) + \omega^2}$$

- familiar "precision-weighted" average form
- when group j provides little information about α_j in group j --- e.g., n_j is small, and so $V(\bar{y}_j)$ is large, relative to the between variance ω^2 --- then the shrinkage factor λ_j grows, and the Bayes estimate of α_j is pulled towards the grand mean μ.
- if the between variance ω^2 is relatively large --- e.g., groups are quite heterogeneous -- then the Bayes estimate of α_j will display less shrinkage, relying more on information in group j and less "borrowing strength" from other groups.

Computation via Markov chain Monte Carlo

- Easy under conjugacy: normal data, normal prior for group-specific α_j, normal for hyperparameter μ, inverse-Gamma for within-variance σ^2 and between-variance ω^2.
- not necessary: e.g., $\sigma_j \sim \text{Unif}(0, k)$
- DAG structure makes hierarchical models well suited for general-purpose solutions like `BUGS`/`JAGS`
- Many other programs too: `MLWin`, `HLM`, `lme4` package in `R`

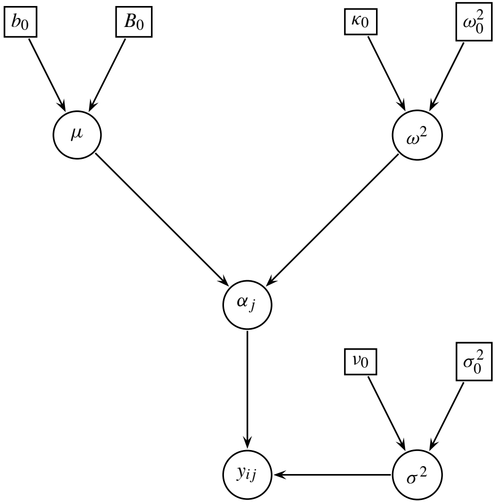
One way ANOVA, conditionally conjugate hierarchical model

$$
\begin{aligned}
y_{ij}|\alpha_j, \sigma^2 &\sim N(\alpha_j, \sigma^2) \\
\alpha_j|\mu_0, \omega^2 &\sim N(\mu_0, \omega^2) \\
\mu_0 &\sim N(b_0, B_0) \\
\sigma^2 &\sim \text{inverse-Gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \\
\omega^2 &\sim \text{inverse-Gamma}(\kappa_0/2, \kappa_0\omega_0^2/2)
\end{aligned}
$$

- $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_J, \mu_0, \sigma^2, \omega^2)$
- prior:

$$
\begin{aligned}
p(\boldsymbol{\theta}) &= p(\alpha_1, \ldots, \alpha_J, \mu_0, \sigma^2, \omega^2) \\
&= p(\alpha_1, \ldots, \alpha_J, |\mu_0, \omega^2)p(\mu_0)p(\sigma^2)p(\omega^2) \\
&= \prod^{J} p(\alpha_j|\mu_0, \omega^2)p(\mu_0)p(\sigma^2)p(\omega^2)
\end{aligned}
$$

DAG for one way ANOVA as hierarchical model



Conditional distributions, Gibbs sampler

- $p(\alpha_j | \mathcal{G} \setminus \alpha_j), j = 1, \ldots, J$: parents of each α_j are μ_0 and ω^2; the children of α_j are the data in unit j, $\mathbf{y}_j = (y_1, \ldots, y_{n_j})'$ and the parents of $\mathbf{y}_j$ are α_j and σ^2.

$$\alpha_j | (\mathcal{G} \setminus \alpha_j) \sim N\left(\frac{\mu_0\, \omega^{-2} + \bar{y}_j \frac{n_j}{\sigma^2}}{\omega^{-2} + \frac{n_j}{\sigma^2}}, \left(\omega^{-2} + \frac{n_j}{\sigma^2} \right)^{-1} \right),$$

- $p(\mu_0 | \mathcal{G} \setminus \mu_0)$. Parents of μ_0 are just its prior hyperparameters, the prior mean and variance b_0 and B_0 respectively. The children of μ_0 are $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$. The $\boldsymbol{\alpha}$ have two parents, μ_0 and ω^2.

$$\mu_0 | (\mathcal{G} \setminus \mu_0) \sim N\left(\frac{b_0 B_0^{-1} + \bar{\mu} \frac{J}{\omega^2}}{B_0^{-1} + \frac{J}{\omega^2}}, \left(B_0^{-1} + \frac{J}{\omega^2} \right)^{-1} \right),$$

where $\bar{\mu} = J^{-1} \sum_{j=1}^{J} \alpha_j$.

Conditional distributions, Gibbs sampler

- $p(\omega^2 | \mathcal{G} \setminus \omega^2)$. The parents of ω^2 are just its prior hyperparameters, κ_0 and ω_0^2. The children of ω^2 are the α_j; the parents of α_j are ω^2 and μ_0.

$$\omega^2 | (\mathcal{G} \setminus \omega^2) \sim \text{inverse-Gamma}\left(\frac{\kappa_0 + J}{2}, \frac{\kappa_0 \omega_0^2 + S_\mu}{2}\right)$$

  where $S_\mu = \sum_{j=1}^{J} (\alpha_j - \mu_0)^2$.

- $p(\sigma^2 | \mathcal{G} \setminus \sigma^2)$. The parents of σ^2 are just its prior hyperparameters, ν_0 and σ_0^2. The children of σ^2 are the y_{ij}; the parents of the y_{ij} are the α_j and σ^2.

$$\sigma^2 | \mathcal{G} \setminus \sigma^2 \sim \text{inverse-Gamma}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + S_\mathbf{Y}}{2}\right)$$

  where $n = \sum_{j=1}^{j} n_j$ is the total number of observations and $S_\mathbf{Y} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \alpha_j)^2$ is the total sum-of-squares of $\mathbf{Y}$.

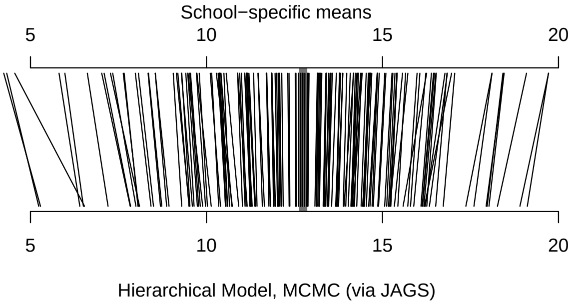
Example 7.6, one way ANOVA , HSB

JAGS code

```
model{
        for(i in 1:N){
              mu.y[i] <- mu[j[i]]
              math[i] ~ dnorm(mu.y[i],tau[1])
        }

        for(p in 1:J){
              mu[p] ~ dnorm(mu0,tau[2])
        }

        mu0 ~ dnorm(0,.0001)
        for(p in 1:2){
              tau[p] <- pow(sigma[p],-2)
              sigma[p] ~ dunif(0,10)
        }
}
```

Example 7.6, one way ANOVA, HSB

- The Bayes estimates of α_j --- means of the marginal posterior densities of the α_j, or $E(\alpha_j|\mathbf{y}, \sigma^2\omega^2)$ --- are "shrunk" towards the grand mean by the hierarchical model relative to the MLEs.
- The MLEs are simply the sample means, i.e., $\hat{\alpha}_j^{(\text{MLE})} = \bar{y}_j$.



2-way ANOVA, state and year effects in presidential elections data, Example 7.7

- Two-levels of grouping: state $i = 1, \ldots, 50$ and year $t = 1984, 1988, \ldots 2004$.
- Model:

$$
\begin{aligned}
y_{it} &\sim N(\mu + \alpha_i + \delta_t, \sigma^2) \\
\mu &\sim N(50, 15^2) \\
\alpha_i &\sim N(0, \sigma^2_\alpha) \\
\delta_t &\sim N(0, \sigma^2_\delta) \\
\sigma_\alpha &\sim \text{Unif}(0, 15) \\
\sigma_\delta &\sim \text{Unif}(0, 15)
\end{aligned}
$$

- Slow-mixing MCMC algorithm

50,000 iterations, thinned by 5:

	Geweke	Heidelberger-Welch	Raftery-Lewis	
Parameter	z	p	N	I
μ	0.74	0.82	256665	68.50
σ	-0.53	0.82	18705	4.99
σ_α	-0.61	0.99	18550	4.95
σ_δ	-1.83	0.89	19170	5.12

Example 7.7: over-parameterization for better mixing

$$
\begin{aligned}
y_{it} &\sim N(\mu + \alpha_i + \delta_t, \sigma^2) \\
\mu &\sim N(0, 100^2) \\
\alpha_i &\sim N(\mu_\alpha, \sigma_\alpha^2) \\
\delta_t &\sim N(\mu_\delta, \sigma_\delta^2) \\
\mu_\alpha &\sim N(0, 100^2) \\
\mu_\delta &\sim N(0, 100^2)
\end{aligned}
$$

- μ, μ_α and μ_δ not identified.
- Map back to identified parameters by imposing the restrictions

$$\sum_{i=1}^{n} \alpha_i = 0 \Rightarrow \bar{\alpha} = 0 \quad \sum_{t=1}^{T} \delta_t = 0 \Rightarrow \bar{\delta} = 0$$

- apply these identifying restrictions in `JAGS` or by *post-processing* the MCMC output in `R`

Example 7.7: over-parameterization for better mixing

- At iteration m, define

$$
\begin{aligned}
\alpha_i^{*(m)} &= \alpha_i^{(m)} - \bar{\alpha}^{(m)}, \quad i = 1, \ldots, n \\
\delta_t^{*(m)} &= \delta_t^{(m)} - \bar{\delta}^{(m)}, \quad t = 1, \ldots, T \\
\mu^{*(m)} &= \mu^{(m)} + \bar{\alpha}^{(m)} + \bar{\delta}^{(m)}.
\end{aligned}
$$

- n.b., simply a re-parameterization; we get the same likelihood contributions either way, since

$$
\begin{aligned}
\mu^{*} + \alpha_i^{*} + \delta_i^{*} &= \mu + \bar{\alpha} + \bar{\delta} + \alpha_i - \bar{\alpha} + \delta_i - \bar{\delta} \\
&= \mu + \alpha_i + \delta_t.
\end{aligned}
$$

Example 7.7: over-parameterization for better mixing

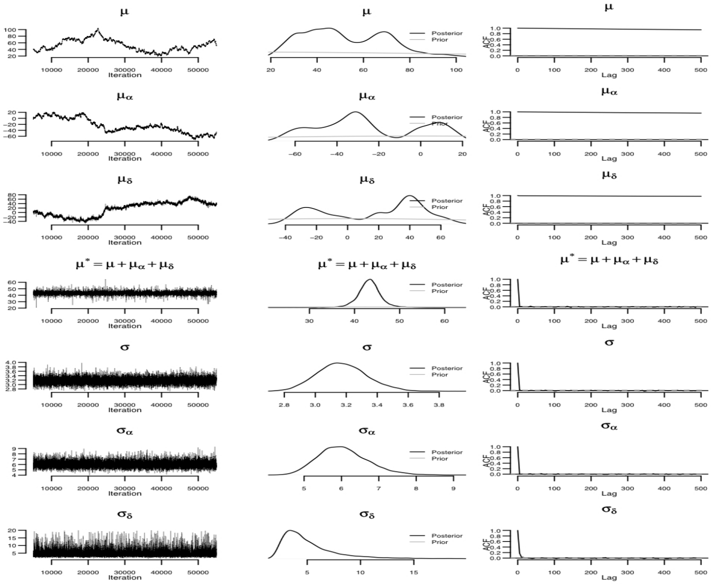
JAGS code

```
model{
        for(i in 1:n){
              mu.y[i] <- mu[1] + alpha[s[i]] + delta[j[i]]
              demVote[i] ~ dnorm(mu.y[i],tau[1])
        }
        sigma[1] ~ dunif(0,20)
        sigma[2] ~ dunif(0,20)
        sigma[3] ~ dunif(0,20)

        for(i in 1:50){
              alpha[i] ~ dnorm(mu[2],tau[2])
        }
        for(i in 1:nyear){
              delta[i] ~ dnorm(mu[3],tau[3])
        }
        for(i in 1:3){
              tau[i] <- pow(sigma[i],-2)
        }
        for(i in 1:3){
              mu[i] ~ dnorm(0,1E-4)
        }

        ## transformations for identified parameters
        mustar <- mu[1] + mean(alpha[]) + mean(delta[])
        for(i in 1:50){
              alphastar[i] <- alpha[i] - mean(alpha[])
        }
        for(i in 1:nyear){
              deltastar[i] <- delta[i] - mean(delta[])
        }
}
```

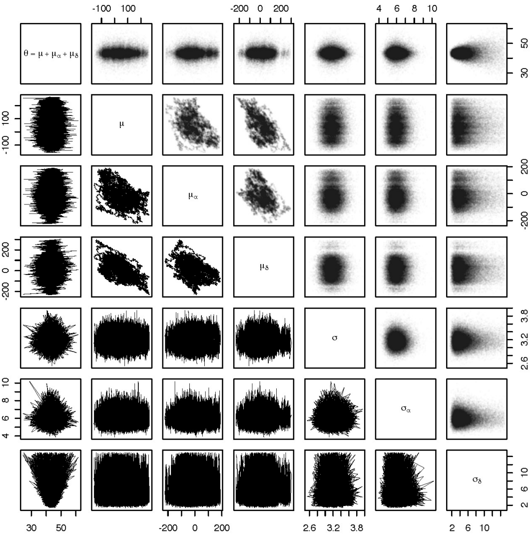
Example 7.7: over-parameterization for better mixing

Parameter	Geweke z	Heidelberger-Welch p	Raftery-Lewis N	Raftery-Lewis I
$\mu^* = \mu + \bar{\alpha} + \bar{\delta}$	-0.58	0.91	19645	5.24
σ	-0.34	0.53	18855	5.03
σ_α	-0.22	0.21	19010	5.07
σ_δ	1.25	0.75	18705	4.99

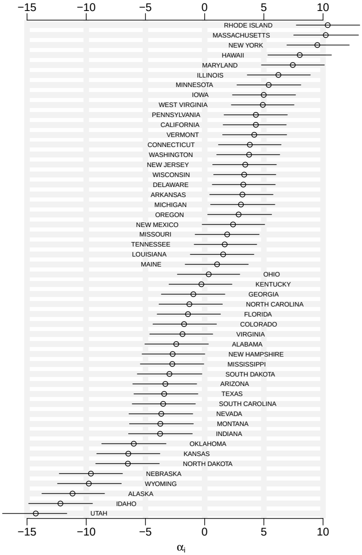
Example 7.7: over-parameterization for better mixing



Ex 7.7: over-parameterization for better mixing



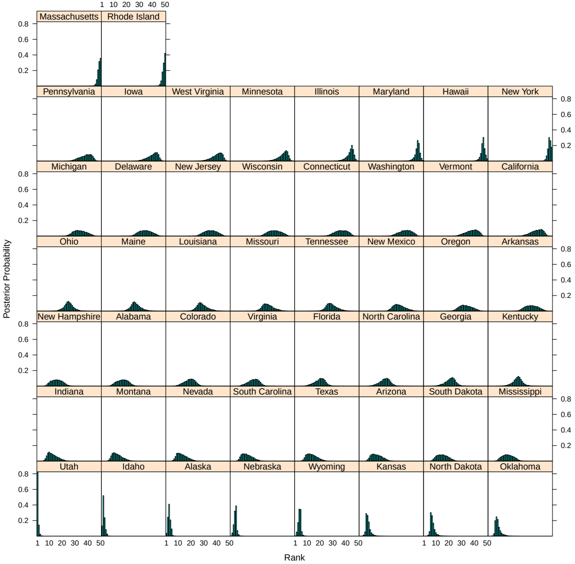
Ex 7.7: marginal posterior densities, state effects α_i



Extension of Ex 7.7: inducing a posterior mass function over ranks of α_i

- At each iteration of the Gibbs sampler we have $\boldsymbol{\alpha}^{(t)} = (\alpha_1^{(t)}, \ldots, \alpha_n^{(t)})'$
- Compute ranks: to produce $\mathbf{r}^{(t)} = (r_1^{(t)}, \ldots, r_n^{(t)})'$, $r_i \in \{1, 2, \ldots, n\} \forall\, i$.
- A simulation consistent estimate of the posterior probability that α_i occupies rank p is simply the proportion of times we see $r_i^{(t)} = p$ over many iterations of the Gibbs sampler, $t = 1, \ldots, T$.
- Demo with code in `alphaSort.R`

Posterior Mass Function over ranks



Other Examples, do in “slow-motion” in R

- multi-level regression, HSB
- Green and Vavreck, “Rock The Vote” cluster-randomized field experiment on voter turnout: hierarchical model for treatment effects in binomial model.
- show superior out-of-sample performance of hierarchical model with linear growth curves; e.g., presidential elections data, rat growth, etc.
- Exercises from Ch 7 in book
- hierarchical models also appear in Ch 8 (e.g., hierarchical model for interviewer effects); Ch 9 (e.g., modeling latent variables as a function of observables).