HIERARCHICAL MODELS

- Consider the number of car accidents over 30 years by a driver (M) in Milano and one (R) in Roma
- We can consider two persons, randomly selected or not, or the average of (a subset of) the population in the two cities but then we round up to an integer
- The event is rare and takes only integer values $\Rightarrow$ Poisson distribution
- $X \sim \mathcal{P}(\lambda) \rightarrow \mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda};\ x \in \mathbb{Z}$
- How should we model our data and prior for M and R?
- We should think if the behaviour of the two drivers is the same, completely different or there are similarities
- How do we transform those situations into a statistical model?

HIERARCHICAL MODELS

- n_M and n_R number of accidents for M and R
- λ_M and λ_R parameters for Poisson distribution for n_M and n_R
- **Equal**: If the two drivers are behaving in the same way, we model the data independently but with a common λ, with gamma prior $\mathcal{G}(\alpha, \beta)$
  - $\Rightarrow \pi(\lambda|n_M, n_R) \propto \lambda^{n_M} e^{-\lambda} \cdot \lambda^{n_R} e^{-\lambda} \cdot \lambda^{\alpha-1} e^{-\lambda\beta}$
  - $\Rightarrow \lambda|n_M, n_R \sim \mathcal{G}(\alpha + n_M + n_R, \beta + 2)$
- **Completely different**: If the two drivers are behaving in a completely different way, we model the data not only independently but also with different λ's, and independent gamma priors
  - $n_M \sim \mathcal{P}(\lambda_M)$ and $\lambda_M \sim \mathcal{G}(\alpha_M, \beta_M) \Rightarrow \lambda|n_M \sim \mathcal{G}(\alpha_M + n_M, \beta_M + 1)$
  - $n_R \sim \mathcal{P}(\lambda_R)$ and $\lambda_R \sim \mathcal{G}(\alpha_R, \beta_R) \Rightarrow \lambda|n_R \sim \mathcal{G}(\alpha_R + n_R, \beta_R + 1)$

HIERARCHICAL MODELS

- **Similar**: If the two drivers are behaving in a similar way, we model the data independently, with different λ's, but drawn from the same exponential (for simplicity) prior, dependent on a parameter θ
  - $\Rightarrow \pi(\lambda_M, \lambda_R | n_M, n_R, \theta) \propto \lambda_M^{n_M} e^{-\lambda_M} \cdot \lambda_R^{n_R} e^{-\lambda_R} \cdot \theta e^{-\lambda_M \theta} \cdot \theta e^{-\lambda_R \theta}$
  - $\Rightarrow \lambda_M | n_M, n_R, \theta \sim \mathcal{G}(n_M + 1, \theta + 1)$ and $\lambda_R | n_M, n_R, \theta \sim \mathcal{G}(n_R + 1, \theta + 1)$
- Two independent gamma posteriors for known θ but what about if unknown?
- We could consider a gamma prior $\theta \sim \mathcal{G}(a, b)$
- $\Rightarrow \pi(\lambda_M, \lambda_R, \theta | n_M, n_R) \propto \lambda_M^{n_M} e^{-\lambda_M} \cdot \lambda_R^{n_R} e^{-\lambda_R} \cdot \theta e^{-\lambda_M \theta} \cdot \theta e^{-\lambda_R \theta} \cdot \theta^{a-1} e^{-b\theta}$
- Gibbs sampling:
  - $\lambda_M | \lambda_R, \theta, n_M, n_R \sim \mathcal{G}(\theta + n_M + 1, \theta + 1)$
  - $\lambda_R | \lambda_M, \theta, n_M, n_R \sim \mathcal{G}(\theta + n_R + 1, \theta + 1)$
  - $\theta | \lambda_M, \lambda_R, n_M, n_R \sim \mathcal{G}(a + 2, b + \lambda_M + \lambda_R)$

HIERARCHICAL MODELS

- We have to integrate out θ if we are just interested in the full conditionals of each λ given the other

$$\begin{aligned}
\pi(\lambda_M, \lambda_R | n_M, n_R) &= \int \pi(\lambda_M, \lambda_R, \theta | n_M, n_R) d\theta \\
&\propto \lambda_M^{n_M} e^{-\lambda_M} \lambda_R^{n_R} e^{-\lambda_R} \int \theta^{a+1} e^{-(b+\lambda_M+\lambda_R)\theta} d\theta \\
&\propto \frac{\lambda_M^{n_M} e^{-\lambda_M} \lambda_R^{n_R} e^{-\lambda_R}}{(b+\lambda_M+\lambda_R)^{a+2}}
\end{aligned}$$

- $\Rightarrow$ We can use Gibbs sampling with Metropolis steps within
  - $\pi(\lambda_M | \lambda_R, n_M, n_R) \propto \dfrac{\lambda_M^{n_M} e^{-\lambda_M}}{(b+\lambda_M+\lambda_R)^{a+2}}$
  - $\pi(\lambda_R | \lambda_M, n_M, n_R) \propto \dfrac{\lambda_R^{n_R} e^{-\lambda_R}}{(b+\lambda_M+\lambda_R)^{a+2}}$
- As proposal distributions we could use $\mathcal{G}(n_M+1, 1)$ and $\mathcal{G}(n_R+1, 1)$, respectively

HIERARCHICAL MODELS

- Empirical Bayes is a practical, although not properly Bayesian, alternative to the choice of a prior on θ
- The idea is to find the value of θ maximising the probability of the data and plug it into the formulas
- The critical aspect, from a strict Bayesian viewpoint, is that data are used twice, first to find a value of θ and then computing the posterior distribution: priors should be independent from the data!
- We have to look for $\widehat{\theta} = \arg\max_\theta f(n_M, n_R|\theta)$
- With the same computations as before for θ known, we plug in $\widehat{\theta}$
  $\Rightarrow \lambda_M|n_M, n_R, \widehat{\theta} \sim \mathcal{G}(n_M + 1, \widehat{\theta} + 1)$ and $\lambda_R|n_M, n_R, \widehat{\theta} \sim \mathcal{G}(n_R + 1, \widehat{\theta} + 1)$

HIERARCHICAL MODELS

$$
\begin{aligned}
f(n_M, n_R|\theta) &= \int f(n_M, n_R|\lambda_M, \lambda_R)\pi(\lambda_M, \lambda_R|\theta)d\lambda_M d\lambda_R \\
&\propto \int \lambda_M^{n_M} e^{-\lambda_M} \cdot \lambda_R^{n_R} e^{-\lambda_R} \cdot \theta e^{-\lambda_M \theta} \cdot \theta e^{-\lambda_R \theta} d\lambda_M d\lambda_R \\
&\propto \theta^2 \int \lambda_M^{n_M} e^{-(\theta+1)\lambda_M} d\lambda_M \int \lambda_R^{n_R} e^{-(\theta+1)\lambda_R} d\lambda_R \\
&\propto \theta^2 \frac{\Gamma(n_M+1)}{(\theta+1)^{n_M+1}} \frac{\Gamma(n_R+1)}{(\theta+1)^{n_R+1}} \\
&\propto \frac{\theta^2}{(\theta+1)^{n_M+n_R+2}} \\
&= h(n_M, n_R, \theta)
\end{aligned}
$$

- $\dfrac{\partial \log h(n_M, n_R, \theta)}{\partial \theta} = \dfrac{2}{\theta} - \dfrac{n_M + n_R + 2}{\theta + 1}$
- $\dfrac{\partial \log h(n_M, n_R, \theta)}{\partial \theta} = 0 \Leftrightarrow \widehat{\theta} = \dfrac{2}{n_M + n_R}$

HIERARCHICAL MODELS

- Is $\widehat{\theta} = \frac{2}{n_M + n_R}$ surprising? Not much!
- We are considering an event described by a Poisson distribution with parameter λ
- For $X \sim \mathcal{P}(\lambda)$ we know that $E(X) = \lambda$
- For $\lambda \sim \mathcal{E}(\theta)$ we know that $E(\lambda) = 1/\theta$
- Since we use $\widehat{\theta} = \frac{2}{n_M + n_R}$, we can think of X somehow approximated (with some mathematical imprecision) by $\frac{n_M + n_R}{2}$, which is very reasonable under our assumptions

HIERARCHICAL MODELS

- In Italy every year students in some grades are taking tests on their knowledge about Italian language and Mathematics. The results of the tests could be affected by the school attended by the students so that it is reasonable to assume that the outcome for students of the same school are modelled by the same distribution while there should be a difference between schools.

- The same model could be used for batches of the same item but produced in different factories or survival times of patients in different hospitals

- We suppose that we observe data from n different groups, with n_i, $i = 1, \ldots, n$, elements in each of them

- Therefore the data are Y_{ij_i}, $i = 1, \ldots, n$ and $j_i = 1, \ldots, n_i$, although we will use Y_{ij} for simplicity

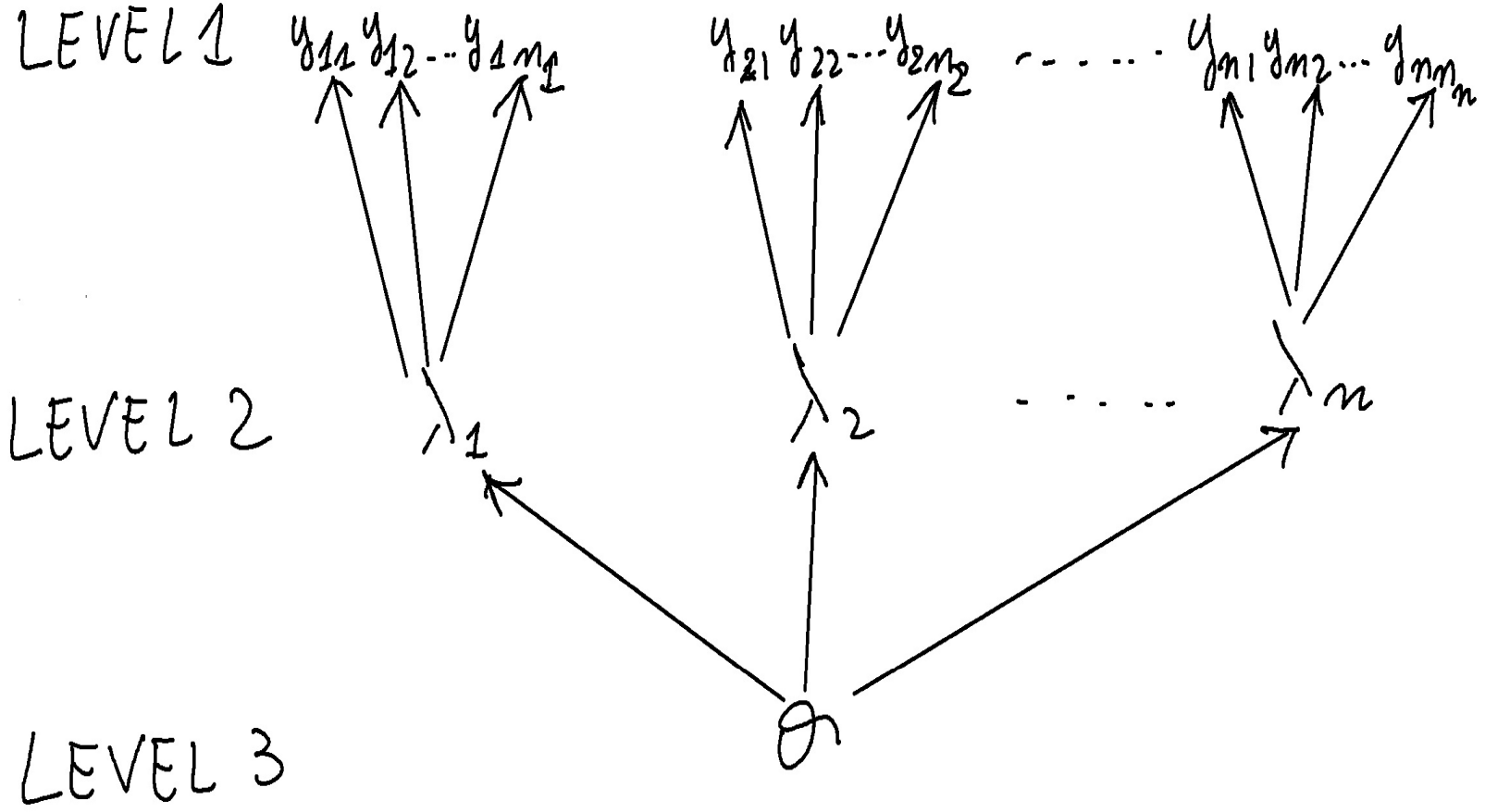
- Notation: $\underline{Y}_i = \{Y_{i1}, \ldots, Y_{i,n_i}\}$, $i = 1, \ldots, n$ data for i-th group

- Hierarchical models related to the notion of exchangeability, i.e. $\mathbb{P}(X_1, \ldots, X_n)$ invariant w.r.t. permutations (but we will not discuss it)

HIERARCHICAL MODELS

- Each group has its own distribution with a common parameter, i.e., the density of Y_{ij} is $f(y_{ij}|\lambda_i), i = 1, \ldots, n, j = 1, \ldots, n_i$
- This assumption implies a common behaviour within the group
- We assume that the functional form of f is not changing between groups (but it could)
- All the parameters λ_i's are supposed different (although sometimes some groups might have the same parameter)
- This assumption implies that the behaviour changes between groups
- All λ_i's come from the same distribution, i.e. $g(\lambda_i|\theta)$, where θ is a parameter in common
- This assumption implies that the behaviour of the groups, although different, is actually similar
- As before, a prior could be chosen for θ or a value could be plugged in, using, e.g., Empirical Bayes

HIERARCHICAL MODELS



HIERARCHICAL MODELS

- $\{y_{i1}, \ldots, y_{in_i} | \lambda_i\} \sim$ i.i.d. $f(y|\lambda_i), i = 1, \ldots, n$

  Within group sampling variability

- $\{\lambda_1, \ldots, \lambda_n\} \sim$ i.i.d. $g(\lambda|\theta)$

  Between groups sampling variability

- $\theta \sim \pi(\theta|\omega)$

  Prior distribution with hyperparameter ω

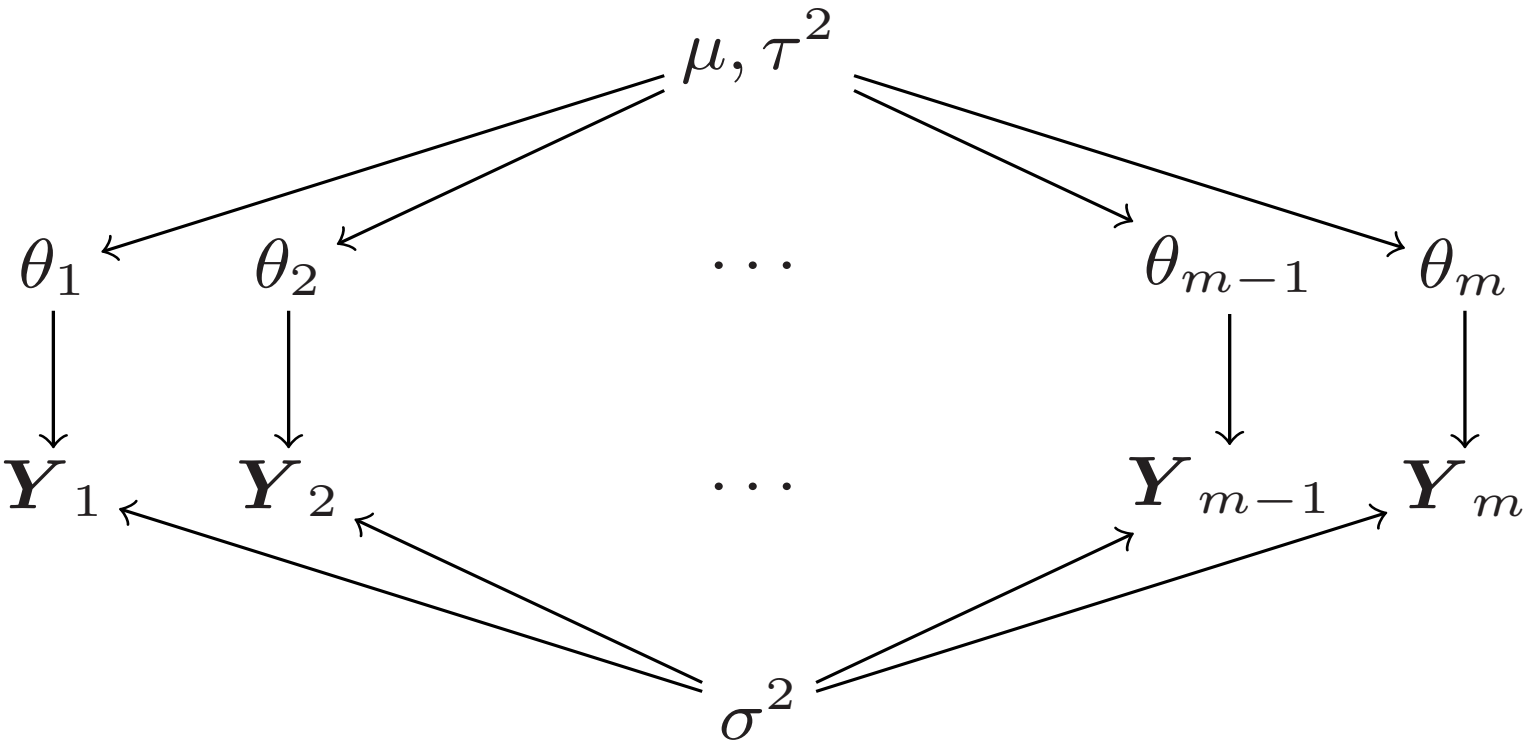
- Sometimes both $f(y|\lambda_i)$ and $g(\lambda|\theta)$ are called *sampling distributions*

- A popular model to describe heterogeneity of means across several populations is a hierarchical Normal model where both sampling distributions are Gaussian

HIERARCHICAL MODELS

- Observations in group $j, j = 1, \ldots, m$: $Y_{ji} \sim \mathcal{N}(\theta_j, \sigma^2)$ (Within group variability)
- Mean of group $j, j = 1, \ldots, m$: $\theta_j \sim \mathcal{N}(\mu, \tau^2)$ (Between groups variability)
- Independent priors on (μ, τ^2, σ^2) : $\pi(\mu)\pi(\tau^2)\pi(\sigma^2)$
  - $\mu \sim \mathcal{N}(\mu_0, \gamma_0^2)$
  - $\tau^2 \sim \mathcal{IG}(\eta_0/2, \eta_0\tau_0^2/2)$
  - $\sigma^2 \sim \mathcal{IG}(\nu/2, \nu\sigma_0^2/2)$
- Note that we assume the same variance for all the observations, while the mean is the same within a group but it changes between groups
- As seen graphically in the next slide, (μ, τ^2) provide information on Y's but, once θ is known, the distributions of Y's do not depend on (μ, τ^2)

HIERARCHICAL MODELS*



*From Hoff (2009), *A First Course in Bayesian Statistical Methods*, Springer

HIERARCHICAL MODELS

- Notation: $Y_i = (Y_{j1}, \ldots, Y_{jn_j})$, $j = 1, \ldots, m$
- $Y = (Y_1, \ldots, Y_m)$ and $\theta = (\theta_1, \ldots, \theta_m)$
- Joint posterior distribution

$$\begin{aligned}\pi(\theta, \mu, \tau^2, \sigma^2|Y) &\propto \pi(\mu, \tau^2, \sigma^2) g(\theta|\mu, \tau^2, \sigma^2) f(Y|\theta, \mu, \tau^2, \sigma^2) \\ &\propto \pi(\mu)\pi(\tau^2)\pi(\sigma^2) \left\{ \prod_{j=1}^{m} g(\theta_j|\mu, \tau^2) \right\} \left\{ \prod_{j=1}^{m} \prod_{i=1}^{n_j} f(y_{ji}|\theta_j, \sigma^2) \right\}\end{aligned}$$

- Full conditionals for μ and τ^2: $\pi(\mu, \tau^2|\theta, \sigma^2, Y) \propto \pi(\mu)\pi(\tau^2) \prod_{j=1}^{m} g(\theta_j|\mu, \tau^2)$
- $\pi(\mu|\theta, \tau^2, \sigma^2, Y) \propto \pi(\mu) \prod_{j=1}^{m} g(\theta_j|\mu, \tau^2)$
- $\pi(\tau^2|\theta, \mu, \sigma^2, Y) \propto \pi(\tau^2) \prod_{j=1}^{m} g(\theta_j|\mu, \tau^2)$

HIERARCHICAL MODELS

- The two full conditionals look very familiar!
  - Sample $(\theta_, \ldots, \theta_m)$ from $\mathcal{N}(\mu, \tau^2)$
  - $\mu \sim \mathcal{N}(\mu_0, \gamma_0^2)$
  - $\tau^2 \sim \mathcal{IG}(\eta_0/2, \eta_0\tau_0^2/2)$

- $\mu|\theta, \tau^2, Y \sim \mathcal{N}\left(\dfrac{m\overline{\theta}/\tau^2 + \mu_0/\gamma_0^2}{m/\tau^2 + 1/\gamma_0^2}, \left[m/\tau^2 + 1/\gamma_0^2\right]^{-1}\right)$

- $\tau^2|\theta, \mu, Y \sim \mathcal{IG}\left(\dfrac{\eta_0 + m}{2}, \dfrac{\eta_0\tau_0^2 + \sum_{j=1}^m (\theta_j - \mu)^2}{2}\right)$

- Here $\overline{\theta} = \displaystyle\sum_{j=1}^m \theta_j/m$

HIERARCHICAL MODELS

- Regarding θ, we can compute the full conditional for each θ_j, as dependent on $\mu, \tau^2, \sigma^2, Y_j$ since it is independent from the other θ_k's and the data from other groups
- $g(\theta_j|\mu, \tau^2, \sigma^2, Y_j) \propto g(\theta_j|\mu, \tau^2) \prod_{i=1}^{n_j} f(y_{ji}|\theta_j, \sigma^2), j = 1, \ldots, m$
- We have the product of Gaussian densities (already done, although in a simpler case)
- $\Rightarrow \theta_j|\mu, \tau^2, \sigma^2, Y_j \sim \mathcal{N}\left(\frac{n_j\overline{y}_j/\sigma^2 + 1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}, \left[n_j/\sigma^2 + 1/\tau^2\right]^{-1}\right)$
- Here $\overline{y}_j = \sum_{i=1}^{n_j} y_{ji}/n_j$

HIERARCHICAL MODELS

- Full conditional of σ^2

$$\begin{aligned}\pi(\sigma^2|\theta, Y) &\propto \pi(\sigma^2)\left\{\prod_{j=1}^{m} g(\theta_j|\mu,\tau^2)\right\}\left\{\prod_{j=1}^{m}\prod_{i=1}^{n_j} f(y_{ji}|\theta_j,\sigma^2)\right\} \\ &\propto (\sigma^2)^{-\nu_0/2+1} e^{-\nu_0\sigma_0^2/(2\sigma^2)} \cdot (\sigma^2)^{-\sum_{j=1}^{m} n_j/2} e^{-\sum_{j=1}^{m}\sum i=1^{n_j}(y_{ij}-\theta_j^2)/(2\sigma^2)}\end{aligned}$$

- $\Rightarrow \sigma^2|\theta, Y \sim \mathcal{IG}\left((\nu_0 + \sum_{j=1}^{m} n_j)/2, (\nu_0\sigma_0^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij}-\theta_j)^2)/2\right)$

- We use the Gibbs algorithm to get a sample from the posterior distribution since all the conditional distributions are properly specified