MA677 Final Project

CASI-Chap 6 - Empirical Bayes

Yuta Tsukumo

May 6th 2024

1 Overview of this Chapter

The Empirical Bayes Method, explored in this chapter, represents a groundbreaking statistical approach pioneered in the 1950s by Robbins, Good and Toulmin, Stein, among others. Its significant breakthrough lies in its ability to estimate parameters solely based on observed data, without relying on assumed probability distributions or prior knowledge, which are key features of traditional Bayesian methods.

Specifically, as demonstrated in Robbins' formula below, this method follows a nonparametric approach that doesn't incorporate prior probability distributions into the equation.

$$\hat{E}\{\theta_i|x_i\} = \frac{(x_i+1)\cdot y_{x_i+1}}{y_{x_i}} \tag{1}$$

Harold Robbins coined the term "Empirical Bayes" because deriving the prior distribution directly from the data epitomizes an empirical approach. In this sense, Empirical Bayes exhibits characteristics of both frequentist and Bayesian methodologies. This chapter also delves into Fisher's approach for The Missing-Species Problem, now recognized as parametric Empirical Bayes.

2 Mathematics

Derivation of Robbins' formula If $\theta \sim g(\theta)$, where $g(\theta)$ is the prior density, and $x|\theta \sim f(x|\theta)$, then according to Bayes rule, the posterior density of θ given x is

$$g(\theta|x) = \frac{g(\theta)f(x|\theta)}{f(x)}, \quad \text{where } f(x) \text{ is the marginal density of } x,$$

$$f(x) = \int_\Theta f(x|\theta)g(\theta)\,d\theta,$$

with Θ the set of possible θ values.

For the Poisson density $f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}$ we get

$$f(x) = \int_\Theta \frac{\theta^x e^{-\theta}}{x!} g(\theta)\,d\theta,$$

and $g(\theta|x) = g(\theta)$.

This yields the posterior expectation

$$\begin{aligned} &E\{\theta|x\} \\ &= \int_0^\infty \theta g(\theta|x)\,d\theta \\ &= \int_0^\infty \frac{e^{-\theta}\theta^{x+1}}{x!} g(\theta)\,d\theta \Big/ f(x) \\ &= (x+1)\frac{\int_0^\infty \frac{e^{-\theta}\theta^{x+1}}{(x+1)!} g(\theta)\,d\theta}{f(x)}. \end{aligned}$$

$$E(\theta|x) = (x+1)f(x+1)/f(x).$$

In the empirical Bayes situation we don't know $f(x)$ or $f(x+1)$ but they are easy to estimate. The expected value of yx/N is $f(x)$. Substituting yx/N for $f(x)$ and $yx+1/N$ for $f(x+1)$ in the bottom line of the equation gives Robbins' formula(1).

3 Historical Progression of Empirical Bayes

The advent of the electronic computer after World War II was a pivotal event in the history of statistics, freeing the field from the constraints of slow machine computation and allowing for the emergence of new statistical techniques. in the 1950s, the pace of change was slow, with continued emphasis on classical statistical topics. Nevertheless, the following decades were marked by significant transformations due to computer-based statistical methodologies. However, the development of Empirical Bayesian methodology, even though its initial stages began in the 1940s, progressed slowly due to a lack of suitable data sets rather than computational limitations. Nevertheless, the abundance of modern scientific data has prompted a resurgence of interest in empirical Bayesian methods, highlighting their new relevance and importance in modern statistical practice.

1943: Fisher

In 1943, Fisher presented a parametric solution to the missing species problem. He analyzed butterfly data collected by Corbet in Malaysia and estimated the number of unknown species. Fisher used a Gamma distribution as a prior to solve the missing species problem.

1956: Robbins

In 1956, Robins proposed Empirical Bayes. He introduced a method to estimate parameters θ_i without directly specifying the underlying probability density function $g(\theta)$. Robins demonstrated that other observations x_j $(j \neq i)$ could also contribute to the estimation of parameter θ_i. Notably, his formulation, captured in Robbins' formula (1), is nonparametric, as it doesn't depend on any explicit parametric form for $g(\theta)$. In fact, $g(\theta)$ is not explicitly estimated in the proof of (1.3). Robins' derivation finesses the role of $g(\theta)$, highlighting the self-sufficiency of his method.

1956: Good and Toulmin

Also in 1956, Good and Toulmin proposed a nonparametric solution to the missing species problem. They assumed that parameters θ_i followed a Poisson process and used Empirical Bayes methods to estimate the number of unknown species.

1961: James and Stein

In 1961, James and Stein introduced the James-Stein estimator as an alternative to maximum likelihood estimators for normal distribution observations. These estimators were

shown to have smaller expected mean squared error than traditional maximum likelihood estimators.

1987: Efron and Thisted

In 1987, Efron and Thisted applied Empirical Bayes methods to analyze vocabulary data from Shakespeare's works, addressing problems such as estimating new works or attributing authorship.

4 Statistical Practice Implications

In the following, we will discuss how Empirical Bayes methods can be used from a statistical practice perspective.

Let's consider the scenario where an automobile insurance company needs to set insurance premiums for policyholders. For the company, making a profit is essential, so it needs to determine how much premium each policyholder should pay, considering the risk involved. Therefore, being able to predict the expected number of insurance claims a policyholder will make annually is of significant importance to the insurance company.

Using traditional Bayesian methods, we can estimate the rate at which policyholders make insurance claims annually for each frequency of claims using the following formula:

$$\text{Posterior}(\lambda_k|n_k, T_k) = \frac{\text{Likelihood}(n_k, T_k|\lambda_k) \cdot \text{Prior}(\lambda_k)}{\text{Marginal Likelihood}(n_k, T_k)}$$

Here's the breakdown of each component:

- $\text{Posterior}(\lambda_k|n_k, T_k)$ represents the posterior distribution, denoting the probability of λ_k given n_k and T_k.
- $\text{Likelihood}(n_k, T_k|\lambda_k)$ is the likelihood function, indicating the probability of observing n_k and T_k given λ_k.
- $\text{Prior}(\lambda_k)$ stands for the prior distribution, indicating our initial belief or knowledge about λ_k.
- $\text{Marginal Likelihood}(n_k, T_k)$ represents the marginal likelihood, showing the probability of observing n_k and T_k irrespective of λ_k.

However, the prior distribution required for this Bayesian method may not always be available, especially in cases where there isn't enough previous data. For instance, for newly established insurance services with limited past data, obtaining a reliable prior distribution could be challenging.

In such cases where past distributions are not available, the real magic of Robbins' advocated empirical Bayes formula lies in its ability to estimate the expected number of insurance claims based on the vast amount of current-year data that insurance companies possess.

Empirical Bayes has numerous practical applications beyond insurance. For example, let's consider a scenario where a city's education board is organizing a marathon event for middle schools for the first time. Although injuries during marathons are rare, it's essential for the education board to estimate the expected number of injuries at a single school during the event to assess the necessary medical support. Even without sufficient past event data, the empirical Bayes formula allows the education board to estimate the injury rates based on data collected from schools participating in the current event.

Overall, empirical Bayes offers a powerful tool for making informed decisions in situations where traditional Bayesian methods face limitations due to the unavailability of reliable prior distributions.

5 Key Computational Elements

Below, I will provide a sample R code demonstrating the method of categorizing policyholders based on the number of claims made this year and estimating the expected number of claims for each category using Robbins' formula.

We assume a sample size of 10 thousand policyholders, with claims following a Poisson distribution with a mean of $\lambda = 2$.

```r
library(tibble)

# Generate samples from a Poisson distribution
lambda <- 2
claims <- rpois(n = 10000, lambda = lambda)

# Estimate the expected number of claims for the next year using Robbins' formula
next_year_claims_estimate <- numeric(11)
for (i in 0:10) {
  next_year_claims_estimate[i + 1] <- (i + 1) * sum(claims == i + 1) / sum(claims == i)
}

# Create a table showing the results including this year's claims
result_table <- tibble(
  Claims = 0:10,
  This_Year_Claims = as.numeric(table(claims))[1:11],
  Next_Year_Claims_Estimate = next_year_claims_estimate
)
print(result_table)
```

A tibble: 11 × 3

Claims <int>	This_Year_Claims <dbl>	Next_Year_Claims_Estimate <dbl>
0	1381	1.919623
1	2651	2.046020
2	2712	2.008850
3	1816	1.971366
4	895	2.167598
5	388	1.747423
6	113	1.920354
7	31	2.322581
8	9	2.000000
9	2	5.000000

1–10 of 11 rows Previous 1 2 Next

Figure 1: Estimated Number of Claims using Robbins' Formula (N = 10,000)

Next, we will show the same calculation with a much smaller sample size (n = 100).

```
library(tibble)

# Generate samples from a Poisson distribution
lambda <- 2
claims <- rpois(n = 100, lambda = lambda)

# Estimate the expected number of claims for the next year using Robbins' formula
next_year_claims_estimate <- numeric(11)
for (i in 0:10) {
  next_year_claims_estimate[i + 1] <- (i + 1) * sum(claims == i + 1) / sum(claims == i)
}

# Create a table showing the results including this year's claims
result_table <- tibble(
  Claims = 0:10,
  This_Year_Claims = as.numeric(table(claims))[1:11],
  Next_Year_Claims_Estimate = next_year_claims_estimate
)
```

```
19 print(result_table)
```

A tibble: 11 × 3

Claims <int>	This_Year_Claims <dbl>	Next_Year_Claims_Estimate <dbl>
0	12	2.333333
1	28	1.785714
2	25	2.640000
3	22	1.272727
4	7	2.857143
5	4	1.500000
6	1	7.000000
7	1	0.000000
8	*NA*	NaN
9	*NA*	NaN

1–10 of 11 rows Previous 1 2

Figure 2: Estimated Number of Claims using Robbins' Formula (N = 100)

The two results with different sample sizes show that the expected value estimated by Empirical Bayes has greater variability when the sample size is small. Therefore, while empirical Bayes is certainly useful, it is considered necessary to be very careful in interpreting its results when the prior distribution estimated from the current data does not have a large enough sample from a reliability standpoint.

6 Questions

1. What distributions are more accurate for prediction by the Empirical Bayes Method, and conversely, what distributions are less accurate for prediction?
2. Can such high and low accuracy distributions be explained mathematically?

Github Repository Link

`https://github.com/yu99t/MA677_FinalProject`

References

[1] Publisher: Cambridge University Press
Online publication date: July 2016
Print publication year: 2016
Online ISBN: 9781316576533
https://doi.org/10.1017/CBO9781316576533

[2] Efron, B. (2021). *Empirical Bayes: Concepts and Methods*. Retrieved from `https://efron.ckirby.su.domains/papers/2021EB-concepts-methods.pdf`

[3] Vindyani Herath(Teachng Fellow)
Advised me about references.