STATS 551

Hierarchical Models

Yang Chen

Department of Statistics

University of Michigan

ychenang@umich.edu

February 10, 2020

Overview

Plan

Hierarchical Models

- Structure of multiple parameters
- Common population distribution – prior
- Hyperparameters
- Fit ‘well’ without overfitting
- Example: analyze an experiment in the context of historical data

Rat Tumor Example

- Estimate θ, the probability of tumor in a population of female laboratory rate of type 'F344' in control group.

Rat Tumor Example

- Estimate θ, the probability of tumor in a population of female laboratory rate of type 'F344' in control group.
- Observation: 4 out of 14 rats develop cancer.

Rat Tumor Example

- Estimate θ, the probability of tumor in a population of female laboratory rate of type 'F344' in control group.
- Observation: 4 out of 14 rats develop cancer.
- Prior: $Beta(\alpha, \beta)$ from historical data.

Rat Tumor Example

- Estimate θ, the probability of tumor in a population of female laboratory rate of type 'F344' in control group.
- Observation: 4 out of 14 rats develop cancer.
- Prior: $Beta(\alpha, \beta)$ from historical data.
- Historical data: 70 groups of rats.
  - for jth experiment, $y_j \sim \text{Binomial}(n_j, \theta_j)$.

Rat Tumor Example

- Estimate θ, the probability of tumor in a population of female laboratory rate of type 'F344' in control group.
- Observation: 4 out of 14 rats develop cancer.
- Prior: $Beta(\alpha, \beta)$ from historical data.
- Historical data: 70 groups of rats.
  - for jth experiment, $y_j \sim \text{Binomial}(n_j, \theta_j)$.
  - Set (α, β) based on mean and sd of historical data.

Rat Tumor Example

- Data used twice?
  - Data used to construct prior
  - Data used to estimate θ

Rat Tumor Example

- Data used twice?
  - Data used to construct prior
  - Data used to estimate θ
- Point estimate for (α, β) is arbitrary.

Rat Tumor Example

- Data used twice?
    - Data used to construct prior
    - Data used to estimate θ
- Point estimate for (α, β) is arbitrary.
- Does it make sense to 'estimate prior' from data?

Plan

Exchangeability

- No information other than data is available to distinguish θ_j.

Exchangeability

- No information other than data is available to distinguish θ_j.
- The parameters $(\theta_1, \ldots, \theta_J)$ are exchangeable if
  - $p(\theta_1, \ldots, \theta_J)$ is invariant to permutations of the indexes $(1, \ldots, J)$.

Exchangeability

- No information other than data is available to distinguish θ_j.
- The parameters $(\theta_1, \ldots, \theta_J)$ are exchangeable if
  - $p(\theta_1, \ldots, \theta_J)$ is invariant to permutations of the indexes $(1, \ldots, J)$.
- Simple form

$$p(\theta) = \int \left(\prod_{j=1}^{J} p(\theta_j|\phi) \right) p(\phi) d\phi. \tag{1}$$

Exchangeability

- No information other than data is available to distinguish θ_j.
- The parameters $(\theta_1, \ldots, \theta_J)$ are exchangeable if
  - $p(\theta_1, \ldots, \theta_J)$ is invariant to permutations of the indexes $(1, \ldots, J)$.
- Simple form

$$p(\theta) = \int \left(\prod_{j=1}^{J} p(\theta_j | \phi) \right) p(\phi) d\phi. \tag{1}$$

- **de Finetti's theorem**: as $J \to \infty$, any suitably well-behaved exchangeable distribution on $(\theta_1, \ldots, \theta_J)$ can be expressed as a mixture of independent and identical distributions as (1).

Exchangeability

Positive correlation

Suppose the distribution of $\theta = (\theta_1, \ldots, \theta_J)$ can be written as a mixture of i.i.d. components,

$$p(\theta) = \int \prod_{j=1}^{J} p(\theta_j|\phi)p(\phi)d\phi.$$

Then the covariances $Cov(\theta_i, \theta_j)$ are all nonnegative.

Exchangeability

Positive correlation

Suppose the distribution of $\theta = (\theta_1, \ldots, \theta_J)$ can be written as a mixture of i.i.d. components,

$$p(\theta) = \int \prod_{j=1}^{J} p(\theta_j|\phi)p(\phi)d\phi.$$

Then the covariances $Cov(\theta_i, \theta_j)$ are all nonnegative.

Let $\mu(\phi) = E(\theta_j|\phi)$ for all j, then for all i, j,

$$cov(\theta_i, \theta_j) = E(cov(\theta_i, \theta_j|\phi)) + cov(E(\theta_i|\phi), E(\theta_j|\phi)) = Var(\mu(\phi)) \geq 0.$$

Exchangeability: Example

Suppose it is known a priori that $2J$ parameters $\theta_1, \ldots, \theta_{2J}$ are clustered into two groups, with exactly half drawn from $N(1,1)$ and the other half from $N(-1,1)$. But we do not know which.

1. Are $\theta_1, \ldots, \theta_{2J}$ exchangeable under the prior distribution?
2. Can this distribution be written as a mixture of i.i.d.s?
3. As $J \to \infty$ is it a counter example to de-Finetti's theorem?

Exchangeability: Example

1. The joint density $p(\theta_1, \ldots, \theta_{2J})$ is

$$\binom{2J}{J}^{-1} \sum_{\sigma(1,\ldots,2J)} \left[\prod_{j=1}^{J} N(\theta_{\sigma(j)}; 1, 1) \prod_{j=J+1}^{2J} N(\theta_{\sigma(j)}; -1, 1) \right],$$

where the sum is over all permutations.

Exchangeability: Example

1. The joint density $p(\theta_1, \ldots, \theta_{2J})$ is

$$\binom{2J}{J}^{-1} \sum_{\sigma(1,\ldots,2J)} \left[\prod_{j=1}^{J} N(\theta_{\sigma(j)}; 1, 1) \prod_{j=J+1}^{2J} N(\theta_{\sigma(j)}; -1, 1) \right],$$

   where the sum is over all permutations.

2. $Cov(\theta_i, \theta_j) < 0$.

Exchangeability: Example

1. The joint density $p(\theta_1, \ldots, \theta_{2J})$ is

$$\binom{2J}{J}^{-1} \sum_{\sigma(1,\ldots,2J)} \left[\prod_{j=1}^{J} N(\theta_{\sigma(j)}; 1, 1) \prod_{j=J+1}^{2J} N(\theta_{\sigma(j)}; -1, 1) \right],$$

   where the sum is over all permutations.

2. $Cov(\theta_i, \theta_j) < 0$.

3. Correlation $\rightarrow 0$. As $J \rightarrow \infty$, the distinction disappears between (1) independently assigning each j to one of two groups, and (2) picking exactly half of the j's for each group.

Plan

Conjugate Hierarchical Models

- Calculate $p(\phi|y)$ as

$$p(\phi|y) = \int p(\theta, \phi|y)d\theta = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}.$$

Conjugate Hierarchical Models

- Calculate $p(\phi|y)$ as

$$p(\phi|y) = \int p(\theta, \phi|y) d\theta = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}.$$

- Draw $\phi_1, \ldots, \phi_M$ from $p(\phi|y)$.

Conjugate Hierarchical Models

- Calculate $p(\phi|y)$ as

$$p(\phi|y) = \int p(\theta, \phi|y) d\theta = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}.$$

- Draw $\phi_1, \ldots, \phi_M$ from $p(\phi|y)$.
- Draw θ from $p(\theta|\phi, y)$ for each ϕ_i.

Conjugate Hierarchical Models

- Calculate $p(\phi|y)$ as

$$p(\phi|y) = \int p(\theta, \phi|y) d\theta = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}.$$

- Draw $\phi_1, \ldots, \phi_M$ from $p(\phi|y)$.
- Draw θ from $p(\theta|\phi, y)$ for each ϕ_i.
- Draw predictive values y.

Plan

Rat Tumor Example

$$y_j \sim \text{Binomial}(n_j, \theta_j), \theta_j \sim \text{Beta}(\alpha, \beta).$$

- Joint posterior distribution $p(\theta, \alpha, \beta|y)$.
- Conditional posterior distribution $p(\theta|\alpha, \beta, y)$.
- Marginal posterior distribution $p(\alpha, \beta|y)$.

Rat Tumor Example

Setting up priors

Reparametrize (α, β) as $(\text{logit}(\frac{\alpha}{\alpha+\beta}), \log(\alpha+\beta))$.

- Flat priors give improper posterior: $\alpha + \beta \to \infty$.

Rat Tumor Example

Setting up priors

Reparametrize (α, β) as $(\text{logit}(\frac{\alpha}{\alpha+\beta}), \log(\alpha+\beta))$.

- Flat priors give improper posterior: $\alpha + \beta \to \infty$.
- Uniform on $(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2})$,

$$p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2},$$

which corresponds to

$$p\left(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta)\right) \propto \alpha\beta(\alpha+\beta)^{-5/2}.$$

Rat Tumor Example

Setting up priors

- Alternatives
  - $p(\frac{\alpha}{\alpha+\beta}, \alpha+\beta) \propto 1$
  - $p(\alpha, \beta) \propto 1$

Rat Tumor Example

Setting up priors

- Alternatives
  - $p(\frac{\alpha}{\alpha+\beta}, \alpha+\beta) \propto 1$
  - $p(\alpha, \beta) \propto 1$
- Flat prior for $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ on a vague but finite range e.g. $[-10^{10}, 10^{10}]^2$ is not an acceptable solution.

In general, when a likelihood is not integrable, setting a faraway finite cutoff to a uniform prior does not necessarily eliminate the problem.

Rat Tumor Example

- R Demon.
- Estimate $E(\alpha|y)$.
- Sample from $p(\theta|y)$ through
  1. Sample $(\log(\frac{\alpha}{\beta}), \log(\alpha+\beta))$ from grid.
  2. Sample θ_j from $p(\theta_j|\alpha, \beta, y)$.

Gaussian Example

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2); i = 1, \ldots, n_j; j = 1, \ldots, J.$$

Gaussian Example

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2); i = 1, \ldots, n_j; j = 1, \ldots, J.$$

- Estimate θ_j with $\bar{y}_{\cdot j}$ or $\bar{y}_{\cdot\cdot}$, or linear combination?

Gaussian Example

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2); i = 1, \ldots, n_j; j = 1, \ldots, J.$$

- Estimate θ_j with $\bar{y}_{\cdot j}$ or $\bar{y}_{\cdot\cdot}$, or linear combination?
- Hierarchical model

$$p(\theta_1, \ldots, \theta_J) = \prod_{j=1}^{J} N(\theta_j; \mu, \tau^2);$$
$$p(\mu, \tau) = p(\mu|\tau)p(\tau).$$

Gaussian Example

- Conditional posterior $\theta_j | \mu, \tau, y$.

Gaussian Example

- Conditional posterior $\theta_j|\mu, \tau, y$.
- Marginal posterior of hyperparameters

$$\bar{y}_{\cdot j} \sim N(\mu, \frac{\sigma^2}{n_j} + \tau^2).$$

Gaussian Example

- Conditional posterior $\theta_j|\mu, \tau, y$.
- Marginal posterior of hyperparameters

$$\bar{y}_{\cdot j} \sim N(\mu, \frac{\sigma^2}{n_j} + \tau^2).$$

- Posterior distribution of μ given τ.

Gaussian Example

- Conditional posterior $\theta_j|\mu, \tau, y$.
- Marginal posterior of hyperparameters

$$\bar{y}_{\cdot j} \sim N(\mu, \frac{\sigma^2}{n_j} + \tau^2).$$

- Posterior distribution of μ given τ.
- Posterior distribution of τ

$$p(\tau|y) = \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)}$$

Gaussian Example

- Prior distribution for τ.

Gaussian Example

- Prior distribution for τ.
- Non-Bayesian estimate of hyper parameters

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \hat{\tau}^2 = (MS_B - MS_W)/n.$$

Gaussian Example

- Prior distribution for τ.
- Non-Bayesian estimate of hyper parameters

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \hat{\tau}^2 = (MS_B - MS_W)/n.$$

Main problem:

  - Ignore uncertainty of (μ, τ^2)
  - Numerically $\hat{\tau}^2$ might not be positive.

Eight Schools Example

- Separate Estimates.
- Pooled Estimates.
- Hierarchical Model.
- R Demon.

Weakly informative priors for variance parameters

- Uniform prior distributions.
- Inverse Gamma (ϵ, ϵ) prior distributions.
- Half Cauchy prior distributions.
- Application to 8-school example.