STA732
Statistical Inference

Lecture 11: Empirical Bayes and Hierarchical Bayes

Yuansi Chen

Spring 2023

Duke University

https://www2.stat.duke.edu/courses/Spring23/sta732.01/

Recap from Lecture 10

1. Discussed conjugate priors
2. Four typical ways to construct priors
    - Prior data experience
    - Subjective prior
    - Objective prior
    - Convenience prior
3. Pros of Bayes:
    - Straighforward construction of Bayes estimators
    - Bayes optimal
    - Detailed output
4. Cons of Bayes:
    - Difficulty in choosing prior
    - Difficulty in specifying the whole model

Goal of Lecture 11

1. Hierarchical Bayes
2. Empirical Bayes
3. James-Stein estimator

Chap. 15.1, 11.1-2 of Keener or 4.5, 4.6 of Lehmann and Casella

Hierarchical Bayes

Motivation

Recall from last lecture that we can construct prior from previous data experience:

- In a standard Bayesian model $X \sim p_\theta(\cdot), \Theta \sim \Lambda$, we only have one draw of Θ
- If we have previous data with similar structure, we can decide the prior better

Predict a batter's batting average from data $X =$ number of hits $\sim$ Binomial(n, θ).

Prior info:

- most batting averages are between 0.1 and 0.3
- 0.8 is rare
- We can specify the prior using a Beta distribution

Q: how to set α, β in Beta(α, β)?

Photo by Chris Chow on Unsplash

Hierarchical Bayes solution to the prior choice

Suppose we have data from m batters (each batter has data $X_i =$ number of hits, $i = 1, \dots, m$)

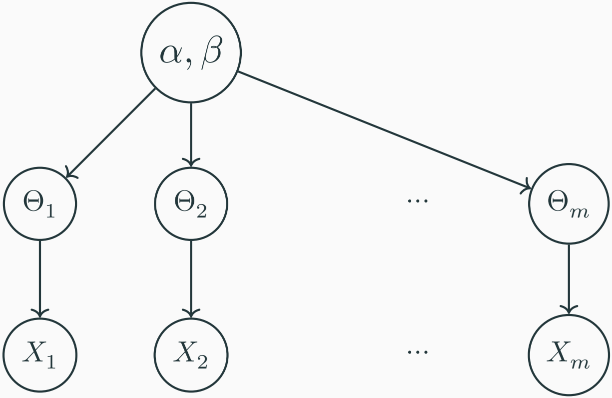
the hierarchical Bayes solution is a hierachical modelling of the batting average by pooling prior info across batters

$$\alpha \sim \text{Exp}(1), \beta \sim \text{Exp}(1), \text{independently}$$

$$\Theta_i \mid \alpha, \beta \overset{\text{i.i.d}}{\sim} \text{Beta}(\alpha, \beta), i = 1, \dots, m$$

$$X_i \mid \Theta_i = \theta_i \overset{\text{indep}}{\sim} \text{Binomial}(n_i, \theta_i), i = 1, \dots, m$$

Graphical model for the hierarchical model



Directed graphical model. The joint density factorizes

$$p(\alpha, \beta, \theta_1, \dots, \theta_m, x_1, \dots, x_m)$$
$$= p(\alpha, \beta) \cdot \prod_{i=1}^{m} p(\theta_i \mid \alpha, \beta) \cdot \prod_{i=1}^{m} p(x_i \mid \theta_i)$$

Posterior computation

To obtain a Bayes estimator, we are interested in the posterior

$$p(\theta_1, \dots, \theta_m \mid x_1, \dots, x_m),$$

It does not have a closed form in this case.

Computational strategy:

Set up a Markov chain with stationary distribution $\propto p(\theta_1, \dots, \theta_m \mid x_1, \dots, x_m)$, run it long enough to get approximate samples

MCMC is not the main focus of this course

Practical implication of the joint modeling of the prior

The posterior for a single parameter also depends on all data

$$p(\theta_1 \mid x_1, \ldots, x_m)$$

Intuitively,

$X_2, \ldots, X_m$ indirectly influence the estimate of θ_1 through the hyperprior, by teaching us what values of θ are more plausible

Examples where hierarchical Bayes may make sense

- Model COVID reproduction number R for multiple countries
- SAT scores collected from five high schools in NC
- Mortality rate after heart attack across 10 hospitals in NYC

Examples where hierarchical Bayes may make sense

- Model COVID reproduction number R for multiple countries
- SAT scores collected from five high schools in NC
- Mortality rate after heart attack across 10 hospitals in NYC

Exercise:

Modelling batting average for players in Major League Baseball

- Shall we always pool the data from all batters?
- If we have batter data from college baseball, should we include them?
- By pooling more data, what estimate is improved and what estimate might deteriorate?

Example: Gaussian hierarchical model

$$\tau \sim \lambda(\tau) \quad \text{e.g. } 1/\tau^2 \sim \text{Gamma}(k, s)$$
$$\theta_i \mid \tau^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2)$$
$$X_i \mid \theta_i, \tau^2 \overset{\text{indep}}{\sim} \mathcal{N}(\theta_i, 1)$$

Compute the posterior mean

- What is the shrinkage factor?
- What is a good estimate of the shrinkage factor?
- How much does the prior on τ matter? The prior on τ does not matter much

Empirical Bayes

Recall the situation in hierarchical Bayes:

- Prior on τ Always ask does the prior matter?
- $\Theta_i \mid \tau$ Key part in hierarchical Bayes
- $X_i \mid \Theta_i, \tau$

Empirical Bayes as a hybrid approach

- Estimate ξ based on all data, e.g. via MLE
- Plug in $\hat{\xi}$ as if it is known

Empirical Bayes applied to Gaussian mean estimation

$$\Theta_i \sim \mathcal{N}(0, \tau^2)$$
$$X_i \mid \Theta_i \sim \mathcal{N}(\theta_i, 1), i = 1, \ldots, m$$

- Compute the posterior mean treating τ is known
- What would be a good estimate of τ^2?

James-Stein estimator

James-Stein estimator as an empirical Bayes estimator

James and Stein proposed a slight different shrinkage factor $m \geq 3$

$$\delta_{\mathsf{JS},i}(X) = \left(1 - \frac{m-2}{\|X\|_2^2}\right) X_i$$

Interpretation

$\frac{m-2}{\|X\|_2^2}$ is UMVU for $\frac{1}{1+\tau^2}$

Prop.

If $Y \sim \mathrm{X}^2_d, d \geq 3$, then

$$\mathbb{E}\left[\frac{1}{Y}\right] = \frac{1}{d-2}$$

JS estimator better than sample mean

In the non-Bayesian Gaussian sequence model, n data points, $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2 \mathbb{I}_d), \theta \in \mathbb{R}^d$ (fixed), $\sigma^2 > 0$ (known), for $d \geq 3$, the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is inadmissible for estimating θ under squared error loss

The JS estimator

$$\delta_{\mathsf{JS}}(X) = \left(1 - \frac{(d-2)\sigma^2/n}{\left\|\bar{X}\right\|_2^2} \right) \bar{X}$$

has strictly lower risk uniformly

Interpretation of the James-Stein paradox

- Could take $n = 1$ by reasoning about sufficient statistics
- The result holds without assumption on the prior model on θ
- There isn't much speciality about 0: for any $\theta_0 \in \mathbb{R}$, we can introduce the estimator

$$\delta' = \theta_0 + \left(1 - \frac{(d-2)}{\|X\|_2^2}\right)(X - \theta_0)$$

- The current justification comes from empirical Bayes. But shrinkage makes sense even without Bayes justification.

Linear shrinkage

Gaussian sequence model $X_i \stackrel{\text{ind.}}{\sim} \mathcal{N}(\theta_i, 1), \theta$ (fixed)
Let $\delta_S(X) = (1 - S)X$ for fixed S.

- Derive the optimal S for the risk under squared error loss
- Give an estimate of the optimal S^*

Stein's lemma

Useful tool for computing risk in Gaussian estimation problems.

Stein's Lemma, univariate, Lem 11.1 in Keener

Suppose $X \sim \mathcal{N}(\theta, \sigma^2)$, $h : \mathbb{R} \to \mathbb{R}$, differentiable, $\mathbb{E}\left|\dot{h}(X)\right| < \infty$, then

$$\mathbb{E}[(X - \theta)h(X)] = \sigma^2 \mathbb{E}[\dot{h}(X)]$$

proof idea: write down the intergrals for $\theta = 0, \sigma^2 = 1$ first

Multivariate Stein's Lemma, Thm 11.3 in Keener

Suppose $X \sim \mathcal{N}(\theta, \sigma^2 \mathbb{I}_d), \theta \in \mathbb{R}^d$, $h : \mathbb{R}^d \to \mathbb{R}^d$, differentiable, $\mathbb{E} \left\| Dh(X) \right\|_F < \infty$.

$$\mathbb{E}\left[(X-\theta)^\top h(X)\right] = \sigma^2 \sum_{i=1}^{d} \mathbb{E} \frac{\partial h_i}{\partial x_i}(X)$$

Stein's unbiased risk estimator (SURE)

We can use Stein's lemma to get unbiased estimator of the risk under squared error loss for any $\delta(X)$, apply with $h(X) = X - \delta(X)$.

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta \left[\|X - \theta - h(X)\|_2^2 \right] \\ &= \mathbb{E}_\theta \|X - \theta\|_2^2 + \mathbb{E}_\theta \|h(X)\|_2^2 - 2\mathbb{E}_\theta \left[(X - \theta)^\top h(X) \right] \\ &= d + \mathbb{E}_\theta \|h(X)\|_2^2 - 2\mathbb{E}_\theta \operatorname{Tr}(Dh(X)) \end{aligned}$$

We get an unbiased estimator for the risk

$$\hat{R}(X) = d + \|h(X)\|_2^2 - 2 \operatorname{Tr}(Dh(X))$$

Calculate the risk of James-Stein

proof idea: apply SURE

Other notes on JS

- δ_{JS} is also inadmissible

$$\delta_{\text{JS+}}(X) = \left(1 - \frac{d-2}{\|X\|_2}\right)_+ X$$

  is strictly better, called positive-part James-Stein estimator
- But the positive-part James-Stein estimator is also inadmissible, although not much improvement can be made, read around Chap 5.5 in Lehmann and Casella.

Summary

- Hierarchical Bayes is good for pooling multiple similar datasets
- Empirical Bayes is similar to Hierarchical Bayes if the hyperprior is not important
- Empirical Bayes gives the James-Stein estimator, which makes the sample mean inadmissible
- Think about shrinkage

What is next?

- Minimax optimality

Thank you