Hierarchical models

Benjamín Garzón

PhD course 3045, May 2018

Karolinska Institutet

Grouped data

- Groups $j = 1, ..., N$, and observations $Y = \{y_{ij}\}$, $i = 1, ..., M$
- E.g. N subjects with M trials per subject
- (Number of observations need not be equal for all groups)
- Group parameters of interest θ_j, $y_{ij} \sim p(y_{ij}|\theta_j)$

Observations x Groups

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1N} \\ y_{21} & y_{22} & y_{23} & \cdots & y_{2N} \\ \hdotsfor{5} \\ y_{M1} & y_{M2} & y_{M3} & \cdots & y_{MN} \end{bmatrix}$$

Complete pooling

Consider all θ_j equal

- $y_{ij} \sim p(y_{ij}|\theta)$
- Robust: uses all data to estimate θ
- Cannot look at inter-group differences
- Does not consider within vs between sources of variation

$$\theta=\theta_1=\ldots=\theta_N$$

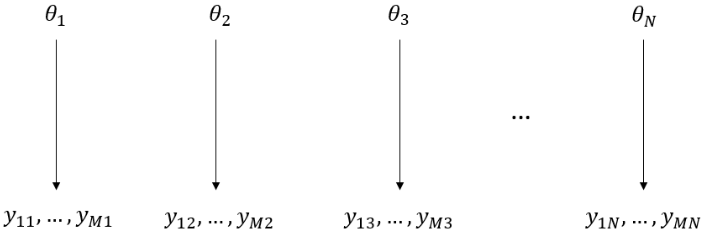
$$\Big\downarrow$$

$$y_{11}, \ldots, y_{M1}, y_{12}, \ldots, y_{M2}, y_{13}, \ldots, y_{M3}, \ldots, y_{1N}, \ldots, y_{MN}$$

No pooling

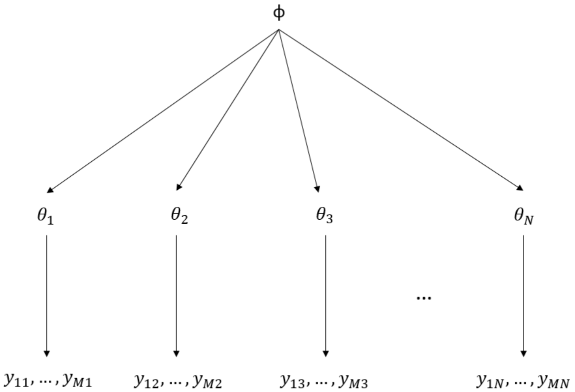
Model each group independently

- One θ_j per group
- $y_{ij} \sim p(y_{ij}|\theta_j)$
- Uses only a subset of the data to estimate θ_j
- Does not exploit similarity between θ_j



Hierarchical model

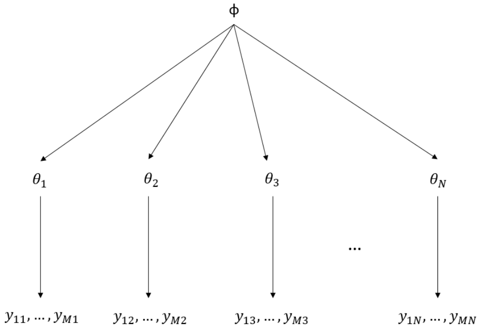
- Some parameters in the model may be related through a common distribution, $\theta_j \sim p(\theta_j|\phi)$
- $p(\theta_j|\phi)$ is governed by its own (hyper)parameters ϕ (population parameters)



Hierarchical model

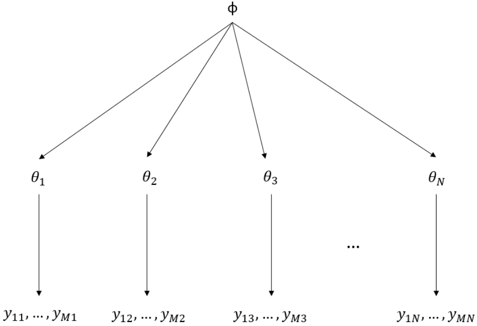
- We want to estimate θ_j (local/group parameters) and ϕ (global/population parameters)
- Joint distribution: $p(Y, \theta_1, ..., \theta_N, \phi)$
- Inference: Bayes rule

$$p(\theta_1, ..., \theta_N, \phi | Y) = \frac{p(\theta_1, ..., \theta_N, \phi) p(Y | \theta_1, ..., \theta_N, \phi)}{p(Y)}$$



Exchangeability

- θ_j exchangeable if we have no information to distinguish between them (e.g. no ordering or grouping)
- *Ignorance implies exchangeability*
- $(\theta_1, ..., \theta_N)$ are exchangeable if the prior $p(\theta_1, ..., \theta_N, \phi)$ is invariant to permutation of the indices (i.e. symmetric wrt θ_j)



Hierarchical model distributions

$$p(\theta_1, ..., \theta_N, \phi | Y) = \frac{p(\theta_1, ..., \theta_N, \phi) p(Y | \theta_1, ..., \theta_N, \phi)}{p(Y)}$$

Assuming exchangeability, the simplest form for a hierarchical model (other possibilities exist):

- Prior:

$$p(\theta_1, ..., \theta_N, \phi) = p(\phi) \prod_{j=1}^{N} p(\theta_j | \phi)$$

  Need to specify a prior for ϕ (hyperprior)

  θ_j are independent given ϕ

- Likelihood:

$$p(Y | \theta_1, ..., \theta_N, \phi) = \prod_{j=1}^{N} p(y_{ij} | \theta_j) = \prod_{j=1}^{N} p(y_{ij} | \theta_j)$$

  y_{ij} in each group j are independent given θ_j

  Likelihood does not depend on ϕ given θ_j

Hierarchical model with normal likelihoods and group distributions

(Hierarchical structure only for means):

- Data y_{ij}
- Group-level parameters: $\theta_j = (\mu_j, \sigma)$, $i = 1, ..., N$
- Hyperparameters: $\phi = (\mu_p, \sigma_p)$

- Likelihood: $p(y_{ij}|\theta_j) = \mathcal{N}(y_{ij}; \mu_j,\, \sigma^2)$ *(common σ)*
- Group distributions $p(\theta_j|\phi)$:
  $p(\mu_j|\phi) = \mathcal{N}(\mu_j; \mu_p,\, \sigma_p^2);\ p(\sigma|\phi) = p(\sigma) \propto 1$
- Hyperprior distribution $p(\phi)$: $p(\mu_p) = \mathcal{N}(\mu_p; \mu_0,\, \sigma_0^2);\ p(\sigma_p) \propto 1$
  (μ_0 and σ_0 are fixed)

Example: Eight schools

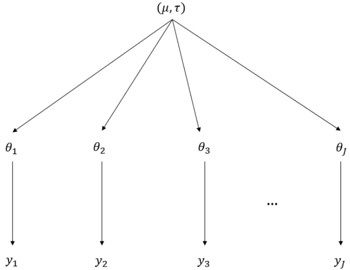
y_j, σ_j : mean and std of the effect of coaching on school performance in school j

$$\mu \sim \mathcal{N}(0, 5^2)$$

$$\tau \sim \mathit{Half} - \mathit{Cauchy}(0, 5)$$

$$\theta_j \sim \mathcal{N}(\mu, \tau^2)$$

$$y_j \sim \mathcal{N}(\theta_j, \sigma_j^2)$$



(Rubin et al., 1981)

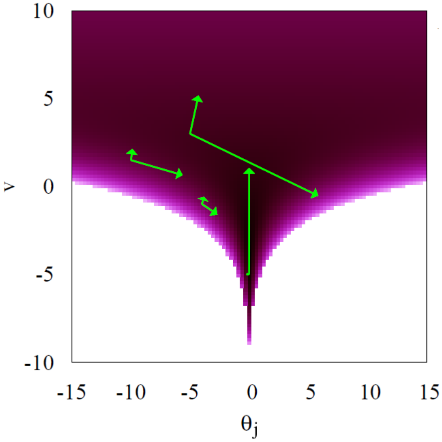
Example: Eight schools

```
data {
 int<lower=0> J;
 real y[J];
 real<lower=0> sigma[J];
}
parameters {
 real mu;
 real<lower=0> tau;
 real theta[J];
}
model {
 mu ~ normal(0, 5);
 tau ~ cauchy(0, 5);
 theta ~ normal(mu, tau);
 y ~ normal(theta, sigma);
}
```

Funnel geometry

$$\theta_j \sim \mathcal{N}(\mu, \tau^2)$$

$$v = log(\tau^2)$$



(Betancourt et al., 2013)

Example: Eight schools, non-centered parameterization (aka Matt trick)

$$\mu \sim \mathcal{N}(0, 5^2)$$

$$\tau \sim \mathit{Half} - \mathit{Cauchy}(0, 5)$$

$$\tilde{\theta}_j \sim \mathcal{N}(0, 1)$$

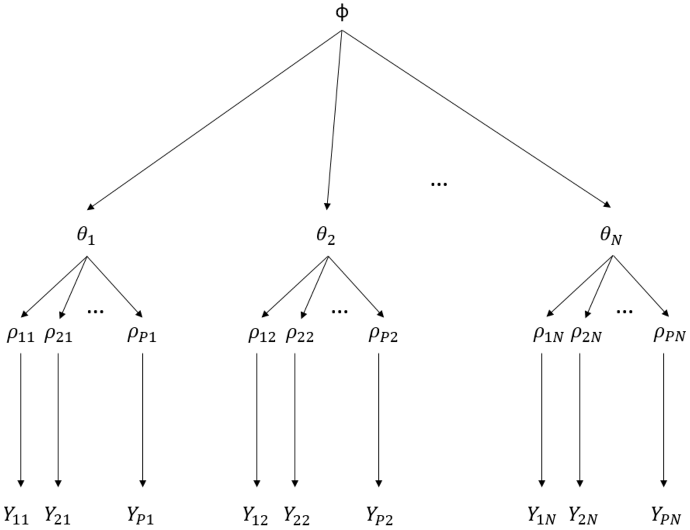
$$\theta_j = \mu + \tau\tilde{\theta}_j$$

$$y_j \sim \mathcal{N}(\theta_j, \sigma^2)$$

Example: Eight schools, non-centered parameterization

```
data {
 int<lower=0> J;
 real y[J];
 real<lower=0> sigma[J];
}
parameters {
 real mu;
 real<lower=0> tau;
 real theta_tilde[J];
}
transformed parameters {
 real theta[J];
 for (j in 1:J)
  theta[j] = mu + tau * theta_tilde[j];
}
model {
 mu ~ normal(0, 5);
 tau ~ cauchy(0, 5);
 theta_tilde ~ normal(0, 1);
 y ~ normal(theta, sigma);
}
```

Multiple levels



Comments

- Hierarchical models tend to *shrink* group parameters (compared to no pooling): tension between data and group mean
- Hierarchical structure most useful when few observations per group (no pooling tends to overfit)
- Hierarchical structure is prior information
- Sampling the posterior is difficult (Hamiltonian Monte Carlo...)