Bayesian Statistics

Empirical Bayes

Brani Vidakovic

Professor

School of Industrial and Systems Engineering and School of Biomedical Engineering, Georgia Tech

Before We Begin…

- Parametric Approach
- Non-Parametric Approach

Empirical Bayes

- Carl Morris (1983, JASA paper) divided Empirical Bayes to <u>parametric</u> and <u>non-parametric</u>.

- **<u>Parametric Approach</u>**

$$X_i \big| \theta_i \overset{ind.}{\sim} f_i(x_\mathrm{i} \big| \theta_i), \qquad i = 1,2,\dots,n$$
$$\theta_i \overset{i.i.d.}{\sim} \pi(\theta_i|\xi), \qquad \xi \text{ is common hyperparameter}$$

Then $m_i(x_i|\xi) = \int f_i(x_i|\theta_i) \cdot \pi(\theta_i|\xi) d\theta_i$.

Also, $m(\underset{\sim}{x}|\xi) = \int \prod_{i=1}^n f_i(x_i|\theta_i) \cdot \prod_{i=1}^n \pi(\theta_i|\xi) d\theta_1 \cdots d\theta_n$

$$= \prod_{i=1}^n \int f_i(x_i|\theta_i) \cdot \pi(\theta_i|\xi) d\theta_i$$
$$= \prod_{i=1}^n m_i(x_i|\theta_i) \quad \boxed{\text{independent}}$$

From $m(\underset{\sim}{x} \mid \xi) = \prod_{i=1}^{n} m_i(x_i \mid \xi)$ $\Rightarrow$ X_i are marginally independent if $\theta_i \overset{i.i.d.}{\sim} \pi(\theta_i \mid \xi)$.

- If $f_i \equiv f$, then X_i are $i.i.d.$ (marginally)
  Also, the posterior is

$$\pi(\theta_i \mid X_i, \xi) = \frac{f(x_i \mid \theta_i) \cdot \pi(\theta_i \mid \xi)}{m(x_i \mid \xi)}.$$

- ξ is unknown, can be estimated from $X_1, X_2, \ldots, X_n$ via
  - MLE (called MLE II approach)
  - MM (moment matching)

Jeremy in Empirical Bayes

Let $\underset{\sim}{X} = (98, 107, 89, 88, 108)$ be Jeremy's scores on $n = 5$ independent IQ tests.

For this data:

$$X_i | \theta_i \overset{ind.}{\sim} N(\theta_i, \sigma^2), \quad \sigma^2 \text{ known and } \sigma^2 = 80.$$

$$\theta_i \overset{i.i.d.}{\sim} N(\mu, \tau^2), \quad \text{Goal: estimate } \theta_i's.$$

$$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2 + \tau^2)$$

$$m\left(\underset{\sim}{x} \middle| \mu, \tau^2\right) = \prod_{i=1}^{n} m(x_i | \mu, \tau^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(\sigma^2 + \tau^2)}} e^{-\frac{(x_i - \mu)^2}{2(\sigma^2 + \tau^2)}}$$

Then, the MLE of μ is $\hat{\mu} = \bar{X}$ and of τ^2 is

$\hat{\tau}^2 = (s^2 - \sigma^2)_+ \equiv \max\{0, s^2 - \sigma^2\}$, s^2 – sample variance for X.

- With these estimators from the data $\underset{\sim}{X}$, the (estimated) posterior becomes:

$$\pi(\theta_i | X_i, \hat{\mu}, \hat{\tau}^2) = N\left(\hat{B}\hat{\mu} + \left(1 - \hat{B}\right)x_i, \left(1 - \hat{B}\right) \cdot \sigma^2\right),$$

where $\hat{\mu} = \bar{X}, \hat{\tau}^2 = (s^2 - \sigma^2)_+$, and $\hat{B} = \frac{\sigma^2}{\sigma^2 + \hat{\tau}^2}$.

Thus, for Jeremy's data: $s^2 = 101$,

$$\hat{B} = \frac{\sigma^2}{\sigma^2 + \hat{\tau}^2} = \frac{80}{80 + (101 - 80)} = \frac{80}{101}$$

$\hat{\theta}_1 = \frac{80}{101} \cdot 98 + \frac{21}{101} \cdot 98 = 98$; $\hat{\theta}_2 = \frac{80}{101} \cdot 98 + \frac{21}{101} \cdot 107 = 99.8713$,

etc.

- **Example:** $X_i \sim \text{Pois}(\lambda_i),\ i = 1,2,\dots,n$
  $\lambda_i \sim \text{Exp}(\mu),\ \pi(\lambda_i) = \mu e^{-\mu\lambda_i}$

  Find EB estimators of λ_i.

- Bayes estimator $X_i \sim \text{Pois}(\lambda_i)$, $\lambda_i \sim \text{Exp}(\mu)$

$$\lambda_i | X_i \sim \text{Ga}(x_i + 1, 1 + \mu)$$

$\mathbb{E}(\lambda_i | X_i) = \frac{x_i+1}{1+\mu}$, but μ may not be known…

- Empirical Bayes

$$m(x_i) = \int_0^{+\infty} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i} \cdot \mu e^{-\lambda_i \mu} d\lambda_i$$

$$m(x_i) = \int_0^{+\infty} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i} \cdot \mu e^{-\lambda_i \mu} d\lambda_i$$

1

$\Gamma(x_i + 1) = x_i!$

$$= \frac{1}{(1+\mu)^{x_i+1}} \cdot \mu \int_0^{+\infty} \frac{(1+\mu)^{x_i+1} \cdot \lambda_i^{x_i}}{\Gamma(x_i+1)} \cdot e^{-(1+\mu)\lambda_i} d\lambda_i$$

as integral of pdf of gamma $\mathrm{Ga}(x_i + 1, 1 + \mu)$

$$= \left(\frac{1}{1+\mu}\right)^{x_i} \frac{\mu}{1+\mu}, \quad x_i = 0,1,\ldots$$

This is geometric distribution!

Denote $\frac{\mu}{1+\mu} = p$ $\Longrightarrow$ $\mathrm{Ge}(p)$: $\quad P(X_i = x_i) = (1-p)^{x_i} \cdot p$

$$L = \prod_{i=1}^{n} m(x_i) = (1-p)^{\sum x_i} \cdot p^n$$

$$l = \log L = \sum x_i \cdot \log(1-p) + n \cdot \log p$$

$$l' = -\frac{\sum x_i}{1-p} + \frac{n}{p} = 0 \Longrightarrow \hat{p} = \frac{n}{n + \sum x_i} = \frac{1}{1 + \bar{x}}$$

Thus, $\frac{\mu}{1+\mu} = \frac{1}{1+\bar{x}}$ $\longrightarrow$ $\boxed{\hat{\mu} = \frac{1}{\bar{x}}}$

Back to Bayes estimator with μ estimated from the data:

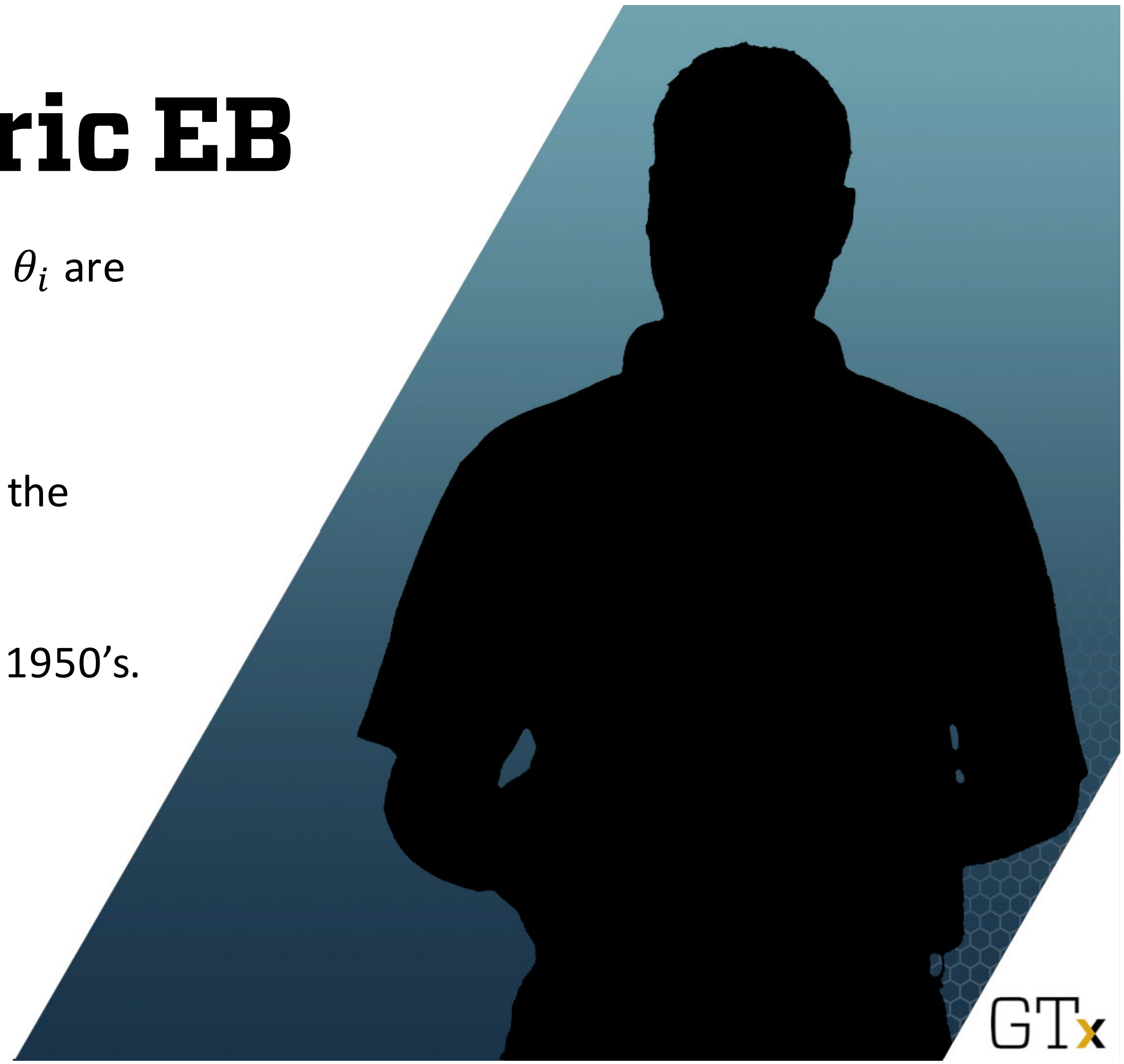
$$\underbrace{\frac{X_i + 1}{1 + \mu}}_{\text{Bayes}} \longrightarrow \underbrace{\frac{X_i + 1}{1 + \hat{\mu}}}_{\text{EB}} = \frac{X_i + 1}{1 + \frac{1}{\bar{X}}} = \frac{\bar{X}}{1 + \bar{X}} \cdot (X_i + 1).$$

Thus,

$$\boxed{\hat{\lambda}_i = \frac{\bar{X}}{1 + \bar{X}}(X_i + 1)}$$

Nonparametric EB

- We assume only that parameters θ_i are $i.i.d.$, no family of distribution is specified.
- Use data to estimate marginal or the prior directly.
- Pioneered by Herbert Robbins in 1950's.



Example:

Let $X_i|\lambda_i \sim \mathrm{Pois}(\lambda_i), i = 1, \dots, n$

$$\lambda_i \overset{i.i.d.}{\sim} \pi(\lambda_i)$$

$$\hat{\lambda}_i = \frac{\int \lambda_i \cdot \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i} \cdot \pi(\lambda_i) d\lambda_i}{\int \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i} \cdot \pi(\lambda_i) d\lambda_i} = \frac{(x_i+1) \int \frac{\lambda_i^{x_i+1}}{(x_i+1)!} e^{-\lambda_i} \cdot \pi(\lambda_i) d\lambda_i}{\int \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i} \cdot \pi(\lambda_i) d\lambda_i} = (x_i + 1) \frac{m_\pi(x_i+1)}{m_\pi(x_i)}$$

Given $X_1, \dots, X_n$, estimate m as $\hat{m}$, and use $\hat{m}$ in

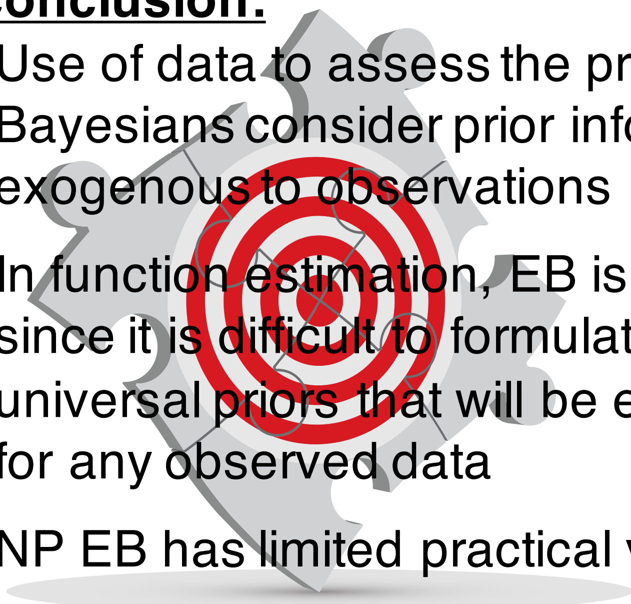
$$(\hat{\lambda})_{\mathrm{EB}} = (x_i + 1) \cdot \frac{\hat{m}(x_i+1)}{\hat{m}(x_i)}$$

- Trivial: $\hat{m}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = x_i)$
  $=$ relative frequency of $x_i's$ in $X_1, \dots, X_n$

$$(\hat{\lambda}_i)_{\mathrm{EB}} = (x_i + 1) \frac{\hat{m}(x_i+1)}{\frac{1}{n}+\hat{m}(x_i)}$$

- Better estimators use smooth estimation of $m(x)$.

Summary



<u>In conclusion:</u>

- Use of data to assess the prior. Bayesians consider prior information exogenous to observations
- In function estimation, EB is popular since it is difficult to formulate universal priors that will be efficient for any observed data
- NP EB has limited practical value
- Instead of EB, will use hierarchical Bayes' models