Empirical Bayes

- Combining frequentist & Bayesian methods
- origin 50-70's "missing species" problem, but resurge now due to high-dimensional analysis problems in eg. biology.
- Key idea: Bayesian methods sensitive to choice of prior
  (can we match parameters of the prior to data?
  Not if) we only have one analysis problem BUT
  if) several analysis problems share a common prior....

Missing species problem

words = Shakespeare's vocabulary

words in 'Complete works of Shakespeare'

// less than, since some words that S knew do not appear in 'Complete works'.

- n_x = # unique words that appear exactly x times in 'Complete works'
  ex. from Efron 2003 $n_1 = 14376$
  $n_2 = 4343$
  $\sum_{x \geq 1} n_x = 31534$
- n_0 = # words not appearing in 'Complete works' but known to Shakespear - the "missing species"

How can we estimate this?

- J = Shakespeare's vocabulary = 31534 + no

- Assume a Poisson model for word counts

$$P(X_j = x) = e^{-\lambda_j} \lambda_j^x / x! \qquad x = 0, 1, \ldots$$

distribution for word j

$G(\lambda) = \#\{\lambda_j \le \lambda\}/J$ is the cdf for the poisson parameters across all words

- Select a word at random. Then, using Bayes rule, we have

$$E(\lambda|x) = \frac{\int_0^\infty e^{-\lambda} \lambda^{x+1}/x!\, dG(\lambda)}{\int_0^\infty e^{-\lambda}\lambda^x/x!\, dG(\lambda)} = (x+1)\frac{\int_0^\infty e^{-\lambda}\lambda^{x+1}/(x+1)!\, dG(\lambda)}{\int_0^\infty e^{-\lambda}\lambda^x/x!\, dG(\lambda)}$$

$$= (x+1)\frac{E_G[n_{x+1}]}{E_G[n_x]}, \text{ where } E_G[n_x] = J \cdot \int_0^\infty e^{-\lambda}\frac{\lambda^x}{x!}\, dG(\lambda)$$

- We substitute the expected value of n_x wrt G by the observed $\Rightarrow$

$$E(\lambda|x) \hat{=} (x+1)\frac{n_{x+1}}{n_x}$$

- Example: $x = 1$ (words that are singletons)

$E(\lambda|x=1) \hat{=} 0.604$ using numbers above from Efron

Interpretation: if we discover a new work by Shakespear, we expect singletons to appear less frequently. In fact, if new work is W words long, we expect the 14376 singletons to appear 0.604×14376 times $\times \left(\frac{W}{\text{total wordlength}}\right)$

- Missing species problem

Define $\lambda_+ = \sum_{j=1}^{J} \lambda_j$, total intensity of words in 'Complete works'

- Can be estimated by $N = 884647$, the total # words, counting repetitions, in 'Complete works'.

Define $r_0 = \sum_{x_j=0} \lambda_j \Big/ \lambda_+$ = proportion of total expectation in the missing class

total intensity of missing words

- From above, $n_0 * E(\lambda | k=0) \hat{=} \frac{n_1}{n_0}$, $n_0 = n_1 \hat{=}$ numerator of r_0

$\Rightarrow \hat{r}_0 = \frac{n_1}{N} = 0.016$ here

Interpretation: next "new" word has probability 0.016 of not being in the 'Complete works'

General problem

$f(\underset{\sim}{x}|\theta)$ data generative distribution

$\pi(\theta)$ prior ⇓ can have a huge impact

$\pi(\theta|\underset{\sim}{x})$ posterior

Empirical Bayes; prior now a <u>family</u> of distributions, indexed by hyper parameter $\pi(\theta|\lambda)$

Example

(A) Bayes: $y \sim Bin(n,p)$ $\quad \pi(p) = 6p(1-p)$

$$\Rightarrow f(y,p) = 6\binom{n}{y} p^{y+1}(1-p)^{n-y+1}$$

$$\Rightarrow \text{marginal } m(y) = 6\binom{n}{y}\frac{\Gamma(y+2)\Gamma(n-y+2)}{\Gamma(n+4)} \quad \text{Beta-Binomial}$$

$$\Rightarrow \pi(p|y) = \frac{\Gamma(n+4)}{\Gamma(y+2)\Gamma(n-y+2)} p^{y+1}(1-p)^{n-y+1} \quad \text{Beta}$$

updated form of the prior

$$\Rightarrow E(p|y) = \left(\frac{n}{n+4}\right)\left(\frac{y}{n}\right) + \left(1-\frac{n}{n+4}\right)\cdot\frac{1}{2}$$

$\hat{p}$: freq. estimate ↑ prior mean

(3) Empirical Bayes

$$\Rightarrow \pi(p|\alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha)}(p(1-p))^{\alpha-1}$$

, family indexed by α

- $\pi(p|y,\alpha) = \text{Beta}(y+\alpha, n-y+\alpha)$

$$E(p|y,\alpha) = \left(\frac{n}{n+2\alpha}\right)\hat{p} + \left(1 - \frac{n}{n+2\alpha}\right)\cdot\frac{1}{2}$$

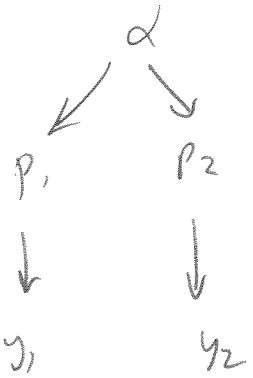
What should α be?

- estimate from marginal $m(y|\alpha)$ (Beta-Binomial)
  - can't - only have one observation.

⇒ Key to EB ⇒ commonality between several analyses.

What if: $y_1|p_1 \sim \text{Bin}(n, p_1)$, $y_2|p_2 \sim \text{Bin}(n, p_2)$

$p_1|\alpha \sim \text{Beta}(\alpha,\alpha)$ ⟺ (shared prior) $p_2|\alpha \sim \text{Beta}(\alpha,\alpha)$



Here, marginal for y_1, y_2 ⇒ $E(y_i) = \frac{n}{2}$

$$V(y_i) = \frac{n}{4}\cdot\frac{n+2\alpha}{2\alpha+1}$$

Use e.g. MoM (method of moments) matching

$$\text{Var}\left(\underline{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) \overset{\wedge}{=} \frac{n}{4}\cdot\frac{n+2\alpha}{2\alpha+1} \Rightarrow \text{solve for } \alpha$$

Clearly, the more parallel problems you have, the better your estimate of α.

Sometimes Stein's problem is used to illustrate the EB principle. Stein deals with estimating a mean-vector in high-dimensional space.

$X_1 \sim N(\theta_1, 1)$
$X_2 \sim N(\theta_2, 1)$ or $\underset{\sim}{X} \sim N_p\left(\underset{\sim}{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}, I\right)$
$\vdots$

$X_p \sim N(\theta_p, 1)$

Using l_2 loss $\Rightarrow$ Find estimate $\hat{\theta}$ to minimize $\sum_{i=1}^{p} (\hat{\theta}_i - \theta_i)^2$

Risk $E_\theta\left[\sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2\right]$

Usual, naive estimate is $\hat{\theta}^0 = \underset{\sim}{X}$ which has risk p.

Stein's estimator $\hat{\theta}_j^S = \left(1 - \frac{p-2}{\sum_{i=1}^{p} X_i^2}\right) X_j \quad , j = 1, \ldots, p \geq 2$

"shrinkage"

Now, the risk of $\hat{\theta}^S$ is less than p $(p > 2)$!!!

That is, we can uniformly improve on the naive estimate using shrinkage.

Relation to EB?

- $X_i \sim N(\theta_i, 1)$
- prior $\theta_i \sim N(0, \tau^2)$ ⇐ shared information across dimension problems θ_i.
- $E(\theta_i \mid x_i, \tau^2) = \frac{\tau^2}{\tau^2+1} x_i = \left(1 - \frac{1}{\tau^2+1}\right) x_i$
- marginal $x_i \sim N(0, 1+\tau^2)$

$$\Rightarrow \frac{\sum_{i=1}^{p} x_i^2}{\tau^2+1} \sim \chi^2_p$$

$$\Rightarrow E\left[\frac{p-2}{\sum_{i=1}^{p} x_i^2}\right] = \frac{1}{\tau^2+1}$$

(mean of inverse-χ^2)

substitute for τ^2

⇒ Stein's estimator

So Stein = EB, where info of prior (spread of θ around 0) was inferred from sharing problem of estimating p d.m mean.

EB - general setup

Posterior - inference when hyperparameters estimated from the marginal $m(y|\lambda)$ across problems.

- may be difficult to compute either MLE of $m(y|\lambda)$ or posterior $\Rightarrow$ EM, MCMC may be needed.

$f_{\theta,\lambda}(y)\,\pi(\theta|\lambda) \Rightarrow \pi(y,\theta|\lambda) \Rightarrow$ posterior $\pi(\theta|y,\lambda) = \frac{\pi(y,\theta|\lambda)}{\pi(y|\lambda)}$

Bayes:

λ assumed known & fixed or from known distribution

$\pi(\theta|y) = \int \pi(\theta|y,\lambda)\, w(\lambda|y)\, d\lambda$

used for inference

EB:

$\hat{\lambda} = \arg\max \pi(y|\lambda)$

$\Rightarrow$ inference from $\pi(\theta|y,\hat{\lambda})$

Example from Efron 2023, multiple testing

- Data on 1640 genes in 24 control and 24 cancer samples
- Which genes are "differentially expressed" - mean level of gene different in cancer compared to control samples
- ALL: - Perform 1640 t-tests
  - Use Bonferroni correction to decide which genes have significantly different mean value between two sample types.

ALL2 EB; prior that gene expression data comes from two distributions

f_0 → no mean difference between control and cancer
e.g. $y_i \sim N(0, \sigma_i^2)$

f_1 → mean difference

Assume $f(y) = p_0 f_0(y) + p_1 f_1(y)$ pdf across 1640 genes.

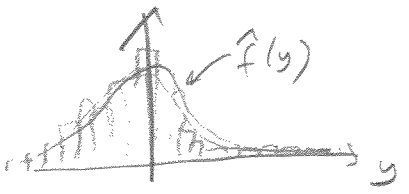
(mixture model)

Bayes rule $p_1(y) = 1 - p_0 f_0(y) / f(y)$

EB; plug in an estimate $\hat{f}(y)$ in the above

$$\hat{p}_1(y) = 1 - p_0 f_0(y) / \hat{f}(y)$$

e.g. NP density estimate of mixture distribution across 1640 genes



Need also choose p_0

- cx $p_0 = 1$ (conservative)
- p_0 s.t. $\hat{p}_1(y)$ estimate always positive: $p_0 = \min\left(\frac{f_0(y)}{\hat{f}(y)}\right)$

Ex - variable selection problem

Check J. Berger, E. George / D. Foster, E. George

and new review paper in AoS 2010

by J. Scott and J. Berger

Idea (sketch)

$$y_i = \alpha + X_{ij1}\beta_{j1} + \quad + X_{ijk}\beta_{jk} + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \phi^{-1})$$

Select $\underline{k}$ out of $\underline{m}$ parameters to be in the model

so $dim(X) = n \times m$

- M_γ = submodel indexed by γ s.t. $\gamma_i = \begin{cases} 0 & \beta_i = 0 \\ 1 & \text{o/w} \end{cases}$

  $\Rightarrow \beta_\gamma$ nonzero coefficient vector

  $k_\gamma \le m$ # nonzeroes

- γ exists on 2^m space of combinations, $p(M_\gamma)$ prior on each model

$$f(y|M_\gamma) = \int f(y|\theta_\gamma, \phi)\, \pi(\theta_\gamma|\theta)\, d\theta_\gamma\, d\phi$$

($\pi(\theta_\gamma|\theta)$: parameter prior, e.g. $N(\theta, i^2)$ etc.)

- Inference, posterior $p(M_\gamma|y) \propto p(M_\gamma) f(y|M_\gamma)$

or

$$p_i = P(\gamma_i \neq 0 | y) = \sum_\gamma 1_{\{\gamma_i = 1\}} P(M_\gamma|y)$$

Bayes $\Longrightarrow$ $P(M_\gamma|p) = p^{k_\gamma}(1-p)^{m-k_\gamma}$ for some choice of p which relates to prior belief of model size

EB $\Rightarrow$ $\hat{p} = \underset{p\in[0,1]}{\operatorname{argmax}} \sum_\gamma p(M_\gamma|p) f(y|M_\gamma)$ by EM or numerical optimizer

$\Downarrow$

$p(M_\gamma|y) \propto \hat{p}^{k_\gamma}(1-\hat{p})^{m-k_\gamma} f(y|M_\gamma)$