# MA677 Final Project

Yuhan Pu

2024-05-07

## Chapter I choose: Chapter 6 from CASI

## Main Points of Chapter 6: Empirical Bayes Method

I summarize the chapter/l Empirical Bayes (EB) Methods as below: Chapter 6 talks about Empirical Bayes methods, which are a refinement of Bayesian statistics. Unlike traditional Bayesian approaches, EB methods do not require a predefined prior; instead, they estimate the prior distribution from the data. This approach is particularly beneficial in situations where prior information is ambiguous or unavailable, making it a versatile tool in modern statistics.

**Robbins' Formula:** A significant portion of the chapter is dedicated to Robbins' formula, a cornerstone of EB methodology. This formula enables the estimation of prior distributions using the data itself, through the marginal maximum likelihood. An interesting question arises from this discussion: How do the assumptions made in Robbins' formula affect the robustness of EB estimations in practical scenarios?

**Applications in Medical and Biological Examples:** The text provides practical applications of EB methods in medical contexts, showcasing their utility in estimating disease probabilities. These examples not only demonstrate the real-world relevance of EB methods but also pose questions about their ethical implications in medical decision-making. For instance, how does the uncertainty in EB estimates impact clinical outcomes, and what measures can be taken to minimize potential risks?

**Utilization of Indirect Evidence:** A noteworthy aspect discussed is the use of indirect evidence in statistical inference, which underscores the capability of EB methods to enhance the accuracy of statistical estimates. This raises a fundamental question about the extent to which indirect evidence should be relied upon, especially when direct data are sparse or noisy. What are the potential biases introduced by heavily depending on indirect evidence, and how can these be mitigated?

**Conclusion:** Overall, Chapter 6 offers a compelling exploration of Empirical Bayes methods, emphasizing their adaptability and power in statistical analysis. It encourages readers to consider not only the mathematical and computational advantages of EB but also the practical implications and challenges posed by its application in various fields. As EB methods continue to evolve, further research is necessary to address the unresolved questions regarding their implementation and impact in more complex datasets and diverse applications. By engaging with these topics, the chapter not only educates on the mechanics of EB methods but also stimulates critical thinking about their broader implications and potential developments.

## Computational Methods in Empirical Bayes

I will include some packages from R and python and how they can help with the computing in EB. Computational methods play a crucial role in the application and understanding of Empirical Bayes techniques. The complexity of deriving priors from observed data necessitates the use of sophisticated statistical software and programming languages. Both R and Python offer robust libraries and packages designed to facilitate the implementation of EB methods, which are essential for statistical analysis in various fields, including bioinformatics, economics, and social sciences.

**R Packages that can be applied for Empirical Bayes:** In R, the "empiricalbayes" package stands out as a specialized tool for EB analysis. This package includes functions that automate the process of estimating priors and calculating posterior distributions based on empirical data. Another noteworthy package is "EBayes", which provides utilities for performing large-scale EB estimations, especially useful in genomic studies where large datasets are common. A practical example in R would involve using the "empiricalbayes" package to estimate the prior distribution from sample data and then apply this prior to perform Bayesian updates. This can be demonstrated by a script that fits an EB model to simulated data, illustrating how the estimated priors influence the posterior outcomes.

**Example of manually applying EB in R:**

```r
# Simulate some data
set.seed(123)
n <- 100
true_rates <- rgamma(n, shape = 2, rate = 1)
counts <- rpois(n, lambda = true_rates)  # Observed counts

# Assume a Gamma prior
# Using method of moments to estimate parameters
mean_counts <- mean(counts)
var_counts <- var(counts)

# Estimate Gamma parameters
alpha_hat <- mean_counts^2 / (var_counts - mean_counts)
beta_hat <- (var_counts - mean_counts) / mean_counts

# Calculate posterior parameters for each count
posterior_shape <- alpha_hat + counts
posterior_rate <- beta_hat + 1

# Compute the posterior means
posterior_means <- posterior_shape / posterior_rate

traditional_means <- counts  # Traditional estimate is the observed count itself
comparison <- data.frame(traditional = traditional_means, empirical_bayes = posterior_means)

head(comparison)
```

```
##   traditional empirical_bayes
## 1           1        2.217101
## 2           2        2.824768
## 3           0        1.609433
## 4           2        2.824768
## 5           3        3.432435
## 6           2        2.824768
```

**Python packages that can be applied for Empirical Bayes:** Python's "PyMC3" library offers a more general approach but is powerful in handling Bayesian statistical modeling, including EB methods. "PyMC3" allows for the construction of complex probabilistic models and performs efficient Bayesian inference with advanced sampling algorithms like Markov Chain Monte Carlo (MCMC). An example usage could be setting up a basic EB model in "PyMC3" to analyze a medical dataset, estimating disease prevalence as a prior, and updating this with new patient data to refine the estimates.[1]

##Mathematics Underlying Empirical Bayes Method

The below are some mathematical contents I found:

At the core of EB methods is Bayesian updating, where prior beliefs are updated with new evidence to form posterior beliefs.In EB, the prior is not specified beforehand but is estimated from the data, changing the formula to include an empirical estimate of the prior.

Bayesian Updating: Posterior(theta|x) = Likelihood(x|theta) * Prior(theta) / Evidence(x)

**An example of applying Bayesian Updating** Suppose you are testing the effectiveness of a new drug. You believe the probability of success of the drug has a Beta prior distribution. You then observe a number of trials and successes, and you want to update your beliefs about theta.

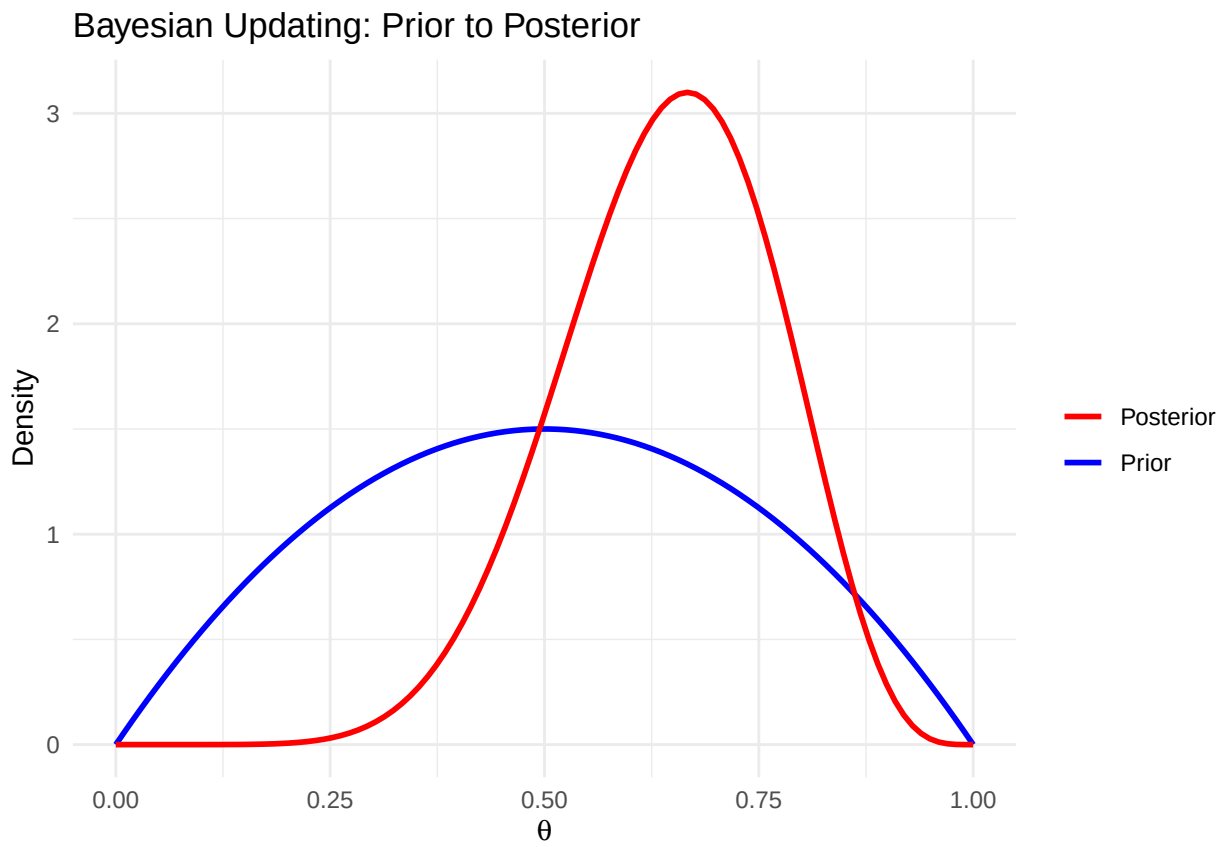
Assumptions: prior: Beta(a,b) likelihood: Binomial(n,x)

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
# Prior assumptions for Beta distribution
alpha_prior <- 2
beta_prior <- 2

# Observed data: n trials and x successes
n <- 10
x <- 7

# Bayesian updating using the conjugate prior
alpha_posterior <- alpha_prior + x
beta_posterior <- beta_prior + n - x

#data for plotting
theta <- seq(0, 1, length.out = 100)
prior <- dbeta(theta, alpha_prior, beta_prior)
posterior <- dbeta(theta, alpha_posterior, beta_posterior)

# Create a data frame for ggplot
df <- data.frame(theta, prior, posterior)
ggplot(df, aes(x = theta)) +
  geom_line(aes(y = prior, color = "Prior"), size = 1) +
  geom_line(aes(y = posterior, color = "Posterior"), size = 1) +
  labs(title = "Bayesian Updating: Prior to Posterior",
       x = expression(theta),
       y = "Density") +
  scale_color_manual(name = "", values = c("Prior" = "blue", "Posterior" = "red")) +
  theme_minimal()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```




The denominator in the Bayesian update formula, often called the evidence or marginal likelihood, plays a crucial role in EB. It integrates the likelihood across all possible values of the parameter, weighted by the prior. This integral normalizes the probability and ensures that the posterior distribution is properly scaled.

Marginal Likelihood(evidence): p(x) = integral of p(x|theta)p(theta) d(theta)

**An example of applyibng Marginal Liklihood** Assumption: prior: Beta(a,b) likelihood: Binomial(n,x)

```
# Prior parameters for Beta distribution
alpha_prior <- 2
beta_prior <- 2

# Observed data: n trials and x successes
n <- 10
x <- 7

# Function to calculate the binomial coefficient
binomial_coefficient <- function(n, k) {
  choose(n, k)
}

# Calculate the marginal likelihood using the Beta-binomial integration
marginal_likelihood <- binomial_coefficient(n, x) *
  beta(x + alpha_prior, n - x + beta_prior) / beta(alpha_prior, beta_prior)
print(marginal_likelihood)
```

```
## [1] 0.1118881
```

The reference for the mathematics is "The Theory of Empirical Bayes" by Robbins and "Bayesian Data Analysis" by Gelman et al. provide comprehensive insights into the theory and application of EB methods.

**Some shortages and challenges that EB is facing:**

Estimations stability: The estimation of priors can be sensitive to outliers or model misspecification. Robust statistical techniques or Bayesian hierarchical models can help mitigate these issues.

Dealing with high dimensional data: In cases of high-dimensional data, the curse of dimensionality can make prior estimation difficult. Dimensionality reduction techniques or assuming sparsity (for example, using LASSO for regularization) can be beneficial.

## Historical Context of Empirical Bayes Methods

**It starts** Empirical Bayes methods are a modern extension of the classical Bayesian framework, which has its roots in the 18th century with Reverend Thomas Bayes. The formal mathematical framework that Bayes developed was posthumously published in 1763 in "An Essay towards solving a Problem in the Doctrine of Chances." However, the EB methods, as discussed in Chapter 6, primarily derive from 20th-century developments.

**Development:** The conceptual foundation for Empirical Bayes methods was laid by Herbert Robbins in the 1950s. Robbins introduced the idea that prior distributions could be estimated from the data itself, rather than being strictly assumed as known a priori. His seminal paper, "An Empirical Bayes Approach to Statistics" (1955), presented a framework where the parameters of the prior distribution are estimated from the data, which was a significant shift from traditional Bayesian methods that required subjective or informative priors.

**When stepping into the computer era:** The advent of the computer age in the late 20th century significantly influenced the development and application of EB methods. With the rise of computational power, it became feasible to implement complex EB models that required intensive calculations for estimating priors and integrating over vast parameter spaces. This computational capability expanded the applicability of EB methods across various fields, including genomics, epidemiology, and machine learning, where large datasets are common.

**Modern:** The integration of EB methods into statistical practice has been influenced by the work of other statisticians like James Stein, who, in the 1960s, introduced ideas such as shrinkage estimators that closely relate to EB principles. These methods, which optimize the trade-off between bias and variance, share a conceptual link with EB in that they adjust estimates based on the data's behavior.

**For future:** Over the decades, the scope of EB methods has broadened significantly. Researchers like Bradley Efron have further developed the theory and application of EB, particularly in the context of large-scale hypothesis testing and non-parametric approaches. Efron's work, including techniques like bootstrap methods introduced in the late 20th century, complements EB methods by providing robust ways to assess uncertainty in the estimates of priors and other parameters.

References are listed at the end of the notes.[2]

## Statistical Practice Implications of Empirical Bayes Methods

Chapter 6's discussion on Empirical Bayes (EB) methods provides several key implications for statistical practice. These methods, which refine Bayesian analysis by estimating prior distributions from data, enhance the flexibility and applicability of Bayesian approaches across various fields. The implications of integrating EB into statistical workflows are profound, ranging from improved estimator accuracy to practical challenges in implementation.

**Improve estimator accuracy:** One of the primary advantages of EB methods is their ability to increase the accuracy of estimators, particularly in complex models where traditional Bayesian methods may be

computationally intensive or where the prior is difficult to specify. EB methods use the data itself to inform the prior, leading to more data-driven and potentially more accurate inferences. This can be particularly advantageous in fields like genomics or economics, where prior information may not be readily available or is too vague to be useful.

**Handling large datasets:** With the advent of big data, EB methods have become increasingly important. They offer a computationally feasible way to handle large datasets by simplifying the estimation process. Unlike full Bayesian methods that may require complex integration over many parameters, EB methods can provide computationally efficient approximations that are crucial in processing and analyzing large volumes of data.

**Implications for Predictive Modeling:** EB methods can significantly enhance predictive modeling by allowing the model to adjust to new data more flexibly. As more data becomes available, EB methods can update the prior distributions in a way that reflects the new information, leading to improved predictive accuracy. This dynamic updating is particularly useful in real-time analytics and applications where data inflow is continuous and voluminous.

**Challenges in prior estimations:** EB methods can significantly enhance predictive modeling by allowing the model to adjust to new data more flexibly. As more data becomes available, EB methods can update the prior distributions in a way that reflects the new information, leading to improved predictive accuracy. This dynamic updating is particularly useful in real-time analytics and applications where data inflow is continuous and voluminous.

**Reality practice:** The use of EB methods raises certain ethical and practical considerations. For example, in medical statistics, where EB methods might be used to make clinical decisions, the uncertainty in prior estimation needs to be carefully considered and communicated to avoid misinterpretations and potential harm. Transparency in how priors are estimated and how conclusions are drawn is crucial to maintaining trust and reliability in statistical analysis.

**For learning and teaching:** The implications of EB methods for statistical practice also extend to education and training. As these methods become more integrated into standard statistical toolkits, there is a growing need for statisticians and data scientists to be trained not only in the technical aspects of EB methods but also in their theoretical foundations and practical applications. This requires updated curricula that balance traditional statistical teachings with modern computational techniques.[3]

**An example of applying Empirical Bayes Methods in reality**

Scenario: Estimating Treatment Success Rates in Hospitals

Suppose we have data from several hospitals on the success rates of a specific treatment. Each hospital has its own observed success rate, but these rates are based on varying numbers of patients treated, leading to potentially unreliable estimates for hospitals with fewer patients.

Data I assume we have: 'hospital_id': identification for each hospital 'patient_treated': number of patients accepting treatment 'success_treated': number of patients successfully recovering

```r
#Let's assum the dataset like this
hospitals <- data.frame(
  hospital_id = 1:10,
  treated_patients = c(150, 180, 100, 50, 300, 230, 90, 120, 200, 160),
  successful_treatments = c(105, 126, 70, 35, 210, 161, 63, 84, 140, 112)
)

#The success rate
hospitals$raw_success_rate <- hospitals$successful_treatments / hospitals$treated_patients

#Applying EB methods
alpha_prior <- 15
beta_prior <- 25
```

```r
# Calculate the posterior estimates
hospitals$posterior_rate <- (hospitals$successful_treatments + alpha_prior) /
                            (hospitals$treated_patients + alpha_prior + beta_prior)

# Compare raw and posterior estimates
print(hospitals[, c("hospital_id", "raw_success_rate", "posterior_rate")])
```

```
##    hospital_id raw_success_rate posterior_rate
## 1            1              0.7      0.6315789
## 2            2              0.7      0.6409091
## 3            3              0.7      0.6071429
## 4            4              0.7      0.5555556
## 5            5              0.7      0.6617647
## 6            6              0.7      0.6518519
## 7            7              0.7      0.6000000
## 8            8              0.7      0.6187500
## 9            9              0.7      0.6458333
## 10          10              0.7      0.6350000
```

```r
# Visualization using ggplot2
library(ggplot2)
ggplot(hospitals, aes(x = as.factor(hospital_id))) +
  geom_bar(aes(y = raw_success_rate, fill = "Raw"), stat = "identity", position = "dodge") +
  geom_bar(aes(y = posterior_rate, fill = "Posterior"), stat = "identity", position = "dodge") +
  labs(title = "Comparison of Raw and Posterior Success Rates per Hospital",
       x = "Hospital ID",
       y = "Success Rate") +
  scale_fill_manual("Estimate Type", values = c("Raw" = "blue", "Posterior" = "red")) +
  theme_minimal()
```


[4]

## References

[1] By Searching through googling and also some resources talking about packages in programming languages for helping with EB

[2] Robbins, H. (1955). "An Empirical Bayes Approach to Statistics." Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife." Bayes, T. (1763). "An Essay towards solving a Problem in the Doctrine of Chances."

[3] This part I summarized the contents that talks about the practice of the EB and fed them into GPT for rewriting in a more academic tone.

[4] For the practice in reality part, I asked GPT for help with the code for ggplot code (parameter adjustment)