HIERARCHICAL MODELS AND SHRINKAGE

September 15, 2019

Jon Fintzi

Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases
National Institutes of Health

In conclusion

Every time we have met, we've talked about:

- Bayesian inference *always* starts with a model for the **joint distribution** of θ and y:.

$$\pi(\theta, y) = f(y|\theta)\pi(\theta) = \pi(\theta|y)m(y).$$

- **Bayes rule** yields the **posterior distribution**

$$\pi(\theta|y) = \frac{f(y,\theta)}{m(y)} = \frac{f(y|\theta)\pi(\theta)}{m(y)} \propto Likelihood \times Prior.$$

.

- All of the information used in the *update* to our prior is encoded in the **likelihood**,

$$L(\mathbf{y}|\theta) = \prod_{i=1}^{N} f(y_i|y_{1,\ldots,i-1}\theta).$$

And last time, we talked about:

- Priors for linear regression parameters.
- Workflow, prior and posterior predictive distributions.
- Failure modes of light tailed priors under poorly chosen scales.
- Weakly informative priors as a starting point.

Lectures 15-17 of Statistical Rethinking

Multilevel/hierarchical models:

- Account for latent structure:
  - Clustering, e.g., students < classrooms < schools < districts, meta-analyses.
  - Heterogeneity, lower level units have individual parameters.
- Shrinkage towards population average.
- Improved out of sample performance, don't want to overfit or underfit.
- Some other topics as well: reparameterization, priors on covariances for subject level parameters,

Plan for today

Shrinkage, hierarchical models, and regularized regression:

- Baseball example[1] - batting ability for players in 1970.
- Three different models - complete, partial, and no pooling of information.
- Briefly talk about sparse regression with horseshoe priors as another example of hierarchical model.

[1] Borrowing heavily from Carpenter (2018)

Take me out to the ballgame

Data from the 1970 Major League Baseball season:

- $N = 18$ players.
- Data: $y_i = Hits_i/AB_i =$ batting average for player i, first $K_i = 45$ at-bats.
- Goal: predict batting average for remainder of the season, $\widetilde{y}_i$.

```
##      Player AB Hits RemainingAB RemainingHits AvgFirst45 AvgRemainder
## 1  Clemente 45   18         367           127  0.4000000    0.3460490
## 2  Robinson 45   17         426           127  0.3777778    0.2981221
## 3    Howard 45   16         521           144  0.3555556    0.2763916
## 4 Johnstone 45   15         275            61  0.3333333    0.2218182
## 5     Berry 45   14         418           114  0.3111111    0.2727273
## 6   Spencer 45   14         466           126  0.3111111    0.2703863
```

Model 1 - Complete Pooling

Use a single quantity, ρ, to represent the probability of a hit for all players.

- Parameter, $\lambda = \text{logit}(\rho) = \log(\rho/(1-\rho)) =$ log-odds of a hit, so the probability of a hit is $\rho = \text{logit}^{-1}(\lambda) = 1/(1+\exp(-\lambda))$.
- Suppose at-bats for each player are independent Bernoulli trials, $y_i \sim Binomial(K_i, \rho) \equiv Binomial(K_i, \text{logit}^{-1}(\lambda))$.
- Complete pooling model, $\pi(\mathbf{Y}, \lambda)$:

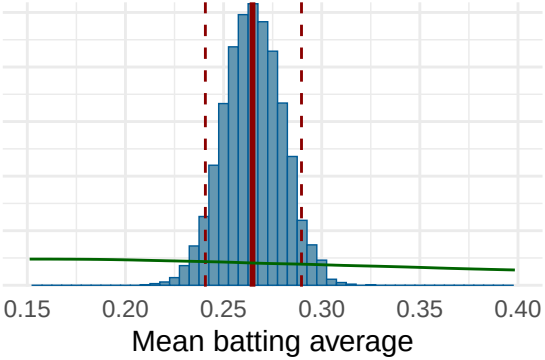
$$\begin{aligned} \pi(\lambda|\mathbf{y}) &\propto \pi(\lambda) L(\mathbf{y}|\boldsymbol{\lambda}), \\ \lambda &\sim N(-1, 1), \\ L(\mathbf{y}|\lambda) &= \prod_{i=1}^{N} Binomial(y_i|K_i, \rho). \end{aligned}$$

- Note, prior and posterior over λ imply a prior and posterior over ρ.
- Prior for λ is weakly informative for ρ; prior median (90% interval) = 0.269 (0.066, 0.656)).

Model 1 - Complete Pooling

Fit the model using `RStanArm` (see Rmarkdown for code).

- Posterior median (90% Credible interval): 0.26 (0.24, 0.29).
- Posterior distribution of mean batting average (blue histogram) and prior (green density):



Model 2 - No Pooling

Use a different quantity, ρ_i, to independently represent the probability of a hit for each player.

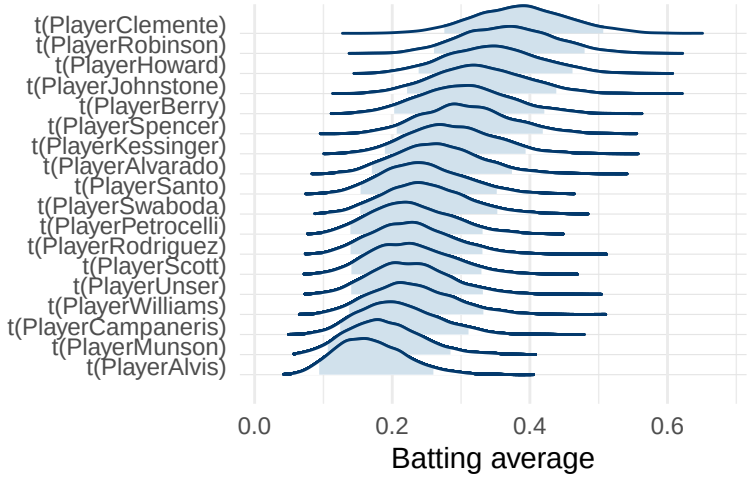
- Parameter, $\lambda_i = \text{logit}(\rho_i) = \log(\rho_i/(1-\rho_i)) =$ log-odds of a hit for player i, so the probability of a hit is $\rho_i = \text{logit}^{-1}(\lambda_i) = 1/(1+\exp(-\lambda_i))$.
- Suppose at-bats for each player are independent Bernoulli trials, $y_i \sim Binomial(K_i, \rho_i) \equiv Binomial(K_i, \text{logit}^{-1}(\lambda_i))$.
- No-pooling model, $\pi(\mathbf{Y}, \boldsymbol{\lambda}) = \pi(\mathbf{Y}|\boldsymbol{\lambda}) \prod \pi(\lambda_i)$:

$$\begin{aligned} \pi(\lambda_i|\mathbf{y}) &\propto \pi(\lambda_i) L(\mathbf{y}|\boldsymbol{\lambda}), \\ \lambda_i &\sim N(-1, 1), \\ L(\mathbf{y}|\lambda_i) &= \prod_{i=1}^{N} Binomial(y_i|K_i, \rho_i). \end{aligned}$$

- Note, independent priors and sampling distributions imply independent posteriors over λ_i.

Model 2 - No Pooling

- Independent posterior distributions of player-specific batting averages are wider.
- Only 45 Bernoulli trials per player, vs. 810 trials with complete pooling.
- Posterior distributions of batting averages:



Model 3 - Partial Pooling

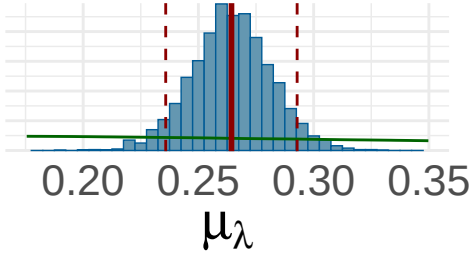
Use a different quantity, ρ_i, for each player while encoding that players are similar, not only to themselves, but also to one another.

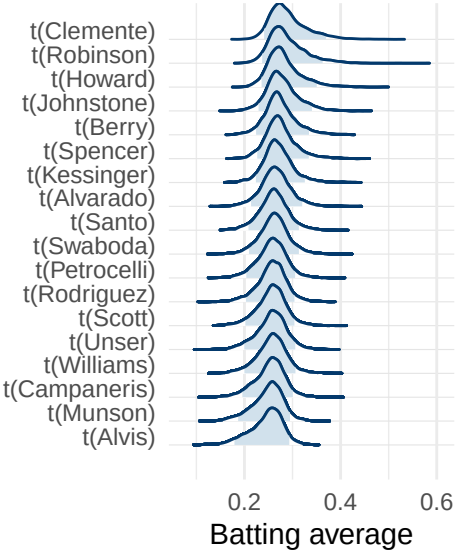
- Now, player-level parameters, λ_i, have a joint distribution, $\lambda_i \sim \pi(\theta_\lambda)$.
- Partial-pooling model:

$$\begin{aligned}
\pi(\lambda_i|\mathbf{y}) &\propto \pi(\lambda_i)L(\mathbf{y}|\boldsymbol{\lambda}), \\
\lambda_i &\sim N(\mu_\lambda, \tau_\lambda^2), \\
\mu_\lambda &\sim N(-1, 1), \\
\tau_\lambda &\sim Exponential(1), \\
L(\mathbf{y}|\lambda_i) &= \prod_{i=1}^{N} Binomial(y_i|K_i, \rho_i).
\end{aligned}$$

- Players are no longer independent in this model, rather, they are *exchangeable*, a priori.
- Exchangeability $\implies \pi(\lambda_1, \lambda_2) = \pi(\lambda_2, \lambda_1)$, and is weaker than independence, $\pi(\lambda_1, \lambda_2) = \pi(\lambda_1)\pi(\lambda_2)$.

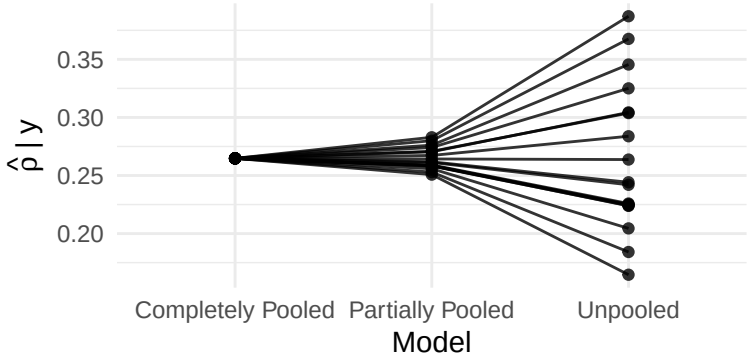
Model 3 - Partial Pooling





Comparing Model Estimates

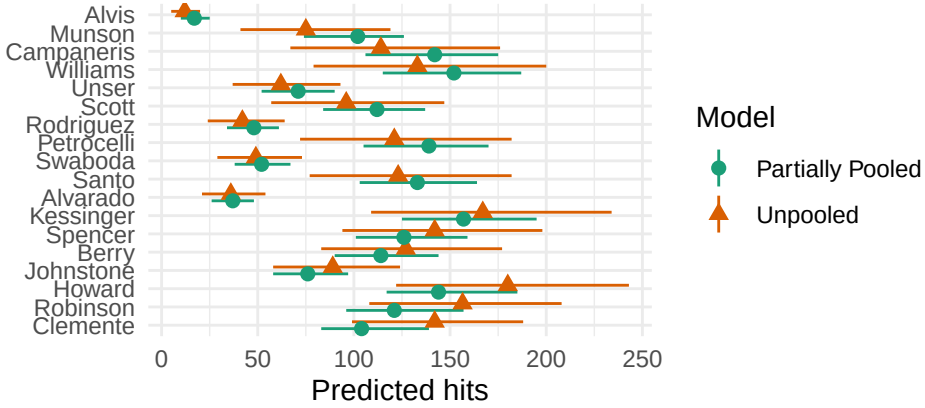
Partially pooled estimates are a weighted average of the unpooled and completely pooled estimates, i.e., we are *shrinking* the unpooled estimates towards the population average hit probability.



Why bother with all this pooling stuff?

A few reasons:

- Arguably more faithful to the data generating process.
- Better out of sample predictive performance:
  - Expected log predictive density: $elpd_{partial\ pool} = -46.6 \pm 2.2; elpd_{nopool} = -53.8 \pm 0.80.$
  - Less uncertainty in prediction intervals:



Summary Before We Continue

Different models for the data and parameters:

- Complete pooling: ignore player labels entirely and lump data.
- No pooling: same as independently analyzing data for each player.
- Partial pooling - players are exchangeable in the prior, estimated batting average depends on each player's data and population average.

Notice that complete and no pooling are the limiting cases of the partial pooling model!

- Partial pooling: $\lambda_i \sim N(\mu_\lambda, \tau_\lambda^2)$
- Complete pooling, all player-level parameters the same: $\lim_{\tau_\lambda \to 0} N(\mu_\lambda, \tau_\lambda^2)$.
- No pooling, player-level parameters unrelated: $\lim_{\tau_\lambda \to \infty} N(\mu_\lambda, \tau_\lambda^2)$.

The Year is 1998 and Steroids are All the Rage!

Fun fact: 1998 was the only year I ever paid attention to baseball.

- Sammy Sosa and Mark Maguire were chasing Roger Maris's home run record.
- I was twelve and made my parents get a newspaper subscription so I could get updated first thing in the morning.
- Turns out sportsmanship didn't make the majors.
- I became disillusioned. Clearly.

The Year is 1998 and Steroids are All the Rage!

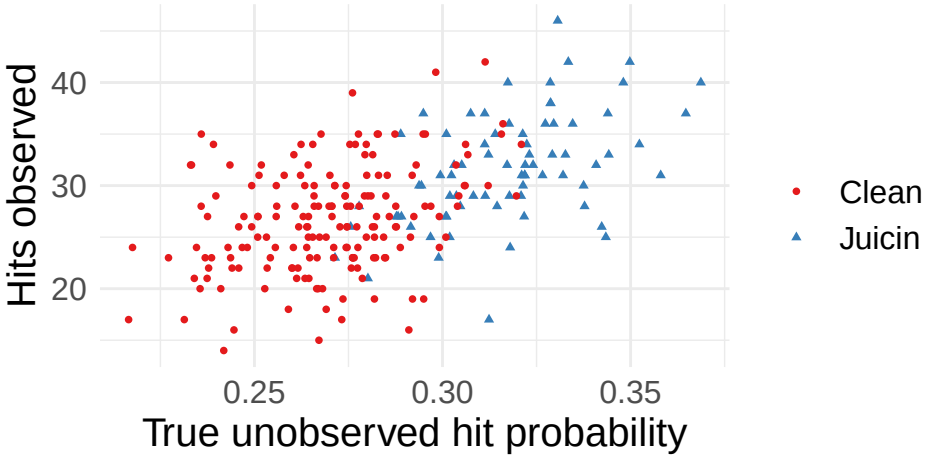
Simulated batting averages over the first 100 at-bats for 250 players under the following model:

$$
\begin{aligned}
Y_i &\sim Binomial(100, \rho_i = \text{logit}^{-1}(\lambda_i)),\ i = 1, \ldots, 250 \\
\lambda_i &= \beta_0 + \beta_1 X_{roids,i} + \beta_2 X_{i,2} + \cdots + \beta_{25} 0 X_{i,25},\ i = 1, \ldots, 250 \\
\text{logit}^{-1}(\beta_0) &= 0.269 \\
\exp(\beta_1) &= 1.25 \\
X_{roids,i} &= 1,\ i = 1, \ldots, 75; X_{roids,i} = 0,\ i = 76, \ldots, 250, \\
\beta_j &\sim N(0, 0.02^2),\ j = 2, \ldots, 25, \\
X_{i,j} &\sim N(0, 1),\ i = 1, \ldots, 250,\ j = 1, \ldots, 25.
\end{aligned}
$$

- The thing that matters is steroid use (assumed observed).
- Nothing else matters, and let's suppose we suspect nothing really matters.

The Year is 1998 and Steroids are All the Rage!

Simulated data:



A Simple Hierarchical Model

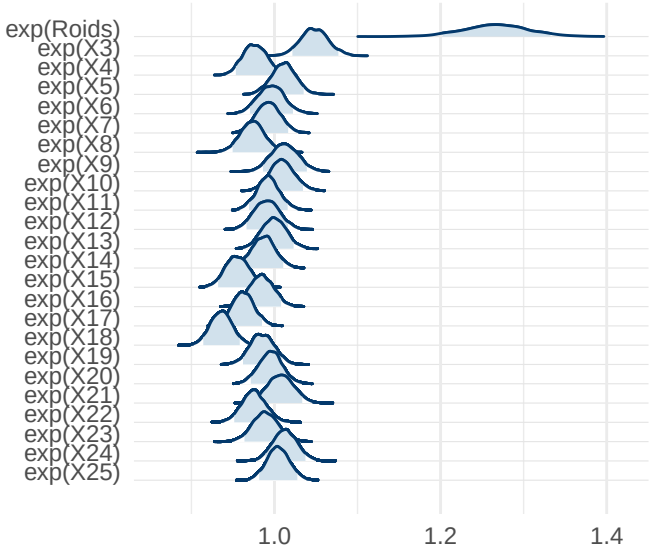
Partial pooling, as before, but with a bit more complexity b/c of covariates:

$$
\begin{aligned}
Y_i &\sim Binomial(100, \rho_i = \text{logit}^{-1}(\lambda_i)),\ i = 1, \ldots, 250 \\
\lambda_i &= \beta_{talent,i} + \beta_1 Z_{roids,i} + \beta_2 X_{i,3} + \cdots + \beta_2 50 X_{i,25},\ i = 1, \ldots, 250 \\
\beta_0 &\sim N(-1, 1), \\
\beta_j &\sim N(0, 0.42),\ j = 1, \ldots, 25.
\end{aligned}
$$

- Priors for μ_λ and τ_λ are weakly informative as before.
- Priors for covariates imply that 90% of their prior mass of the odds ratio of a hit is betwee 0.5x and 2x.
- Note: too many parameters. We know even without fitting this that we're looking for trouble.

Posterior distributions of model parameters

Posterior distributions of model parameters are overly diffuse. Uncertainty is sure to propagate into other distributions of interest, e.g., posterior predictive distributions.



Hierarchical Shrinkage via Horseshoe Priors

Want to incorporate prior knowledge that many parameters are essentially zero, but we don't know which ones.

- When β_i is essentially zero, shrink $\pi(\beta_i|\mathbf{y})$ strongly to zero.
- When β_i is not essentially zero, shrink $\pi(\beta_i|\mathbf{y})$ very little without leaking posterior mass way out into the tails.
- Encode prior information about various aspects of the sparsity, e.g., effective # of non–zero terms, conditional independence structure, etc.
    - Computational tractability.

This is a tall order!

Hierarchical Shrinkage via Horseshoe Priors

Big idea: prior scale for each model component is a product of *global* scale and its own *local* scale.

$$
\begin{aligned}
Y_i &\sim Binomial(100, \rho_i = \text{logit}^{-1}(\lambda_i)),\ i = 1, \ldots, 250 \\
\lambda_i &= \beta_0 + \beta_1 Z_{roids,i} + \beta_2 X_{i,3} + \cdots + \beta_2 50 X_{i,25},\ i = 1, \ldots, 250 \\
\beta_0 &\sim N(-1, 1), \\
\beta_j &\sim N(0, \tau^2 \sigma_j^2),\ j = 1, \ldots, 25, \\
\sigma_j &\sim Cauchy^{+}(0, 1).
\end{aligned}
$$

Hierarchical Shrinkage via Horseshoe Priors

Intuition:

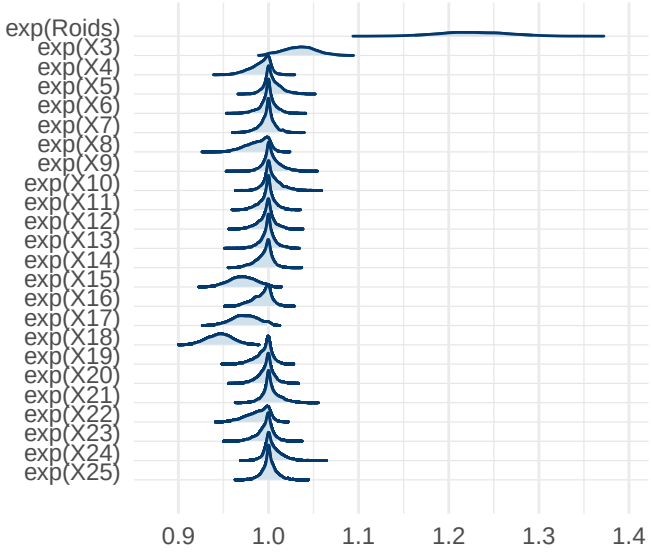
- Global scale parameter τ shrinks β_j globally to 0.
  - Local scales σ_j have Cauchy tails, allowing some β_j to escape shrinkage.
  - Varying $\tau \implies$ more or less sparsity.

Why horseshoe?

- Good theoretical properties,
- Good computational properties,
  - Not gonna talk about either, see Bhadra (2019).
  - Also here for a nice case study.

Hierarchical Shrinkage via Horseshoe Priors

Posteriors for irrelevant parameters are strongly shrunk towards zero, but not the parameter for steroids. Just like we wanted!



Hierarchical Shrinkage via Horseshoe Priors

Shrinking irrelevant parameters results in better out of sample predictive performance:

- Expected log predictive density: $elpd_{HS} - elpd_{partialpool} =$ -5.89 $\pm$ 3.57.
- Less uncertainty in prediction intervals.

Summary

Iteration on the same idea: construct a joint distribution for data and parameters.

- Everything starts with a joint distribution.
- We didn't talk about this a lot today, but incredibly important to simulate from the prior and interrogate the joint prior distribution.
- Posterior predictive checks for model comparison.

Next week is the last lecture. We'll chat about Bayesian handling of missing data.

- Sneak peak, can think of hierarchical models covered today as a sort of missing data problem.
- Lecture 20 of statistical rethinking.

References

A. Bhadra, et al. “Horseshoe Regularization for Machine Learning in Complex and Deep Models.” arXiv preprint arXiv:1904.10939 (2019).

B. Carpenter, et al. “Hierarchical Partial Pooling for Repeated Binary Trials.” `https://cran.r-project.org/web/packages/rstanarm/vignettes/pooling.html` (2018).

B. Efron and C. Morris. “Data analysis using Stein’s estimator and its generalizations.” *Journal of the American Statistical Association* 70.350 (1975): 311-319.

A. Gelman, et al. Bayesian data analysis. Chapman and Hall/CRC, 2013.