5-2024

Bayesian Estimation of Reproduction Numbers from Distributions of Outbreaks Sizes: Branching Process Approach

Alberta Araba Johnson
The University of Texas Rio Grande Valley

BAYESIAN ESTIMATION OF REPRODUCTION NUMBERS FROM DISTRIBUTIONS OF OUTBREAK SIZES: BRANCHING PROCESSES APPROACH

A Thesis

by

ALBERTA ARABA JOHNSON

Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: Applied Statistics and Data Science

The University of Texas Rio Grande Valley

May 2024

BAYESIAN ESTIMATION OF REPRODUCTION NUMBERS FROM DISTRIBUTIONS OF OUTBREAK SIZES: BRANCHING PROCESSES APPROACH

A Thesis
by
ALBERTA ARABA JOHNSON

COMMITTEE MEMBERS

Dr. George P. Yanev
Chair of Committee

Dr. Tamer F. Oraby
Committee Member

Dr. Santanu Chakraborty
Committee Member

Dr. Zhuanzhuan Ma
Committee Member

May 2024

ABSTRACT

Johnson, Alberta A., Bayesian Estimation of Reproduction Numbers from Distributions of Outbreak Sizes: Branching Processes Approach. Master of Science (MS), May, 2024, 55 pp., 4 tables, 7 figures, references, 25 titles.

The Generalized Poisson distribution is useful in modeling epidemiological processes as a branching stochastic processes problem. Our goal is to construct accurate and reliable estimators for the reproduction number (R_0) (i.e., the number of secondary infections), particularly in the context of disease outbreaks modeled by a Galton-Watson process. Towards this goal, we construct the classical Bayes estimator, the Maximum Likelihood estimator, and the Empirical Bayes (EB) estimator under the Square Error Loss function in Chapter II. We prove that the Empirical Bayes estimator is asymptotically optimal and estimate the rate of convergence. We then proceed to monotonize the Empirical Bayes estimator in Chapter III using the Van Houwelingen method (Van Houwelingen 1977) and the Isotonic Regression method (Barlow, Brunk, and Bremner 1972), then introduced the concept of the risk and regret risk associated with our estimators. For the numerical study in Chapter IV we assume a Poisson distribution for the reproduction number and that the initial number of infected individuals follows a Poisson distribution. Simulation results indicate that the empirical estimator suffers from "jumpiness", hence the need for monotonization. We then compare the regret risks of each of the estimators and find out that the monotonized estimate outperforms the others.

DEDICATION

To my dear parents, Anna Yankey and the Late Mr. Albert Johnson whose unconditional love, support, motivation, and belief in my abilities have been the guiding lights of my journey. Their sacrifices have not gone unnoticed, and this achievement is as much theirs as it is mine.

ACKNOWLEDGMENTS

I extend my deepest gratitude to all those who have made this journey not just possible but also rewarding.

First of all, I would like to thank my mentor and committee chair Dr. George P. Yanev for his invaluable guidance, patience, and insightful criticism throughout this research. Your insights and feedback were instrumental in shaping this work.

I am also immensely grateful to my committee members, Dr. Tamer F. Oraby, Dr. Santanu Chakraborty, and Dr. Zhuanzhuan Ma for serving on the committee and for whose expertise and constructive criticism have significantly contributed to the depth and quality of this study.

I want to acknowledge the School of Mathematical and Statistical Sciences and the College of Science of UTRGV, thank you for giving me the opportunity and funding to pursue a Master's and this research through the DGA award. Your investment in my education is deeply appreciated.

To my family, for their unconditional love, understanding, and sacrifices. Thank you for standing by me, for being my source of strength and inspiration, and for believing in me when I doubted myself.

And finally, to all who have contributed directly or indirectly to this work, your roles have been indispensable.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

1.1 Stochastic Modeling of Epidemic Diseases

Epidemic disease modeling serves as a fundamental building block in global health for understanding and managing infectious outbreaks. Many of the infections are harmless and even beneficial (for example, the bacteria we carry in our intestines which assist in the digestion of food). However, some like the pathogenic infectious agents harm their hosts and cause disease. The distinction between benign and harmful infections is crucial in epidemic modeling, as it informs disease management and control strategies. Pathogenic agents, such as viruses, bacteria, and parasites, often lead to epidemics when they invade a susceptible population and spread rapidly, overwhelming the host's defenses and public health systems. The transmission between human or animal hosts occurs in a variety of ways including, by direct contact (scabies, leprosy), the respiratory route (whooping cough, influenza, tuberculosis), through sexual contact (HIV, gonorrhea), etc (Vynnycky, White, and Fine 2010). In recent times, we have witnessed a resurgence of diseases previously under control, as exemplified by the recent surge in measles cases in the UK. This resurgence highlights the importance of constant vigilance in infectious disease control and the impact of vaccination programs on public health. This recent crisis has prompted authorities to declare a national health incident. As reported by Geneva Abdul in January 2024, the UK Health Security Agency (UKHSA) has warned of further outbreaks across Britain due to the decrease in the uptake of the measles, mumps, and rubella (MMR) vaccine. The (MMR) vaccine coverage has significantly dropped, with the average uptake falling to around 85%, significantly lower than the desired threshold of 95% for herd immunity. This decrease has led to numerous measles cases,

particularly in regions like the West Midlands and London, where vaccination rates are alarmingly low. Data released by the agency showed that, since last October, there were 216 lab-confirmed cases in the West Midlands, with 103 cases likely. About 80% of the cases were in Birmingham and 10% were in Coventry, according to the agency, citing low vaccination rates. Most of the cases were among children aged under 10. The UKHSA emphasizes the need for parents to ensure their children are vaccinated, highlighting the critical role of vaccination in preventing widespread outbreaks.

To better understand the spread and control of diseases the concept of the effective reproduction number (R_o) is vital. This parameter, which represents the average number of secondary infections produced by a single infectious case in a population, is crucial in epidemiological modeling. For diseases like measles, maintaining (R_o) below unity is essential for achieving eliminations (De Serres, Nigel J. Gay, and C. Paddy Farrington 2000). However, as the UK's situation illustrates, this is a dynamic threshold, heavily dependent on vaccination coverage and public health intervention. It has been observed that large outbreaks become increasingly likely as the reproductive number approaches one, a situation termed as being at 'criticality'. This reproductive number is influenced by the fraction of the population that is not immunized. In scenarios where vaccine uptake declines, the population witnesses larger and more frequent outbreaks, potentially leading to the re-establishment of measles as an endemic disease (Jansen et al. 2003). Another significant threat in the landscape of epidemic diseases is avian influenza, particularly the H5N1 strain. This strain, while primarily affecting birds, has shown the capability to infect humans. The public health risk associated with avian H5N1 influenza is a subject of extensive study. The reproduction number R_o of human infections with avian H5N1 virus is assumed to be below unity in the absence of viral reassortment. This means that while individual cases may lead to small human-to-human transmission clusters, the transmission rate is not high enough to sustain an outbreak. Understanding the dynamics of these 'subcritical' outbreaks is crucial for predicting and managing potential public health risks associated with avian influenza. It involves modeling the outbreak size distributions and using maximum-likelihood methods to estimate R_o. This kind of modeling helps in identifying any

significant changes in the transmission dynamics of the virus, which could indicate an increase in its ability to spread among humans (Ferguson et al. 2004a; Ferguson et al. 2004b).

In addressing epidemic diseases, mathematical models play a critical role. They provide frameworks for analyzing surveillance data, especially concerning diseases post-elimination of sustained endemic transmission. Branching process models have been utilized for the surveillance of infectious diseases controlled by mass vaccination programs (see (Christine 2010)). These models help in understanding the threshold behavior of epidemics and in calculating the critical vaccination threshold. They are particularly relevant in the context of estimating the effective reproduction number, which is a key indicator of whether an infectious disease will continue to spread or die out in a population. The effective reproduction number being below one is indicative of the disease not persisting but presenting itself in varying outbreak sizes, triggered by external factors such as importations (C. P. Farrington, Kanaan, and N. J. Gay 2003). To better describe transmission in a small population, we need to develop a stochastic model that incorporates the effects of chance on the possible outcome. There are several kinds of stochastic models which include discrete-time compartmental models. It keeps track of the total number of susceptible and infectious persons at each time step. Random numbers are used to determine the total number of susceptible infected by the infectious persons in each generation, assuming that this number follows some distribution (Vynnycky, White, and Fine 2010).

In the event of a potential pandemic, understanding the dynamics of the disease spread is critical. This is where the offspring mean θ comes into play. It represents the R_o in a disease outbreak modeled by a Galton-Watson process. The primary goal in such a scenario is to construct a reliable and accurate estimator, for θ denoted as $\hat{\theta}$. The significance of $\hat{\theta}$ lies in its ability to guide public health responses. When $\hat{\theta}$ is close to 0, it indicates that public intervention may not be necessary as the outbreak will eventually go to extinction. This scenario leads to a sigh of relief among public health officials and the public as there is a decline in the outbreak and things start to move to normal. However, the situation when $\hat{\theta}$ is close to 1, leads to a sustained level of disease

transmission. Here health measures must be taken to curb the spread of a disease. For this reason, we shall construct a series of quality estimates to make reliable inferences about the population.

1.2 Generalized Poisson Distribution

Out of all the power series distributions, the Poisson distribution is uniquely characterized as having equal mean and variance (P. Consul C. 1989); however, in populations that are supposed to be Poisson, researchers have observed that this is not always the case. In addressing these particular issues, Consul and Jain, in 1970 (P.C. Consul and Jain 1970), introduced a Generalized Poisson distribution (GPD). This distribution extends the classic Poisson model, accommodating a greater variability.

The GPD has parameters $0 \leqslant \theta < 1$ and $\tau > 0$, and probability mass function (pmf)

$$P(x;\theta,\tau) = \frac{\tau}{x!}(\tau+\theta x)^{x-1}e^{-(\tau+\theta x)}, \quad x = 0,1,\ldots \tag{1.1}$$

The distribution has mean $\frac{\tau}{1-\theta}$ and variance $\frac{\tau}{(1-\theta)^3}$. Note that the GPD is a member of the family of Abel series distributions (see (Charalambides 1990)) and it reduces to Poison distribution for $\theta = 0$.

Since its introduction the GPD has been a versatile tool in many fields. In epidemiology, the GPD is instrumental in modeling the spread of diseases, accounting for the variable infection rates where θ is particularly significant as it quantifies the average number of secondary infections generated by one case, reflecting the potential of the disease transmission. Meanwhile, τ represents the scale of the initial outbreak which we need to understand the early stages of the disease spread. This is often observed in real-world scenarios, focusing on statistical modeling of epidemic diseases through branching processes and Bayesian inference (Yanev 2001; Albertsen, Steffensen, and Kirstensen 1992). The GPD has also been essential in understanding and predicting the spread of cyber threats like viruses and worms (Sellke, Shroff, and Bagchi 2008), by modeling the variable rates of virus spread. The GPD aids in developing more effective security measures to protect against these digital threats. Additionally, the traffic flow analysis also benefits from the application of the GPD. Understanding and managing traffic congestion and flow patterns is crucial in urban

planning and road safety (Koorey 2007). The distribution's capacity to model variable traffic density and flow rate helps in designing better traffic management systems and infrastructure. (for other applications, see (Gipps 1976; Nirei, Stamatiou, and Sushko 2012; Iešmantas and Alzbutas 2014; Aldous 1999)). The motivation for this paper stems from the role of GPD distribution in modeling epidemics.

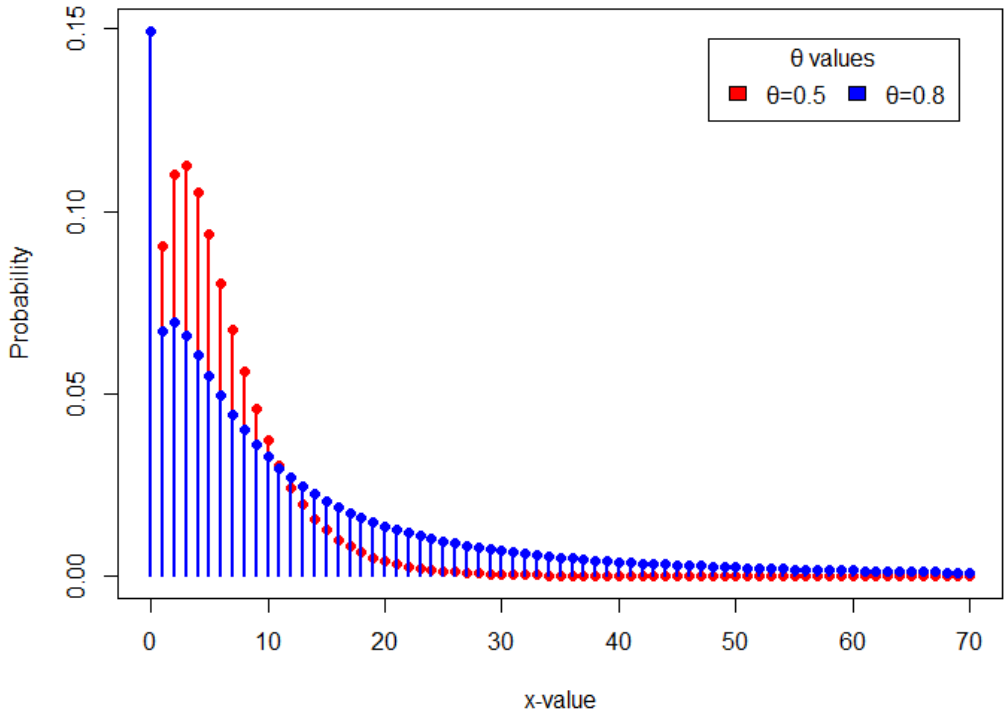

Figure 1.1: Generalized Poisson pmf with $\tau = 3$.

1.3 Total Progeny of Branching Processes

In the 19th century, Victorian England's aristocratic families posed a question to mathematician Sir Francis Galton:

> *How many male children (on average) must each generation of a family have in order for the family name to continue in perpetuity?* (Albertsen, Steffensen, and Kirstensen 1992)

The answer to this question became the oldest, and simplest branching process known as the Galton–Watson (GW) process. Also, it is known as the Bienayme–Galton–Watson process dating as far back as 1845 to the work of statistician Bienayme. By definition, a branching process is a system in which individuals (or entities) live for a random time, producing a random number of progenies (offspring). These processes are applicable in many areas such as gene propagation, neutron chain reactions in nuclear fusion, cell biology, and epidemiology (Yanev 2001). In this paper, we apply this concept to epidemiology by emphasizing how the progeny, or total number of infected individuals of a communicable disease, can be modeled as a variable of a GPD.

The Galton-Watson branching process (GWP) is defined by the recurrence formula:

$$Z_{n+1} = \sum_{i=1}^{Z_n} \xi_{i,n}, \qquad n = 0,1,2,\ldots, \tag{1.2}$$

where $\xi_{i,n}$, $n = 0,1,2,\ldots$ are independent and identically distributed (iid) non-negative integer random variables (rv). The process follows two fundamental assumptions

(i) The number of offspring $\xi_{i,n}$ produced by a single parent particle is independent of the history of the process, and of other individuals existing at the present.

(ii) The offspring distribution is consistent across all individuals in all process generations.

The total progeny distribution in a GWP is a member of the family of Lagrange Distributions with pmf (see Pakes paper).

$$l(x;f,g) = \sum_{r=0}^{x} \frac{r}{x} f^{x*}(x-r)g(r), \quad x = 1,2,\ldots \tag{1.3}$$

where f and g are discrete probability distribution. Setting in 1.3

$$f(x) = \frac{\theta^x}{x!}e^{-\theta} \quad \text{and} \quad g(r) = \frac{\tau^r}{r!}e^{-\tau}, \quad r,x = 0,1,\ldots \tag{1.4}$$

We obtain the GPD's pmf 1.1.

Consider $Z_0, Z_1, \ldots, Z_n$ to be the sizes of the first n generations in a GWP and let $X_n := Z_0 + Z_1 + \ldots + Z_n$. Assume $X = lim_{n\to\infty} X_n$ is the total progeny of the process if the initial number of individuals is r then the distribution of X is known as Borel-Tanner distribution given by $P(X = x \,|\, Z_0 = r) = \frac{rx^{x-r-1}}{(x-r)!}\theta^{x-r}e^{-\theta x}$. Since $0 \leqslant r \leqslant x$ we obtain

$$
\begin{aligned}
P(X = x) &= \sum_{r=0}^{x} P(X = x \mid Z_0 = r) P(Z_0 = r) \qquad (1.5)\\
&= \sum_{r=0}^{x} \frac{rx^{x-r-1}}{(x-r)!}\theta^{x-r}e^{-\theta x}\frac{\tau^r e^{-\tau}}{r!}\\
&= \frac{e^{-(\tau+\theta x)}}{x!}\sum_{r=1}^{x}\frac{(x-1)!}{(x-r)!(r-1)!}(\theta x)^{x-r}\tau^r \quad (set\ k = r-1)\\
&= \frac{\tau e^{-(\tau+\theta x}}{x!}\sum_{k=0}^{x-1}\frac{(x-1)!}{(x-1-k)!k!}(\theta x)^{x-1-k}\tau^k\\
&= \frac{\tau(\tau+\theta x)^{x-1}}{x!}e^{-(\tau+\theta x)}.
\end{aligned}
$$

Hence the total progeny of this GWP follows the Generalized Poisson distribution. More importantly, the parameters θ and τ are crucial. With θ representing the reproduction number or number of secondary infections caused by a parent (infected individual) and τ indicating the initial number of infections. The rest of the thesis is organized as follows: In Chapter II we discuss the Bayes estimators for θ when Z_0 has an arbitrary discrete distribution. In Chapter III we monotonize the empirical Bayes estimator discussed in Chapter II Chapter IV presents a numerical study when the outbreak size follows the GPD.

CHAPTER II

BAYES ESTIMATORS FOR θ WHEN Z_0 HAS ARBITRARY DISCRETE DISTRIBUTION

Bayesian statistical methods play an important role in estimating the parameters, especially in the context of the Generalized Poisson Distribution (GPD). The core of this method lies in formulating a prior distribution $G(\theta)$ which represents the initial belief or knowledge about the parameter θ. This prior distribution captures the variability of θ. After experimenting, we observe data x which is indicative of θ and taken from the population. This form the sample distribution $p(x \mid \theta)$. This distribution illustrates our belief in the likelihood of observing x given θ. Using the experimental data, we then update the prior and create a posterior distribution $G(\theta \mid x)$. This is then derived using the Bayes Rule:

$$G(\theta \mid x) = \frac{p(x \mid \theta)G(\theta)}{m(x)} \qquad \theta \in \Omega, \tag{2.1}$$

In the equation, $m(x)$ denotes the marginal distribution of X that is, $m(x) = \int_{\Omega} p(x,\theta)\mathrm{d}\theta$ and $p(x,\theta)$ is the joint probability mass function which is integral in understanding the joint probability mass function. This posterior distribution is then used to make further inferences about θ.

2.1 Loss Functions–Square Error Loss

Within the Bayesian framework, accurately estimating the unknown parameter θ, represented as a random variable (r.v.) with posterior distribution G is essential. The parameter value drawn from $G(\theta \mid x)$, the posterior distribution, serves as a possible realization of the true parameter. It is therefore important to consider how accurate and precise the estimation is by computing the expected loss of the given estimate. To do this, we use a loss function.

A loss function $L(\theta, \hat{\theta})$, is defined as the difference between a parameter's estimated and true value. This function represents the "cost" or "loss" associated with some random event. In contrast to the frequentist theory, errors are minimized but usually do not consider the loss associated with the error. And so there is a level of ignorance in one's sureness of the parameter. Bayesian estimation aims to minimize posterior loss, and so if one is to be unsure or wrong in their estimation, then it is best to be on the side of *least* wrong. In this paper, we introduce the following Square Error Loss function defined as follows:

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2 \tag{2.2}$$

The Square Error Loss function (2.2 is a symmetric loss function that equally penalizes overestimation and underestimation. The symmetry comes from the fact that the loss is squared, so then, it does not matter whether the predicted value $\hat{\theta}$ is above or below the true value θ; this loss is the same for an equal magnitude of error in either direction. It is crucial in Bayesian estimation for its ability to provide a clear and quantifiable measure of the estimation accuracy. A visual depiction of this symmetry is given below:

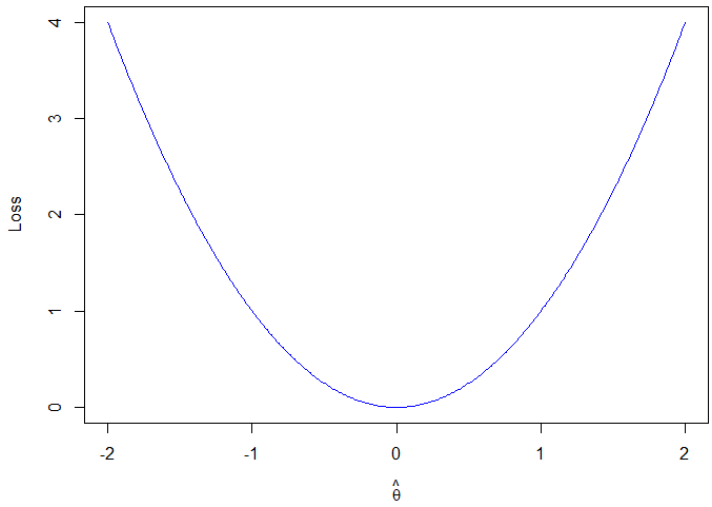

Figure 2.1: This graph illustrates the symmetric nature of the Square Error Loss function.

We chose the Square Error Loss function, particularly for its capacity to provide a clear measure of the estimation accuracy. This is particularly important in fields such as epidemiology, where precise and accurate estimates of the reproduction number are often used to advise public health officials on the possible severity of an outbreak.

2.2 Classical Bayes Estimators

A more detailed Bayes mathematical framework consists of the following elements (e.g. Stijnen (Stijnen 1980)). We observe a random variable or vector X, with distribution θ which is unknown. The problem is what decision to take concerning the true value of θ.

(i) **Sample Space** We define a sample space S of observations, complete with a $\sigma-$algebra on $\mathcal{S}$.

(ii) **Probability Measures** The collection of probability measures on the space $(S,\mathcal{S})$ denoted by $\mathcal{P}$ is usually parameterized by some set suitable parameters $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$.

(iii) **Action Space** The action space $\mathcal{A}$ represents a set A of possible actions that a statistician might take upon observing some $x \in S$. The set A is equipped with a $\sigma-$algebra on $\mathcal{A}$.

(iv) **Decision Rules** A collection D of decision rules. Decision rules in this context are defined as $\mathcal{S}$-$\mathcal{A}$ measurable maps from S into A. Upon observing $x \in S$, the statistician will take action $d(x) \in A$ based on the decision rule $d \in D$.

(v) **Loss Function** The loss function $L : \Omega \times A \longrightarrow \mathbb{R}$ is critical for measuring the cost of decisions. For each $\theta \in \Omega$, the function $L(\theta,\cdot)$ must be $\mathcal{A}$ measurable and bounded from below on A. The incurred loss when taking action $d(x) \in A$, if θ is the true parameter value is represented by $L(\theta, d(x))$.

(vi) **Prior Distribution** The prior distribution G, a probability measure on Ω equipped with the $\sigma-$ algebra $\mathcal{W}$ reflects the initial belief about the parameter space.

Adopting the Bayesian model, we will define the following Bayes estimator θ_G for θ. Suppose $\theta \in \Omega$ is a realization of a random variable (r.v) Θ. Under the squared error loss function and with a

prior distribution G it is well known that the Bayes estimator θ_G for θ is $\theta_G(x) = E[\Theta|X = x]$ that is the posterior expectation of Θ given $X = x$.

Proposition 1. *Consider the Galton-Watson process (1.2) with Poisson offspring f in (1.4). The Bayes estimator $\theta_G(x)$ for θ is given by*

$$\frac{\sum_{r=0}^{\infty} P(Z_0 = r)c_r(x)\left(\int_0^1 \theta^{x-r+1}e^{-\theta x}dG(\theta)\right)}{\sum_{r=0}^{\infty} P(Z_0 = r)c_r(x)\left(\int_0^1 \theta^{x-r}e^{-\theta x}dG(\theta)\right)} =: \frac{\psi_G(x)}{q_G(x)},$$

where for $r = 0, 1, \ldots$

$$c_r(x) := \frac{r}{x}\frac{x^{x-r}}{(x-r)!}, \quad x = r, r+1, \ldots. \tag{2.3}$$

Proof. We have

$$\begin{aligned}
\theta_G(x) &= E[\Theta|X = x] \\
&= \int_0^1 \theta P(\theta|X = x)\, d\, G(\theta) \\
&= \frac{1}{P(X = x)}\int_0^1 \theta P(\Theta = \theta, X = x)\, dG(\theta) \\
&= \frac{\sum_{r=0}^{\infty} P(Z_0 = r)\int_0^1 \theta P(X = x|Z_0 = r)d\, G(\theta)}{\sum_{r=0}^{\infty} P(Z_0 = r)\int_0^1 P(X = x|Z_0 = r)d\, G(\theta)} \\
&= \frac{\sum_{r=0}^{\infty} P(Z_0 = r)c_r(x)\left(\int_0^1 \theta^{x-r+1}e^{-\theta x}dG(\theta)\right)}{\sum_{r=0}^{\infty} P(Z_0 = r)c_r(x)\left(\int_0^1 \theta^{x-r}e^{-\theta x}dG(\theta)\right)}.
\end{aligned}$$

Remarks. Recall that if Z_0 follows $\mathrm{Poi}(\tau)$, then (1.5) is the Generalized Poisson distribution. In the

case of GPD, the Bayes estimator simplifies to

$$\theta_G(x) = \frac{\int_0^1 \theta(\tau+\theta x)^{x-1} e^{-(\tau+\theta x)}\, dG(\theta)}{\int_0^1 (\tau+\theta x)^{x-1} e^{-(\tau+\theta x)}\, dG(\theta)}.$$

Furthermore, if G is $Uni(a_1, a_2)$ then

$$\theta_G(x) = \frac{\int_{a_1}^{a_2} \theta(\tau+\theta x)^{x-1} e^{-(\tau+\theta x)}\, d\theta}{\int_{a_1}^{a_2} (\tau+\theta x)^{x-1} e^{-(\tau+\theta x)}\, d\theta}. \tag{2.4}$$

2.3 Empirical Bayes Estimators: Construction and Properties

To make accurate inferences about the population parameter, it is often important to specify a prior distribution for it. However, sometimes this prior distribution is assumed to exist but is unknown. The empirical Bayes approach addresses this by leveraging a series of comparable past experiments to inform about the prior distribution. This method is particularly applicable when an experiment is part of a sequence of similar investigations, where past data can shed light on the unknown prior distribution. Consider a series of n independent copies of the random triple (X, Z_0, Θ) denoted as (X_1, Z_{01}, Θ_1), (X_2, Z_{02}, Θ_2), $\ldots$ (X_n, Z_{0n}, Θ_n) where Θ has a (prior) distribution G.

Assuming τ is known, (X_i, Z_{0i}), $i = 1, 2, \ldots$ are observable, but Θ_i, $i = 1, 2, \ldots$ are not. The empirical Bayes method then raises the question: it is possible or not to infer the approximate form of the unknown G or directly of the Bayes estimator $\theta_G(x)$, from the set of values $(X_1, Z_{01}), (X_2, Z_{02})$ (Robbins 1964)? And the answer is yes.

In what follows, we will adopt the empirical Bayes method of estimation (Carlin 2000), which relies on the assumption of the existence of a prior, which however is unspecified except that it is also i.i.d. from an unknown distribution, with cumulative distribution function G. Our goal is to construct a point estimate for θ given the sequence of past data. Such an estimator is called empirical Bayes (EB) estimator. We will seek a direct (independent of G) estimate of the Bayes

estimator θ_G.

Following Robbins, we consider the case where $(X_1,Z_{01}),(X_2,Z_{02})\ldots,(X_n,Z_{0n})$ is a sequence of independent random vectors, independent from (X,Z_0,Θ) and with the same BT marginal distribution as $X|Z_0$. Consider past observed data $(x,z_0)(n) := \{(x_1,z_{01}),(x_2,z_{02}),\ldots,(x_n,z_{0n})\}$ generated by an unobserved set of parameter values $\{\theta_1,\theta_2,\ldots,\theta_n\}$ according to the GPD p.m.f. $p(x;\theta,\tau)$ given in (P.C. Consul and Jain 1970).

Let x be the present observation and θ be the present parameter value of Θ. An EB estimator $\theta_n((x,z_0)(n);x) =: \theta_n(x)$ for the parameter θ is a function of the currently observed x and the past data $(x,z_0)(n)$. Define

$$\psi_{nj}(x) = \frac{c_{Z_j}(x)c_1(X_j-x)}{c_{Z_j}(X_j)} I\{Z_j \leqslant x < X_j\}, \quad j = 1,2,\ldots,n$$

and

$$q_{ni}(x) = \frac{c_{Z_i}(x)}{c_{Z_i}(X_i)} I\{Z_i \leqslant x = X_i\}, \quad i = 1,2,\ldots,n.$$

Now, consider

$$\psi_n(x) := \left(\frac{1}{n}\sum_{j=1}^{n}\psi_{nj}(x)\right) \quad \text{and} \quad q_n(x) := \frac{1}{n}\sum_{j=1}^{n} q_{nj}(x). \tag{2.5}$$

In the next lemma, we show that statistics (2.5) are unbiased and consistent estimators for the numerator and denominator of $\theta_G(x)$, respectively.

Lemma 2. *Let $E_n[\cdot]$ and $Var_n[\cdot]$ denote the expectation and variance with respect to $(X_1,Z_1),(X_2,Z_2),\ldots,(X_n,Z_n)$. Then*

$$(i) \quad E_n\left[\frac{1}{n}\sum_{j=1}^{n}\psi_{nj}(x)\right] = \psi_G(x) \quad \textit{and} \quad E_n[q_n(x)] = q_G(x).$$

$$(ii) \quad Var_n[\psi_n(x)] \leqslant \frac{\psi_G(x)}{n} \quad \textit{and} \quad Var_n[q_n(x)] \leqslant \frac{q_G(x)}{n}.$$

Proof. (i) By the Law of Total Expectation, we have

$$\begin{aligned} E_n[\psi_{nj}(x)] &= \sum_{r=0}^{x} E_n[\psi_{nj}(x)|Z_j = r]P(Z_j = r) \\ &= \sum_{r=0}^{x} P(Z_j = r)c_r(x)E_n\left[\frac{c_1(X_j - x)}{c_r(X_j)} I\{X_j \geqslant x+1\}\big|Z_j = r\right] \\ &= \sum_{r=0}^{x}\sum_{t=x+1}^{\infty} P(Z_j = r)c_r(x)\frac{c_1(t-x)}{c_r(t)}\int_0^1 c_r(t)\theta^{t-r}e^{-\theta t}\, dG(\theta). \end{aligned}$$

Setting $y = t - x$, we obtain

$$\begin{aligned} E_n[\psi_{nj}(x)] &= \sum_{r=0}^{x}\sum_{y=1}^{\infty} P(Z_j = r)c_r(x)\int_0^1 c_1(y)\theta^{y+x-r}e^{-\theta(y+x)}\, dG(\theta) \\ &= \sum_{r=0}^{x} P(Z_j = r)c_r(x)\int_0^1 \theta^{x-r+1}e^{-\theta x}\left(\sum_{y=1}^{\infty} c_1(y)\theta^{y-1}e^{-\theta y}\right) dG(\theta) \\ &= \sum_{r=0}^{x} P(Z = r)c_r(x)\int_0^1 \theta^{x-r+1}e^{-\theta x}\, dG(\theta) \\ &= \psi_G(x). \end{aligned}$$

Similarly, we obtain

$$\begin{aligned} E_n[q_{nj}(x)] &= \sum_{r=0}^{x} E_n[q_{nj}(x)|Z_j = r]P(Z_j = r) \\ &= \sum_{r=0}^{x} P(Z_j = r)E_n\left[\frac{c_r(x)}{c_r(X_j)} I\{X_j = x\}\big|Z_j = r\right] \\ &= \sum_{r=0}^{x} P(Z_j = r)\frac{c_r(x)}{c_r(t)}\int_0^1 c_r(t)\theta^{x-r}e^{-\theta x}\, dG(\theta) \\ &= q_G(x). \end{aligned}$$

(ii) We will find upper bounds for the variances of $q_{nj}(x)$ and $\psi_{nj}(x)$. First, for $Var[q_{nj}(x)]$

we have

$$\begin{aligned}
\mathrm{Var}[q_{nj}(x)] &= \mathrm{Var}\left[\frac{c_{Z_j}(x)}{c_{Z_j}(x)} I\{Z_j \leqslant x = X_j\}\right] \\
&= P(Z_j \leqslant x = X_j)(1 - P(Z_j \leqslant x = X_j)) \\
&= \sum_{r=0}^{x} P(Z_j = r) P(X_j = x \mid Z_j = r)(1 - P(X_j = x \mid Z_j = r)) \\
&\leqslant \sum_{r=0}^{x} P(Z_j = r) P(X_j = x \mid Z_j = r) \\
&\leqslant q_G(x). \qquad (2.6)
\end{aligned}$$

Therefore,

$$Var[q_n(x)] = Var\left[\frac{1}{n}\sum_{j=1}^{n} q_{nj}(x)\right] \leqslant \frac{q_G(x)}{n} \to 0 \quad \text{as} \quad n \to \infty.$$

Now, consider $Var[\psi_{nj}(x)]$. We will prove that for $j = 1, 2, \ldots, n$ and $x \geqslant 0$

$$0 \leqslant \psi_{nj}(x) = \frac{c_{Z_j}(x) c_1(X_j - x)}{c_{Z_j}(X_j)} I\{Z_j \leqslant x < X_j\} \leqslant 1.$$

Set $z := z_j = x_j - x$. We have for any $1 \leqslant r \leqslant x$

$$\begin{aligned}
\frac{c_1(x_j - x)}{c_r(x_j)} &= \frac{c_1(z)}{c_r(z+x)} \\
&= \frac{z+x}{rz} \frac{z^{z-1}}{(z-1)!} \frac{(z+x-r)!}{(z+x)^{z+x-r}} \\
&= \frac{z+x}{rz} z^{z-1} \frac{(z+x-r)(z+x-r-1)\ldots z}{(z+x)^{z+x-r+1-1}} \\
&= \frac{z+x}{rz} \frac{z^{z-1}}{(z+x)^{z-1}} \frac{(z+x-r)(z+x-r-1)\ldots z}{(z+x)^{x-r+1}} \\
&= \frac{1}{r}\left(\frac{z}{z+x}\right)^{z-2} \frac{(z+x-r)(z+x-r-1)\ldots z}{(z+x)^{x-r+1}}. \qquad (2.7)
\end{aligned}$$

Hence,

$$c_r(x)\frac{c_1(z)}{c_r(z+x)} = \frac{r}{x}\frac{x^{x-r}}{(x-r)!}\frac{1}{r}\left(\frac{z}{z+x}\right)^{z-2}\frac{(z+x-r)(z+x-r-1)\dots z}{(z+x)^{x-r+1}} < 1.$$

Therefore, for any $j = 1,2,\dots,n$

$$Var_n[\psi_{nj}(x)] = E_n[\psi_{nj}^2(x)] - (E_n[\psi_{nj}])^2 \leqslant E_n[\psi_{nj}^2(x)] \leqslant E_n[\psi_{nj}(x)] = \psi_G(x).$$

Thus,

$$Var[\psi_n(x)] = Var\left[\frac{1}{n}\sum_{j=1}^{n}\psi_{nj}(x)\right] \leqslant \frac{\psi_G(x)}{n} \to 0 \quad \text{as} \quad n \to \infty.$$

□

The lemma is proved.

Let us construct the EB estimator θ_n for θ given by (see also Liang (Liang 2009))

$$\theta_n(x) := \min\left\{\frac{\psi_n(x)}{q_n(x)}, 1\right\} \qquad x = r, r+1, \dots \tag{2.8}$$

Theorem 3. *For each prior distribution G, the EB estimator θ_n is asymptotically optimal.*

Proof. We have

$$S(\theta_n, \theta_G) = \sum_{x=0}^{\infty} E_n[\theta_n(x) - \theta_G(x)]^2 p_G(x),$$

where $\sum\limits_{x=0}^{\infty} p_G(x) = 1$. It is sufficient to show that

$$\lim_{n\to\infty} E_n[\theta_n(x) - \theta_G(x)]^2 = 0. \tag{2.9}$$

Recall that the second moment of a non-negative r.v. Z is given by

$$E[Z^2] = \int_0^{\infty} 2t(1 - P(Z \leqslant t))dt.$$

It follows then

$$E_n[\theta_n(x)-\theta_G(x)]^2 = \int_0^\infty 2tP\left(\left|\theta_n(x)-\theta_G(x)\right| > t\right)dt,$$

$$= \int_0^{\theta_G(x)} 2tP(\theta_n(x)-\theta_G(x) < -t)dt + \int_0^{1-\theta_G(x)} 2tP(\theta_n(x)-\theta_G(x) > t)dt.$$

It suffices then, in order to prove (2.9), to show that $\forall\, t > 0$ both

$$\lim_{n\to\infty} P(\theta_n(x)-\theta_G(x) < -t) = 0 \text{ and } \lim_{n\to\infty} P(\theta_n(x)-\theta_G(x) > t) = 0.$$

Without loss of generality, let's consider the limit of the right tail probability. For $t > 0$, we rearrange the terms to get

$$P(\theta_n(x)-\theta_G(x) > t) = P\Big(\frac{\psi_n(x)}{q_n(x)} \wedge 1 - \frac{\psi_G(x)}{q_G(x)} > t\Big)$$

$$\leqslant P\Big(\frac{\psi_n(x)}{q_n(x)} - \frac{\psi_G(x)}{q_G(x)} > t\Big) = P\Big(\psi_n(x) - \Big(t + \frac{\psi_G(x)}{q_G(x)}\Big)q_n(x) > 0\Big)$$

$$= P\left(\left[\psi_n(x)-\psi_G(x)\right] - \Big(t + \frac{\psi_G(x)}{q_G(x)}\Big)\left[q_n(x)-q_G(x)\right] > tq_G(x)\right). \tag{2.10}$$

Next, we use the following inequality. For any r.v. V and W, and $c > 0$

$$P(V-W > c) \leqslant P\Big(V > \frac{c}{2}\Big) + P\Big(W < -\frac{c}{2}\Big). \tag{2.11}$$

Indeed, for $c > 0$

$$P(V-W > c) = P\Big(V-W > c, V > \frac{c}{2}\Big) + P\Big(V-W > c, V \leqslant \frac{c}{2}\Big)$$

$$\leqslant P\Big(V > \frac{c}{2}\Big) + P\Big(V-W > c, V \leqslant \frac{c}{2}, W < -\frac{c}{2}\Big) + P\Big(V-W > c, V \leqslant \frac{c}{2}, W \geqslant -\frac{c}{2}\Big).$$

But $P\Big(V-W > c, V \leqslant \frac{c}{2}, W \geqslant -\frac{c}{2}\Big) = 0$, which implies

$$P(V-W > c) \leqslant P\Big(V > \frac{c}{2}\Big) + P\Big(W < -\frac{c}{2}\Big),$$

i.e., (2.11) holds. Applying equation (2.11) to equation (2.10), we obtain

$$P(\theta_n(x)-\theta_G(x)>t) \leqslant P\Big(\psi_n(x)-\psi_G(x)>\frac{tq_G(x)}{2}\Big)+P\left(q_n(x)-q_G(x)<\frac{-tq_G(x)}{2(t+\frac{\psi_G(x)}{q_G(x)})}\right). \quad (2.12)$$

Now by Lemma 2(i), we have $E[\psi_n(x)] = \psi_G(x)$ and $E[q_n(x) = q_G(x)]$; applying Chebysher inequality and Lemma 2(ii), we have for $t > 0$

$$P(\psi_n(x)-\psi_G(x)) > \frac{tq_G(x)}{2} \leqslant \frac{Var[\psi_n(x)]^2}{tq_G(x)/2}$$

$$\leqslant \frac{4}{t^2q_G^2(x)}\frac{\psi_G(x)}{n} \to 0 \text{ as } n\to\infty. \quad (2.13)$$

Similarly,

$$P(q_n(x)-q_G(x)) \leqslant \frac{-tq_G(x)}{2\Big(t+\frac{\psi_G(x)}{q_G(x)}\Big)} \leqslant \frac{4\Big(t+\frac{\psi_G(x)}{q_G(x)}\Big)^2}{t^2q_G^2(x)}Var\Big[q_n(x)\Big]$$

$$\leqslant \frac{4\Big(t+\frac{\psi_G(x)}{q_G(x)}\Big)^2}{t^2q_G^2(x)}\frac{q_G(x)}{n} \to 0 \text{ as } n\to\infty. \quad (2.14)$$

Therefore by (2.10)-(2.14), for any $t > 0$ we conclude

$$\lim_{n\to\infty} P(\theta_n(x)-\theta_G(x)>t) = 0.$$

Following a similar process, one can show that for any $t > 0$

$$\lim_{n\to\infty} P(\theta_n(x)-\theta_G(x)<-t) = 0,$$

the details of which are left to the reader. Hence the proof is complete. □

2.4 Bayes Risk and Regret Risk

The Bayes risk of an estimator $\hat{\theta}$ under Squared Error Loss $L(\Theta,\hat{\theta}) = (\hat{\theta}-\Theta)^2$ is defined as

$$R(\theta,\hat{\theta}) = E(\hat{\theta}-\Theta)^2 \tag{2.15}$$

where the expectation is taken with respect to both X and Θ. Then it follows by definition of $\theta_G(x)$ that the minimum Bayes Risk is given by

$$R(\theta,\theta_G) = E(\theta_G(x)-\Theta)^2$$

Therefore,

$$risk(\hat{\theta}_G(x)) = R(\theta,\theta_G) = \int_0^1 \sum_{x=0}^{\infty} (\hat{\theta}_G(x)-\theta)^2 f_{GP}(x;\theta,\tau)\, dG(\theta),$$

where $\hat{\theta}_G(x)$ is an estimator and τ is fixed and known.

Let us turn to the EB estimator θ_n defined in (2.8). When we have fixed values for $(X_1,Z_1),(X_2,Z_2)\ldots,(X_n,Z_n)$, the risk of $\theta_n((X_1,Z_1),\ldots,(X_n,Z_n);X_{n+1}) =: \theta_n(X_{n+1})$, denoted by $\tilde{R}(G,\theta_n)$, is expressed under the square error loss and given by

$$\tilde{R}(\theta,\theta_n) = E_{(X_1,Z_1),\ldots,(X_n,Z_n)}\left[E_{X_{n+1},\theta_{n+1}}\left(\theta_n(X_{n+1})-\Theta\right)^2 \mid (X_1,Z_1),\ldots,(X_n,Z_n)\right]$$

The formulation, $\tilde{R}(\theta,\theta_n)$, is known as the conditional Bayes risk of θ_n and is treated as a random variable due to its dependency on the random observed data $X_1,\ldots,X_n$.

Definition 1. The overall Bayes risk of the EB estimator θ_n is then defined by

$$R(\theta,\theta_n) := E_n\left[\tilde{R}(\theta,\theta_n)\right]$$

Here, $E_n[\cdot]$ denotes the expectation taken with respect to $((X_1,Z_1),\ldots,(X_n,Z_n))$.

In practice, the selection of the estimators often involves various criteria to determine their optimality. One such criterion is called the regret risk associated with an EB estimator θ_n, defined as the non-negative difference between the Bayes risk of the EB estimator and the Bayes risk of the Bayes estimator $\theta_G(x)$

$$S(\theta_n) := R(\theta, \theta_n) - R(\theta, \theta_G) \geqslant 0,$$

This regret risk is a standard measure of the quality (optimality) of an EB estimator. A sequence of EB estimators $\{\theta_n\}_{n=1}^{\infty}$ is defined as asymptotically optimal for a given distribution G if $\lim_{n\to\infty} S(\theta_n) = 0$. Under certain conditions, it is shown that θ_n is asymptotically optimal with a rate of convergence characterized by $O\left(n^{-\theta/2}\right)$ for some $\theta \in (0,2)$ (Liang 2009). We will use this measure of estimator quality to determine the best estimator for θ.

CHAPTER III

MONOTONE EB ESTIMATORS FOR θ IN CASE OF GPD

3.1 Monotone Likelihood Ratio Property

The EB estimator is not monotone with respect to x. We provide an illustration of θ_n in Chapter IV. This is unwanted behavior for this estimator because the GPD has a monotone likelihood ratio as it is given in the proposition below.

Proposition 4. *The GPD distribution has a monotone likelihood ratio, i.e.,*

$$q(x) = \frac{f_{GP}(x;\theta_2,\tau)}{f_{GP}(x;\theta_1,\tau)} \tag{3.1}$$

which is increasing with respect to x whenever $0 < \theta_1 < \theta_2 < 1$ *and* $x > \tau/(1-\theta)$.

Proof. Since for $x = 0, 1, \ldots$

$$f_{GP}(x;\theta,\tau) = \frac{\tau}{x!}(\tau+\theta x)^{x-1}e^{-(\tau+\theta x)} = \frac{\tau x^{x-1}}{x!}(\tau/x+\theta)^{x-1}e^{-x(\tau/x+\theta)},$$

we have

$$q(x) = \frac{f_{GP}(x;\theta_2,\tau)}{f_{GP}(x;\theta_1,\tau)} = \left[\frac{\tau/x+\theta_2}{\tau/x+\theta_1}\right]^{x-1} e^{-x(\theta_2-\theta_1)}.$$

Taking a natural logarithm, we obtain

$$\ln q(x) = (x-1)\ln\left(\frac{\tau/x+\theta_2}{\tau/x+\theta_1}\right) - x(\theta_2-\theta_1).$$

Differentiating with respect to x, we get

$$\begin{aligned}
&\frac{\partial \ln q(x)}{\partial x} = \ln\left(\frac{\tau/x+\theta_2}{\tau/x+\theta_1}\right) + (x-1)\frac{\tau}{x}\left(-\frac{1}{\tau+\theta_2 x} + \frac{1}{\tau+\theta_1 x}\right) + \theta_1 - \theta_2 \\
&= \ln(\tau/x+\theta_2) - \ln(\tau/x+\theta_1) + \theta_1 + \frac{\tau}{x} - \theta_2 - \frac{\tau}{x} + (x-1)\frac{\tau}{x}\left(\frac{1}{\tau+\theta_1 x} - \frac{1}{\tau+\theta_2 x}\right) \\
&= \ln\left[(\tau/x+\theta_2)e^{-(\tau/x+\theta_2)}\right] - \ln\left[(\tau/x+\theta_1)e^{-(\tau/x+\theta_1)}\right] + (x-1)\frac{\tau}{x}\left(\frac{1}{\tau+\theta_1 x} - \frac{1}{\tau+\theta_2 x}\right) \\
&:= A_1(x) - A_2(x) + B(x), \qquad \text{say.}
\end{aligned}$$

Since $0 < \theta_1 < \theta_2 < 1$ and $\tau > 0$, we have that $B(x)$ is positive for any positive x. It remains to show that $A_1(x) > A_2(x)$ for any $x \geqslant 1$. We will prove that the function $f(y) = ye^{-y}$ is increasing for $0 < y < 1$. Indeed, we have for $0 < y < 1$

$$f'(y) = (ye^{-y})' = e^{-y} - ye^{-y} = (1-y)e^{-y} > 0.$$

Since both terms of the derivative are positive for all $x > \tau/(1-\theta)$, we conclude that:

$$\frac{d}{dx}\ln q(x) > 0.$$

This proves that $\ln q(x)$, and hence $q(x)$, is increasing with respect to x under the given conditions, confirming the monotone likelihood ratio property of the GPD. □

The MLR property reveals a relationship between the magnitude of the observed variable and the distribution it draws from. If a distribution $f(x;\theta)$ obeys the MLR property, then the higher the observed value x the more likely it was drawn from the distribution $f(x;\theta_2$ then from $f(x;\theta_1)$ for $\theta_2 > \theta_1$.

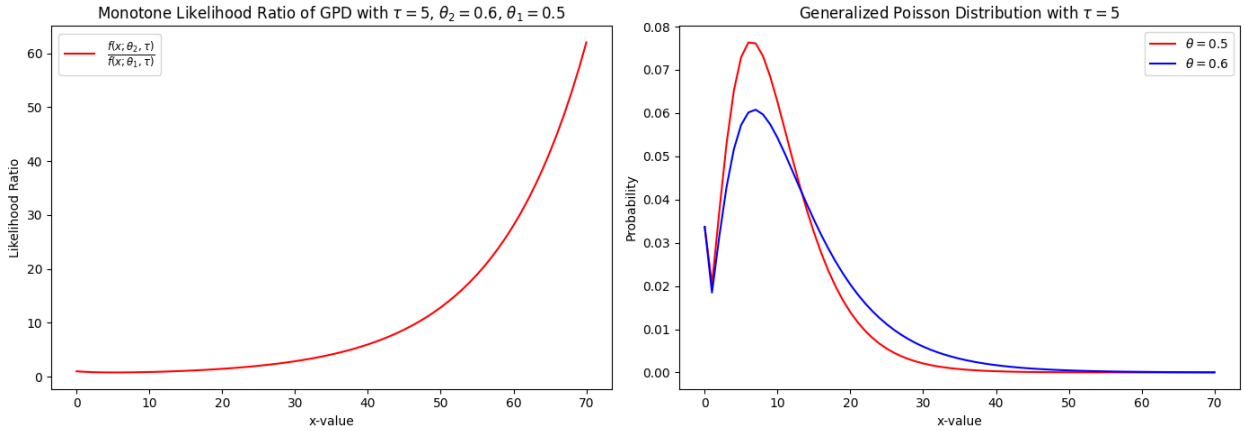

Figure 3.1: Monotone Likelihood Ratio and Generalized Poisson Distribution.

3.2 Van-Houwilengen's Monotonization Procedure

Seeing as how the monotonicity property of the Generalized Poisson Distribution (GPD) is desirable, our estimates need to have this quality as well. However, as highlighted by Van Houwalingen (Van Houwelingen 1977), the Empirical Bayes (EB) estimator θ_n, does not naturally exhibit this monotonic behavior in the context of GPD. To address this issue, Van Houwalingen outlined a method for monotonizing the EB estimator. Moreover, he demonstrated that the monotonized EB estimator, θ_n^*, not only aligns with the monotonicity of the GPD but also possesses a smaller Regret Risk than the original EB estimator θ_n, making θ_n^* a "better" estimator. In our study, we adopt this approach to monotonize θ_n for the GPD, enhancing its accuracy and reliability. In Chapter IV, we discuss yet another example of this classical construction by monotonizing the EB estimator for GPD distribution.

Estimators for discrete distributions with MLR can be made monotone by applying a procedure developed in (Van Houwelingen 1977) (see also (Yanev and Colson 2017)). Consider a simple randomized version of the estimator $\hat{\theta}_n(x)$ represented by the following function $D(a \mid x)$ for $a \in (0,1)$:

$$D(a \mid x) = \begin{cases} 0 & \text{if } \theta_n(x) > a, \\ 1 & \text{if } \theta_n(x) \leqslant a. \end{cases}$$

The number $D(a \mid x)$ is the probability that an estimate $\theta_n(x)$ less than or equal to a is selected given $X = x$. In other words, $D(a \mid x)$ is a cdf on the action space $(0,1)$ for every $X = x$. Then define for $a \in (0,1)$

$$\alpha(a) := \mathbb{E}[D(a \mid X)] = \sum_{\{x:\theta_n(x)\leqslant a\}} \mathrm{P}(x \mid a) = \sum \frac{\tau}{x!}(\tau + \theta x)^{x-1} e^{-(\tau+\theta x)} \tag{3.2}$$

Denote $F(x \mid \theta) = \sum_{k=r}^{x} \frac{\tau}{x!}(\tau + \theta x)^{x-1} e^{-(\tau+\theta x)}$ for $x \geqslant r$ and assume $F(r-1 \mid \theta) = 0$. Now, we can construct a randomized estimator with $D^*(a \mid x)$ as follows

$$D^*(a \mid x) = \begin{cases} 0 & \text{if } \alpha(a) < F(x-1 \mid a), \\ \frac{\alpha(a) - F(x-1|a)}{F(x|a) - F(x-1|a)} & \text{if } F(x-1 \mid a) \leqslant \alpha(a) \leqslant F(x \mid a), \\ 1 & \text{if } F(x \mid a) < \alpha(a), \end{cases} \tag{3.3}$$

$D^*(1 \mid x) = 1$, and $D^*(0 \mid x) = \lim_{a \to 0} D^*(a \mid x)$. Let $a \in (\theta_0, \theta_1)$ be fixed. It follows from the construction of D^*, that $\mathbb{E}[D^*(a \mid X)] = \mathbb{E}[D(a \mid X)]$.

The next proposition shows that using the monotone estimator D^*, one can construct another (non-random) monotone estimator θ_n^*, say, with risk less than or equal to the risk of the θ_n.

Proposition 5. *Let $D^*(a \mid x)$ be the monotone estimator constructed in (3.3).We introduce a non-random monotone estimator $\theta_n^*(x)$:*

$$\theta_n^*(x) := \int_0^1 a \, \mathrm{d}D^*(a \mid x). \tag{3.4}$$

The monotone non-random estimator $\theta_n^*(x)$ dominates $D^*(a \mid x)$, which in turn dominates the initial estimator $D(a \mid x)$ in terms of the Bayes risk under Square Error Loss:

$$R(\theta, \theta_n^*) \leqslant R(\theta, D^*) \leqslant R(\theta, D). \tag{3.5}$$

Proof. The proof is based on the properties of the GPD, particularly its monotone likelihood ratio. Following the theorem in (Van Houwelingen 1977) and ensuring that all assumptions are satisfied for the GPD, we can proceed with the proof: For the second inequality in (3.5), it is established that $D^*(a \mid x)$, as a monotone estimator, dominates the initial estimator $D(a \mid x)$ for all θ in the interval [0,1]. Under the Square Error Loss, we focus on showing that $D^*(a \mid x)$ is dominated by θ_n^*. The overall Bayes risk for θ_n^* is given by the expected square error. Applying Jensen's inequality we obtained

$$\begin{aligned}
R(\theta, \theta_n^*) &= E\left[(\Theta - \theta_n^*)^2\right] \\
&= E\left[\left(\Theta - \int_0^1 a\, dD^*(a \mid X)\right)^2\right] \\
&= E\left[\left(\Theta - \int_0^1 a dD^*(a \mid X)\right)^2\right] \\
&\leqslant E\left[\int_0^1 (\Theta - a)^2\, dD^*(a \mid X)\right] \\
&= E\left[\int_0^1 (\Theta - a)^2 dD^*(a \mid X)\right] \\
&= R(\theta, D^*(a, X)) \qquad (3.6)
\end{aligned}$$

□

3.3 Isotonic Regression Monotonization Procedure

As an alternative to the monotone estimator in Section 3.2, we monotonize the EB estimator using the Isotonic Regression method. The isotonic regression provides a non-decreasing sequence that best fits the data under the given constraints.

Definition. (Barlow, Brunk, and Bremner 1972) Let X be the finite set $\{x_1, \ldots, x_k\}$ with the sample order $x_1 < x_2 < \ldots < x_k$. A real-valued function f on X is isotonic if $x, y \in X$ and $x < y$, then $f(x) \leqslant f(y)$. (The term "non-decreasing" would serve equally well here). Let g be a function on X and w a given positive function on X. An isotonic function g^* on X is an isotonic regression of g with weights w with respect to the simple ordering $x_1 < x_2 < \ldots < x_k$ if it minimizes in the class of

isotonic functions f on X the sum

$$\sum_{x \in X} [g(x) - f(x)]^2 \cdot w(x)$$

When the weight function and the simple ordering are understood, we call g^* simply an isotonic regression of g.

Isotonic Regression by way of an example.

Example: Sample Isotonic Regression (Barlow, Brunk, and Bremner 1972)

Let $X = \{x_1, x_2, \ldots, x_k\}$ where $x_1 < x_2 < \ldots < x_k$. For $i = 1, 2, \ldots, k$, let $y_i(x_i)$, $j = 1, 2, \ldots, m(x)$ be a set of measurements of some quality. That is, for $x \in X$, $y_1(x), \ldots, y_{m(x)}(x)$ are observations on a distribution. Let $\mu(x)$ denote the mean of the distribution. If μ is known or assumed to be linear in x, it may be desired to estimate $\mu(x)$ by the sample linear regression. This is the solution of the problem of linear regression: to fit the data in the sense of least squares by a linear function of x, i.e., to minimize

$$\sum_{x \in X} \sum_{j=1}^{m(x)} [y_j(x) - f(x)]^2$$

in the class of linear functions f. Let

$$\bar{y}(x) = \frac{1}{m(x)} \sum_{j=1}^{m(x)} y_j(x), \qquad x \in X.$$

.

Since

$$\sum_{j=1}^{m(x)} [y_j(x) - f(x)]^2 = \sum_{j=1}^{m(x)} [y_j(x) - \bar{y}(x)]^2 + m(x)[\bar{y}(x) - f(x)]^2,$$

an equivalent problem is to minimize

$$\sum_{x \in X} [\bar{y}(x) - f(x)]^2 m(x) \tag{3.7}$$

in the class of linear functions f on X.

If no restriction were to be placed on $\mu(x)$, its least squares estimate would be obtained by minimizing 3.7 in the case of arbitrary functions f on X. The solution is clearly the function $\bar{y} : \{\bar{y}(x), x \in X\}$. In another situation, it might be known or assumed that μ is nondecreasing in x; that is, isotonic with respect to the simple order on X. A least squares estimate of $\mu(x)$ would be obtained by minimizing the weighted sum if squares 3.7 in the class of nondecreasing functions f on X, the class of functions isotonic with respect to the simple order on X: functions f such that $x_i \leqslant x_j$ implies $f(x_i) \leqslant f(x_j)$. The solution may be called the Sample Isotonic Regression.

Suppose for example that $X = \{1,2\}$, i.e, $x_1 = 1$, $x_2 = 2$. Suppose one measurement $\bar{y}_1 = \bar{y}(1) = 5$ is made on a first quantity, and one measurement, $\bar{y}_2 = \bar{y}(2) = 3$ on a second (see Figure). Then $m(1) = m(2) = 1$. Set $f_i = f(i)$, $\mu_i = \mu(i)$, $i = 1,2$. Suppose it is known that $\mu_1 \leqslant \mu_2$. Here $\bar{y}_1$ and $\bar{y}_2$ do not satisfy $\bar{y}_1 \leqslant \bar{y}_2$ and so will not serve as estimates for μ_1 and μ_2

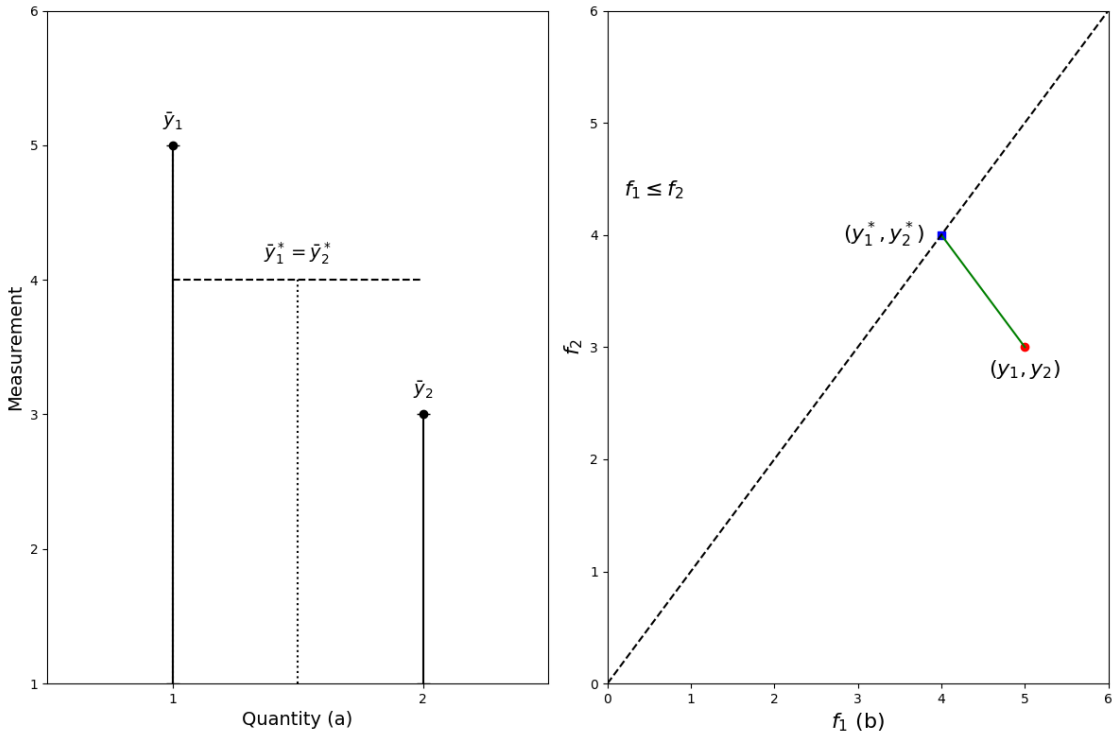

Figure 3.2: Isotonic Regression

subject to $\mu_1 \leqslant \mu_2$. In Figure 3.2 on the right figure, $(\bar{y}_1, \bar{y}_2)$ is plotted as a point in the Cartesian plane. It follows from the Pythagorean theorem that the foot $(\bar{y}_1^*, \bar{y}_2^*)$of the perpendicular onto the region $\{f_1 \leqslant f_2\}$ minimizes

$$\sum_{i=1}^{2} (\bar{y}_i - f_i)^2 = \sum_{x \in X} [\bar{y}(x) - f(x)]^2 m(x) \tag{3.8}$$

subject to $\{f_1 \leqslant f_2\}$.

CHAPTER IV

NUMERICAL STUDY: THE CASE OF GPD

In this section, we employ simulations to assess the performance of various estimators within the Generalized Poisson distribution (GPD) framework, notably focusing on the impact of the Square Error Loss. The estimators under comparison are the Bayes estimator $\theta_G(x)$, the initial EB estimator θ_n, the Van Houwelingen monotone EB estimator θ_n^*, the Isotonic Regression monotone estimator θ_n^{**}, and the maximum likelihood estimator θ_{mle}. The algorithm for the simulations is provided in Appendix A. Given the application context, particularly in epidemiological modeling where the GPD is also used, there is a compelling argument as noted in (Liang 2009) for the parameter θ to take on values in a sub-interval of $(0, 1)$. This restriction is relevant as θ typically represents a rate, which naturally falls within this range. Additionally, we consider the parameter r representing a real-world quantity such as the initial number of infected individuals entering a country with a communicable disease. Lastly, we prioritize the importance of accurate estimations in our epidemiological framework by focusing on the Square Error Loss function. This function is particularly critical as it penalizes errors in estimation, with a heightened focus on underestimations. In the context of public health, underestimating parameters like the initial number of infected individuals can have serious repercussions, potentially leading to insufficient preparedness for outbreaks. Thus, our simulations are designed to critically assess the performance of estimators in minimizing such underestimations within the GPD framework.

4.1 Bayes Estimator

In our simulation study, we adopt a Uniform prior distribution $Uni(0.5, 0.8)$ for the parameter θ. This range is selected for its epidemiological significance: at the lower end, a reproductive number of $\theta = 0.5$ suggests a dwindling epidemic, likely to extinguish without intervention. Conversely, at the higher end, a reproductive number of $\theta = 0.8$ indicates a potentially escalating viral outbreak that could become an epidemic. Setting $\tau = 5$ which represents the scenario such as the initial count of infected individuals in an outbreak, we evaluate $\theta_G(x)$. Under these assumptions we have for $x = 5, 6, \ldots, 25$, the Bayes estimator $\theta_G(x)$ is evaluated using:

$$\theta_G(x) = \frac{\int_{0.5}^{0.8} \theta(5+\theta x)^{x-1} e^{-(5+\theta x)}\, d\theta}{\int_{0.5}^{0.8} (5+\theta x)^{x-1} e^{-(5+\theta x)}\, d\theta}.$$

For example if $x = 0$ then

$$\theta_G(0) = \frac{\int_{0.5}^{0.8} \theta\, d\theta}{\int_{0.5}^{0.8} d\theta} = \frac{0.8+0.5}{2} = 0.65.$$

Thus under our settings, the minimum Bayes risk is given by

$$R(\theta, \theta_G) = \frac{1}{0.8-0.5} \sum_{x=5}^{25} \frac{5}{x!} \left(\int_{0.5}^{0.8} (\theta_G(x) - \theta)^2 (5+\theta x)^{x-1} e^{-(5+\theta x)}\, d\theta \right) \approx 0.0051.$$

where $c_5(x)$ is from (2.3).

4.2 Maximum Likelihood Estimator

Next, we will find the maximum likelihood estimator (MLE) for θ.

Proposition. The MLE for θ is given as

$$\hat{\theta}_{MLE}(x) = \max\left\{0, \frac{x-\tau-1}{x}\right\}, \qquad x \neq 0.$$

Proof. The log-likelihood of (1.1) is

$$\ln f_{GP}(x) = \ln\tau + (x-1)\ln(\tau+\theta x) - (\tau+\theta x) - \ln(x!)$$

and its partial derivative with respect to θ equals

$$\frac{\partial \ln f_{GP}(x)}{\partial\theta} = \frac{x(x-1)}{\tau+\theta x} - x.$$

Finally, setting the above derivative equals 0 and solving for θ we obtain for the MLE $\hat{\theta}_{MLE}(x)$, say

$$\hat{\theta}_{MLE}(x) = \max\left\{0, \frac{x-\tau-1}{x}\right\}, \qquad x \neq 0. \tag{4.1}$$

The maximum likelihood estimator with $\tau = 5$ is given by

$$\theta_{MLE}(x) = \max\left\{0, \frac{x-6}{x}\right\}, \qquad x = 5,6,\ldots,25.$$

The $\hat{\theta}_{MLE}(x)$ has a risk of approximately 0.0235 and a regret risk $R(\theta,\theta_{MLE}) - R(\theta,\theta_G) = 0.0184$.

The calculations of the Bayes estimator $\theta_G(x)$ and values of the maximum likelihood estimator $\theta_{MLE}(x)$ for each x from 5 to 25 are presented in Table 4.1 and Figure 4.1 to demonstrate the behavior of the Bayes and MLE estimators.

Table 4.1: Bayes and MLE estimates

x	$\theta_G(x)$	$\theta_{MLE}(x)$	x	$\theta_G(x)$	$\theta_{MLE}(x)$	x	$\theta_G(x)$	$\theta_{MLE}(x)$
5	0.63	0	12	0.64	0.50	19	0.66	0.68
6	0.63	0	13	0.64	0.54	20	0.66	0.70
7	0.63	0.14	14	0.64	0.57	21	0.66	0.71
8	0.63	0.25	15	0.65	0.60	22	0.66	0.73
9	0.63	0.33	16	0.65	0.63	23	0.67	0.74
10	0.63	0.40	17	0.65	0.65	24	0.67	0.75
11	0.64	0.45	18	0.65	0.67	25	0.67	0.76

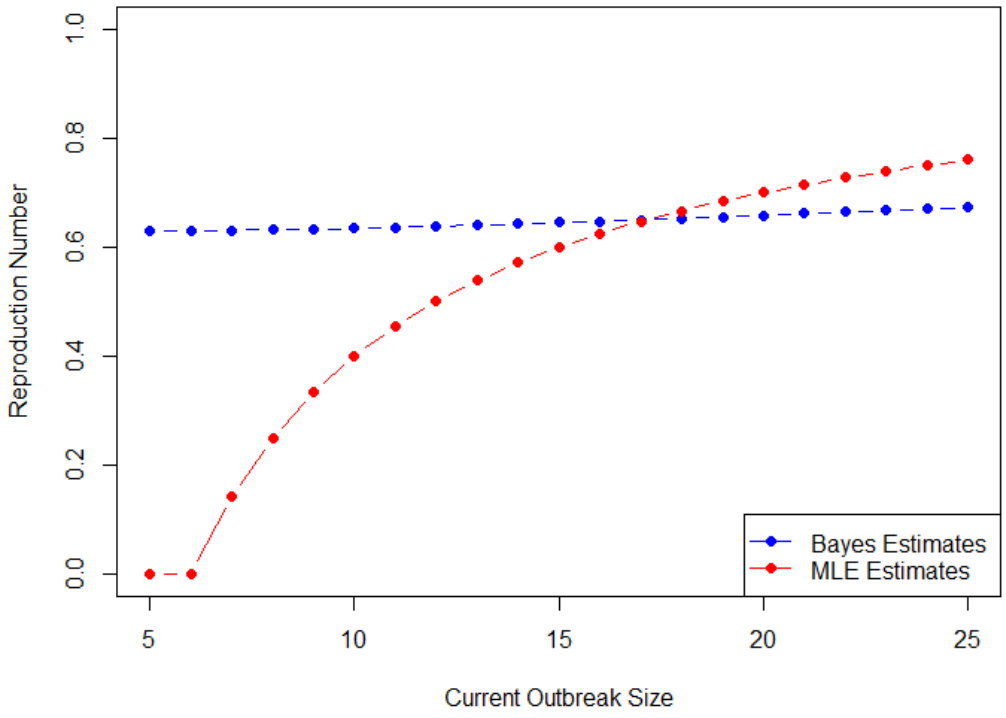

Figure 4.1: Bayes and MLE Estimators given $n = 80$, $\tau = 5$, and prior $U(0.5, 0.8)$.

4.3 Empirical Bayes Estimator: Simulation

Following the empirical Bayes framework, for the numerical study, we consider $n = 20, 40, 60, 80$ independent copies

$$(X_1, Z_1, \Theta_1), (X_2, Z_2, \Theta_2), \ldots, (X_n, Z_n, \Theta_n) \tag{4.2}$$

of the random triple (X, Z, Θ), where Θ is $Uni(0.5, 0, 8)$ variable and, given Θ, X follows the GPD distribution (1.1) and Z follows a $Poi(5)$. We assume that (X_i, Z_i) values are observable in our simulations, but Θ_i are not. We produce $m = 10$ sets of the n triples above. For each set of triples we calculate the EB estimate $\theta_n^{(j)}(x)$ where $j = 1, 2, ..., 10$ and $x = 5, 6, \ldots, 25$. This way we obtain for each $j = 1, \ldots, 10$ the following EB estimates

$$\theta_n^{(j)}(5), \theta_n^{(j)}(6), \ldots, \theta_n^{(j)}(25).$$

Next, we calculate the conditional EB risk for the j^{th} estimate above using the formula

$$\tilde{R}(\theta, \theta_n^{(j)}) = \frac{1}{0.8 - 0.5} \sum_{x=5}^{25} \int_{0.5}^{0.8} (\theta_n^{(j)}(x) - \theta)^2 GPD(x, \theta) d\,\theta, \qquad j = 1, 2, \ldots, 10.$$

After computing all 10 conditional EB risks, we estimate the overall Bayes risk $R(\theta, \theta_n)$ by

$$\hat{R}(\theta, \theta_n) = \frac{1}{10} \sum_{j=1}^{10} \tilde{R}(\theta, \theta_n^{(j)}).$$

Finally, the estimated regret risk is given by

$$\hat{S}(\theta_n) = \hat{R}(\theta, \theta_n) - R(\theta, \theta_G).$$

We repeat the above simulation procedure for $n = 20, 40, 60, 80$.

4.4 Monotonized Empirical Bayes Estimators

We proceed to monotonise the EB estimator and compute the estimate using the Van-Houwilengen's $\theta^*_{80}(x)$ and the Isotonic Regression Monotonization Procedure $\theta^{**}_{80}(x)$. We first applied Van-Houwilengen's Procedure to our empirical Bayes estimators and subsequently applied the Isotonic Regression Procedure.

Subsequently, the values of the empirical Bayes estimator $\theta_n(x)$, Van-Houwilengen's $\theta^*_n(x)$ estimator and Isotonic Regression monotonized estimator $\theta^{**}_n(x)$ for each x from 5 to 25 are presented in Table 4.2. Also, a visual depiction of their trends is provided in Figure 4.2.

Table 4.2: Empirical Bayes, Van Houwilengen, and Isotonic Regression estimates

x	$\theta_n(x)$	$\theta^*_n(x)$	$\theta^{**}_n(x)$	x	$\theta_n(x)$	$\theta^*_n(x)$	$\theta^{**}_n(x)$	x	$\theta_n(x)$	$\theta^*_n(x)$	$\theta^{**}_n(x)$
5	1.00	0.42	0.70	12	0.21	0.85	0.70	19	1.00	0.90	0.73
6	0.67	0.54	0.70	13	0.83	0.85	0.73	20	0.75	0.90	0.73
7	1.00	0.59	0.70	14	0.90	0.85	0.73	21	0.75	0.90	0.73
8	0.72	0.60	0.70	15	1.00	0.89	0.73	22	0.17	0.90	0.73
9	0.83	0.66	0.70	16	0.26	0.90	0.73	23	1.00	0.91	0.78
10	0.44	0.74	0.70	17	1.00	0.90	0.73	24	1.00	0.95	0.78
11	0.76	0.81	0.70	18	0.63	0.90	0.73	25	0.33	0.95	0.78

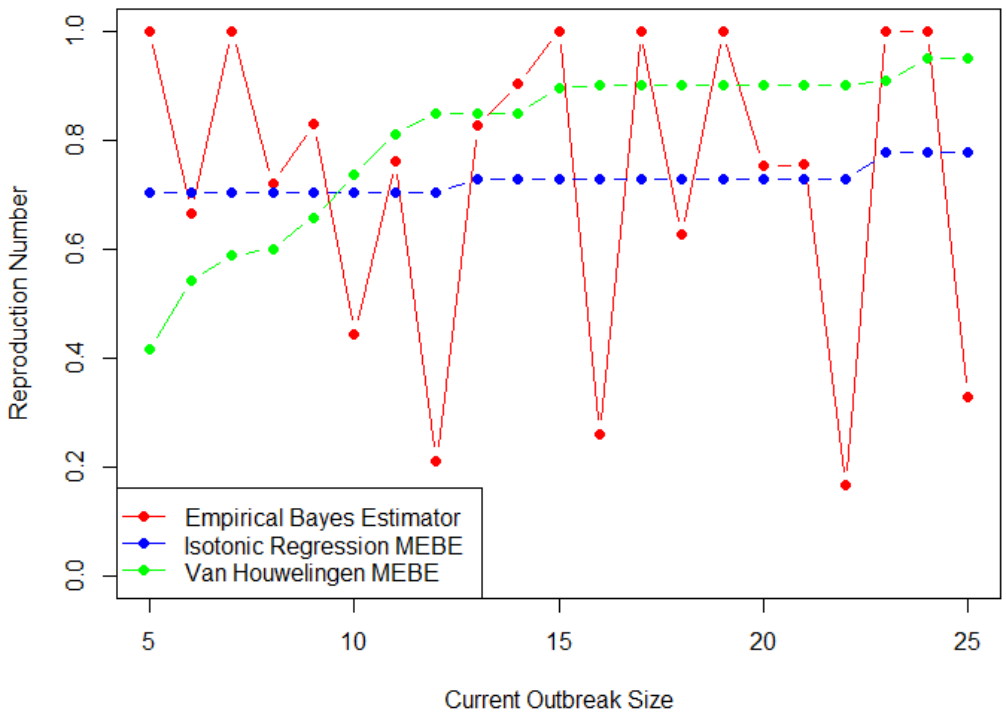

Figure 4.2: Empirical Bayes and MEB estimators given $n = 80$, $\tau = 5$, and prior $U(0.5, 0.8)$.

Similar to $S(\theta_{80})$, we also estimate the regret risk for the monotonized EB estimator, $S(\theta^*_{80})$ and $S(\theta^{**}_{80})$ by the average $\hat{S}(\theta^*_{80})$ and $\hat{S}(\theta^{**}_{80})$ respectively. For the EB estimator monotonized using the Van-Houwilengen's Monotonization Procedure, the $S(\theta^*_{80})$ by the average $\hat{S}(\theta^*_{80})$ was calculated to be -0.0011. Similarly, for the EB estimator monotonized using the Isotonic Regression Monotonization Procedure, the $S(\theta^{**}_{80})$ by the average $\hat{S}(\theta^{**}_{80})$ was estimated to be 0.00295. We repeat the entire procedure for $n = 20$, 40, and 60 as well.

We report the numerical results for the regret risks ratios w.r.t that of θ_{mle} in Table 4.3 below. The improvement of θ^*_n and θ^{**}_n over θ_n is quite substantial in terms of their regret risk.

Notice that from Table 4.3, the monotone EB estimators θ^*_n and θ^{**}_n show a substantial decrease in regret risk over θ_n, with θ^{**}_n showing a slightly higher reduction in regret risk compared to θ^*_n in most cases.

Table 4.3: Change of Regret Risks of θ_n, θ_n^*, and θ_n^{**} in terms of percent from $\hat{S}(\theta_{mle}) = 0.0184$.

n	$\hat{S}(\theta_n)$	$\hat{S}(\theta_n^*)$	$\hat{S}(\theta_n^{**})$
20	▲ 21.52%	▼ -85.29%	▼ -102.83%
40	▼ -13.01%	▼ -112.29%	▼ -113.51%
60	▼ -19.31%	▼ -116.51%	▼ -115.26%
80	▼ -27.21%	▼ -104.89%	▼ -113.17%

Note. All standard errors are less than 10^{-4} and $\tau = 5$.

Additionally, we present the estimates based on a single set of size $n = 80$ triples from (4.2) along with the maximum likelihood and Bayes estimate in Figure 4.3 to illustrate the estimators' behavior.

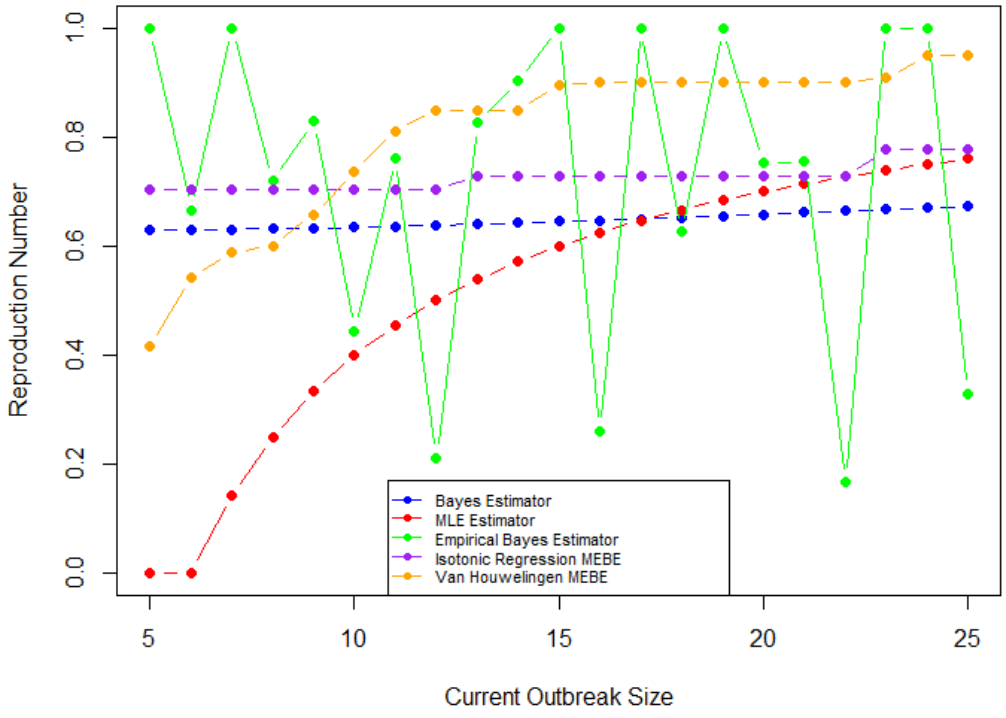

Figure 4.3: Estimators given $n = 80$, $\tau = 5$, and prior $U(0.5, 0.8)$.

CHAPTER V

CLOSING REMARKS

In this paper, we studied the estimation problem for the reproduction parameter θ of Generalized Poisson distribution. Our interest stemmed from applying branching processes as models of epidemic outbreaks where θ equals the average number of secondary infections caused by a host. Using the Isotonic Regression method (Barlow, Brunk, and Bremner 1972) and Van Houwelingen method (Van Houwelingen 1977), we constructed a monotone empirical Bayes estimators θ_n^{**} and θ_n^{*} for θ based on the empirical Bayes estimator θ_n proposed by Liang (Liang 2009). These new monotone estimators are strictly better than the original empirical estimation supported by having a smaller regret risk than both the empirical and maximum likelihood estimates with θ_n^{**} showing a slightly higher reduction in regret risk compared to θ_n^{*}. The non-monotone empirical Bayes estimator θ_n turns out to be quite jumpy (see Figure 4.3) and does not have good small sample properties (see Table 4.3). Simulation results show that θ_n^{**} and θ_n^{*} perform much better than θ_n, especially when the number of past observations and/or the epidemic size are small. This confirms the major positive effect of the monotonization procedure. In addition, the square error loss function is incredibly powerful for epidemic analysis. Due to its symmetric nature and the capacity to provide a clear measure of the estimation accuracy both underestimating and overestimating are penalized equally. When running simulations, we saw that the monotone estimators θ_n^{*} and θ_n^{**} again outperformed the other estimates, indicated by the smaller regret risk.

Generally, the comparison of various estimators—Bayes estimator, EB estimator, monotone EB estimators, and the maximum likelihood estimator—underlines the superiority of the monotone EB estimators in minimizing square error loss which is a crucial aspect in epidemiological modeling where underestimation can have significant public health implications.

REFERENCES

Albertsen, K., J.F Steffensen, and E. Kirstensen (Nov. 1992). *THREE PAPERS ON THE HISTORY OF BRANCHING PROCESSES*. Tech. rep. 242. Translated from Danish by Peter Guttorp. Seattle, Washington: Department of Statistics, GN-22 University of Washington.

Aldous, David J. (Feb. 1999). "Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilists". In: *Bernoulli* 5.1, pp. 3–48. URL: https://projecteuclid.org:443/euclid.bj/1173707093.

Barlow, R. E., H. D. Brunk, and J. M. Bremner (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley.

Carlin Bradley, P. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd. Chapman and Hall,

Charalambides, Ch. A. (1990). "Abel Series Distributions with applications to Fluctuations of Sample Functions of Stochastic Processes". In: *Commun. Statist. -Theory Meth.* 19.1, pp. 317–335.

Christine, J. (2010). "Branching Processes: Their Role in Epidemiology". In: *Int. J. Environ. Res. Public Health* 1186. URL: 10.3390/ijerph7031204.

Consul C., P. (1989). *Generalized Poisson Distribution Properties and Applications*. Marcel Dekker Inc.

Consul, P.C. and G.C. Jain (1970). "On the generalization of Poisson distribution". In: *Annals of Mathematical Statistics* 41, p. 1387.

De Serres, Gaston, Nigel J. Gay, and C. Paddy Farrington (2000). "Epidemiology of Transmissible Diseases after Elimination". In: *American Journal of Epidemiology* 151, pp. 1039–1048.

Farrington, C. P., M. N. Kanaan, and N. J. Gay (2003). "Branching process models for surveillance of infectious diseases controlled by mass vaccination". In: *Biostatistics* 4.2, pp. 279–295.

Ferguson, Neil M. et al. (2004a). "Public Health Risk from the Avian H5N1 Influenza Epidemic". In: *Science Supporting Online Material*, pp. 1–5.

— (May 2004b). "Public Health Risk from the Avian H5N1 Influenza Epidemic". In: *Science* 304.5673, pp. 968–969.

Gipps, P. G. (1976). "An Abbreviated Procedure for Estimating Equilibrium Queue Lengths in Rural Two Lane Traffic." In: *Transportation Science* 10.4, p. 337. ISSN: 00411655. URL: http://ezhost.utrgv.edu:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=6704521&site=ehost-live.

Iešmantas, Tomas and Robertas Alzbutas (2014). "Bayesian assessment of electrical power transmission grid outage risk". In: *International Journal of Electrical Power & Energy Systems* 58, pp. 85–90. ISSN: 0142-0615. DOI: https://doi.org/10.1016/j.ijepes.2014.01.006. URL: http://www.sciencedirect.com/science/article/pii/S0142061514000076.

Jansen, V. A. A. et al. (Aug. 2003). "Measles Outbreaks in a Population with Declining Vaccine Uptake". In: *Science* 301.5634, p. 804.

Koorey, Glen (2007). "Passing Opportunities at Slow-Vehicle Bays". In: *Journal of Transportation Engineering* 133.2, pp. 129–137. DOI: 10.1061/(ASCE)0733-947X(2007)133:2(129). eprint: https://ascelibrary.org/doi/pdf/10.1061/%28ASCE%290733-947X%282007%29133%3A2%28129%29. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%290733-947X%282007%29133%3A2%28129%29.

Liang, TaChen (2009). "Empirical Bayes estimation for Borel–Tanner distributions". In: *Statistics & Probability Letters* 79.20, pp. 2212–2219. ISSN: 0167-7152. DOI: https://doi.org/10.1016/j.spl.2009.07.018. URL: http://www.sciencedirect.com/science/article/pii/S0167715209002727.

Nirei, Makoto, Theodoros Stamatiou, and Vladyslav Sushko (Feb. 2012). *Stochastic Herding in Financial Markets Evidence from Institutional Investor Equity Portfolios*. Tech. rep. 371. Bank for International Settlements.

Robbins, H. (1964). "The empirical Bayes approach to statistical decision problems". In: *The Annals of Mathematical Statistics* 35, pp. 1–20.

Sellke, S. H., N. B. Shroff, and S. Bagchi (Apr. 2008). "Modeling and Automated Containment of Worms". In: *IEEE Transactions on Dependable and Secure Computing* 5.2, pp. 71–86. ISSN: 1545-5971. DOI: 10.1109/TDSC.2007.70230.

Stijnen, Th. (1980). "On the asymptotic behaviour of monotonized empirical Bayes rules". PhD thesis. The Netherlands: University of Utrecht.

Van Houwelingen, J. C. (1977). "Monotonizing empirical Bayes estimators for a class of discrete distributions with monotone likelihood ratio". In: *Statistica Neerlandica* 31.3, pp. 95–104. DOI: 10.1111/j.1467-9574.1977.tb00756.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9574.1977.tb00756.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.1977.tb00756.x.

Vynnycky, Emilia, G. White Richard, and E. Fine Paul (2010). *An Introduction to Infectious Disease Modelling*. 1st. Oxford University Press Inc., New York.

Yanev, George P. (2001). "Statistical Modeling of Epidemic Disease Propagation via Branching Processes and Bayesian Inference". PhD thesis. Tampa, Florida: University of South Florida.

Yanev, George P. and Roberto Colson (Oct. 2017). "Monotone Empirical Bayes Estimators for the Reproduction Number in Borel-Tanner Distribution". In: *Pliska Studia Mathematica* 27, pp. 115–122.

APPENDIX A

APPENDIX A

```r
install.packages("pracma")  # pracma package, which provides the integral2 function for numerical integration.
library(pracma)
if(!require(VGAM)) install.packages("VGAM", repos = "http://cran.us.r-project.org") #install required packages
library(VGAM)
library(ggplot2)
options(scipen = 999) # suppresses scientific notation in the output

# Parameters
tau <- 5
max_x <- 25
a <- 0.5
b <- 0.8

# Function to integrate for numerator
psi_G <- function(theta, tau, x) {
  return(theta * (tau + theta*x)^(x-1) * exp(-(tau + theta*x)))
}

# Function to integrate for denominator
q_G <- function(theta, tau, x) {
  return((tau + theta*x)^(x-1) * exp(-(tau + theta*x)))
}

# Bayes estimate for single x
theta_G <- function(tau, x, a, b) {
  integral_num <- integrate(psi_G, lower = a, upper = b, tau = tau, x = x)$value
  integral_den <- integrate(q_G, lower = a, upper = b, tau = tau, x = x)$value
  return(integral_num / integral_den)
}

# Vectors to store x values and Bayes estimate results
x_values <- 5:max_x
theta_G_estimates <- numeric(length(x_values)) # Fix here: use length(x_values) instead of max_x

# Compute Bayes estimate for each x in range 5 to max_x
for (i in 1:length(x_values)) { # Use i as the index
  estimate <- theta_G(tau, x_values[i], a, b)
  theta_G_estimates[i] <- estimate # Store sequentially
}

# Plot Bayes estimates
plot(x_values, theta_G_estimates, type="b", ylim=c(0,1), col="blue", pch=16, xlab="Current Outbreak Size", ylab="Reproduction Number", main="Bayes Estimates")

```

```r

# Print the values of x_values and their corresponding theta_G_estimates
for (i in 1:length(x_values)) {
  cat("For x =", x_values[i], ", Theta_G estimate =", theta_G_estimates[i], "\n")
}


# Now, plot the Bayes estimates
plot(x_values, theta_G_estimates, type = "b", ylim = c(0,1), col = "blue", pch = 19,
     xlab = "Current Outbreak Size", ylab = "Reproduction Number",
     main = "Bayes Estimates")


# Function to compute risk for single x
risk_single_x <- function(tau, x, a, b) {
  theta_G_x <- theta_G(tau, x, a, b)
  integrand <- function(theta, tau, x, theta_G_x) {
    return(((theta_G_x - theta)^2) * (tau + theta*x)^(x-1) * exp(-(tau + theta*x)))
  }
  integral <- integrate(integrand, lower = a, upper = b, tau = tau, x = x, theta_G_x = theta_G_x)$value
  return(5 / factorial(x) * integral)
}

# Compute minimum risk
min_risk <- 0
for (x in 5:max_x) {
  min_risk <- min_risk + risk_single_x(tau, x, a, b)
}
min_risk <- min_risk / (b - a)

print(paste("Minimum Bayes risk:", min_risk))


```

```r

# Maximum Likelihood Estimator
# Parameters
tau <- 5
max_x <- 25

# Function to compute MLE estimate for a single x
theta_MLE <- function(x, tau) {
  if (x > tau) {
    return(max(0, (x - tau - 1) / x))
  } else {
    return(0) # Ensure theta_MLE does not return negative values
  }
}

# Vectors to store x values and MLE estimate results
x_values <- seq(5, max_x) # Adjusted to start from 5
theta_MLE_estimates <- numeric(length(x_values))

# Compute MLE estimate for each x in range 5 to max_x
for (i in seq_along(x_values)) {
  x <- x_values[i]
  estimate <- theta_MLE(x, tau)
  theta_MLE_estimates[i] <- estimate
  print(paste("Theta_MLE for x =", x, "is", estimate))
}

# Plot MLE estimates
plot(x_values, theta_MLE_estimates, type="b", ylim=c(0,1), col="blue", pch=16,
     xlab="Current Outbreak Size", ylab="Reproductive Number",
     main="MLE Estimates for Reproductive Number")

# risk calculation function
risk_single_x <- function(tau, x, a, b) {
  theta_MLE_x <- theta_MLE(x, tau)
  if (is.na(theta_MLE_x)) {
    return(0)
  }
  integrand <- function(theta, tau, x, theta_MLE_x) {
    return(((theta_MLE_x - theta)^2) * (tau + theta * x)^(x - 1) * exp(-(tau + theta * x)))
  }
  integral <- integrate(integrand, lower = a, upper = b, tau = tau, x = x, theta_MLE_x = theta_MLE_x)$value
  return(5 / factorial(x) * integral)
}
```

```r

# Define the limits for theta
a <- 0.5
b <- 0.8 # Adjust as per the actual limits of theta

# Compute minimum risk
min_risk_mle <- 0
for (x in 5:max_x) {
  min_risk_mle <- min_risk_mle + risk_single_x(tau, x, a, b)
}
min_risk_mle <- min_risk_mle

print(paste("Minimum Bayes risk:", min_risk_mle))

##############################################################################################

# First, plot the Bayes estimates
plot(x_values, theta_G_estimates, type="b", ylim=c(0,1), col="blue", pch=16,
     xlab="Current Outbreak Size", ylab="Reproduction Number",
     main="Comparison of Bayes and MLE Estimates")

# Then, add the MLE estimates to the same plot
points(x_values, theta_MLE_estimates, type="b", col="red", pch=16, lty=1)

# Add a legend
legend("bottomright", legend=c("Bayes Estimates", "MLE Estimates"),
       col=c("blue", "red"), pch=c(16, 16), lty=c(1, 1))


##############################################################################################
```

```r
# Parameters
n <- 80
tau <- 5  # Corrected comment to match the tau value
x_min <- 5
x_max <- 25
x <- x_min:x_max
m <- 10  # Number of realizations
seeds <- c(9, 28, 54, 55, 976, 442, 28, 1, 151, 38)

# Initialize a matrix to store results for each m and x
results_n <- matrix(NA, nrow = m, ncol = length(x), dimnames = list(1:m, as.character(x)))

# Loop for m realizations
for (j in 1:m) {
  set.seed(seeds[j])

  # Generate Z, theta, and X values
  Z <- rpois(n, tau)
  theta <- runif(n, 0.5, 0.8)
  X <- numeric(n)
  for (i in 1:n) {
    X[i] <- rbort(1, Qsize = tau, a = theta[i])
  }

  # Defining the C functions with corrected error handling
  c1 <- function(x, X) {
    term <- X - x
    ifelse(term <= 0, 1, (1 / term) * (term^(term - 1)) / factorial(term - 1))
  }

  cZx <- function(x, Z) {
    term_diff <- x - Z
    ifelse(term_diff < 0, 0, ifelse(term_diff == 0, 1, (Z / x) * (x^term_diff) / factorial(term_diff)))
  }

  cZX <- function(X, Z) {
    term_diff <- X - Z
    ifelse(term_diff < 0, 0, ifelse(term_diff == 0, 1, (Z / X) * (X^term_diff) / factorial(term_diff)))
  }

  psi_nj <- function(x, X, Z) {
    numerator <- cZx(x, Z) * c1(x, X)
    denominator <- cZX(X, Z)
    if (is.na(denominator) || is.nan(denominator) || denominator == 0) return(0)
    indicator <- as.integer(Z <= x & x < X)
    return((numerator / denominator) * indicator)
  }

  qnj <- function(x, X, Z) {
    value <- cZx(x, Z) / cZX(X, Z)
    if (is.nan(value) || is.infinite(value) || value == 0) return(0)
    indicator <- as.integer(Z <= x & x == X)
    return(value * indicator)
  }

  psi_n <- function(x) {
    total <- sum(mapply(psi_nj, rep(x, n), X, Z))
    return(total / n)
  }

  qn <- function(x) {
    total <- sum(mapply(qnj, rep(x, n), X, Z))
    return(total / n)
  }

  theta_n <- function(x) {
    psi_val <- psi_n(x)
    qn_val <- qn(x)
    if (qn_val == 0) return(1)
    return(min(psi_val / qn_val, 1))
  }

  # Calculate and store the results
  for (xi in x) {
    results_n[j, as.character(xi)] <- theta_n(xi)
  }
}

# Print the results matrix
print(results_n)

# Assuming results_n is already calculated
plot(x, results_n[1, ], type = 'b', ylim = c(0, 1), col = 'blue', pch = 16, lty = 1,
     xlab = 'X', ylab = 'Theta_n',
     main = 'Empirical Bayes Estimates for the First Set')
grid()
```

```r
# Parameters
tau <- 5 # Define the value of τ based on your model's specification
x_max <- 25
m <- 10  # Number of realizations


# Define the GPD function
gpd <- function(x, theta, tau) {
  if (x == 0) {
    return(tau * exp(-tau))
  } else {
    return(tau / factorial(x) * (tau + theta * x)^(x - 1) * exp(-(tau + theta * x)))
  }
}


# Function to calculate conditional EB risk for a given j
conditional_eb_risk <- function(j, results_n, tau, x_range) {
  integrand <- function(theta) {
    risk_sum <- 0
    for (x in x_range) {
      estimated_theta <- results_n[j, as.character(x)]
      risk_sum <- risk_sum + (estimated_theta - theta)^2 * gpd(x, theta, tau)
    }
    return(risk_sum)
  }

  # Integrate over theta from 0.5 to 0.8 (as per your uniform distribution range for theta)
  risk <- integrate(integrand, lower = 0.5, upper = 0.8)$value
  return(risk)
}

# Calculate conditional EB risks and overall Bayes risk
x_range <- 5:25  # Adjusted x range from 5 to 25
conditional_risks <- sapply(1:m, function(j) conditional_eb_risk(j, results_n, tau, x))
print(conditional_risks)
overall_bayes_risk <- mean(conditional_risks)

# Print the overall Bayes risk
print(overall_bayes_risk)
```

```r
# Apply isotonic regression to each set of estimates in results_n
iso_results_n <- matrix(NA, nrow = m, ncol = length(x), dimnames = dimnames(results_n))

# Loop through each set of estimates
for (j in 1:m) {
  # Apply isotonic regression
  iso_fit <- isoreg(x, results_n[j, ])

  # Store the fitted (isotonic regression) values
  iso_results_n[j, ] <- iso_fit$yf
}

# Print the isotonic regression-adjusted results matrix
print(iso_results_n)

# Optionally, plot the original and isotonic regression-adjusted estimates for a specific set
set_to_plot <- 1  # Change this to plot a different set

# Plotting the original estimates
plot(x, results_n[set_to_plot, ], type = 'b', col = 'red', pch = 16, lty = 1,
     main = paste('Original vs Isotonic Regression Estimates for Set', set_to_plot),
     xlab = 'X', ylab = 'Theta_n')

# Adding the isotonic regression-adjusted estimates
lines(x, iso_results_n[set_to_plot, ], type = 'b', col = 'blue', pch = 18, lty = 2)

# Adding a legend
legend("bottomright", legend = c("Original", "Isotonic Regression"),
       col = c("red", "blue"), pch = c(19, 18), lty = c(1, 2))


# Parameters
tau <- 5 # Define the value of τ based on your model's specification
x_max <- 25
m <- 10  # Number of realizations


# Define the GPD function
gpd <- function(x, theta, tau) {
  if (x == 0) {
    return(tau * exp(-tau))
  } else {
    return(tau / factorial(x) * (tau + theta * x)^(x - 1) * exp(-(tau + theta * x)))
  }
}
```

```r
# Function to calculate conditional EB risk for a given j
conditional_ieb_risk <- function(j, iso_results_n, tau, x_range) {
  integrand <- function(theta) {
    risk_sum <- 0
    for (x in x_range) {
      estimated_theta <- iso_results_n[j, as.character(x)]
      risk_sum <- risk_sum + (estimated_theta - theta)^2 * gpd(x, theta, tau)
    }
    return(risk_sum)
  }

  # Integrate over theta from 0.5 to 0.8 (as per your uniform distribution range for theta)
  risk <- integrate(integrand, lower = 0.5, upper = 0.8)$value
  return(risk)
}

# Calculate conditional EB risks and overall Bayes risk
x_range <- 5:25  # Adjusted x range from 5 to 25
conditional_risks <- sapply(1:m, function(j) conditional_ieb_risk(j, iso_results_n, tau, x_range))
print(conditional_risks)
overall_bayes_risk <- mean(conditional_risks)

# Print the overall Bayes risk
print(overall_bayes_risk)
```

```r

##########################################################################################################################################

#Monotonizing the Empirical Bayes Estimate using Van Houwelingen
# Define the lower and upper bounds of G prior ~ uniform(a,b)
a <- 0.5
b <- 0.8

# Initialize parameters for monotonization
r <- 5 # initial outbreak size
Xmax <- 25 # max no. current OB size

# Set the seeds for reproducibility
seeds <- c(9, 28, 54, 55, 976, 442, 28, 1, 151, 38)

pmax <- 20
alpha <- matrix(0, length(seeds), pmax)
aG <- seq(0, 1, length.out = pmax + 2) # Create a grid value
aG <- aG[-c(1, pmax + 2)]

# Alpha calculation loop
for (j in 1:length(seeds)) {
  set.seed(seeds[j]) # Set the seed for reproducibility
  for (i in 1:pmax) {
    for (k in 1:length(x)) {
      # Add a check to ensure theta_values[k] is not NA or NaN
      if (!is.na(results_n[k]) && !is.nan(results_n[k]) && results_n[k] <= aG[i]) {
        alpha[j, i] <- alpha[j, i] + sum(rgenpois(1, lambda1 = r, lambda2 = aG[i])) # compute alpha
      }
    }
  }
}

# FGT calculation
FGT <- matrix(0, length(x), pmax)
for (i in 1:pmax) {
  FGT[1, i] <- rgenpois(1, lambda1 = r, lambda2 = aG[i]) # BT cdf for x=r
  for (k in 2:length(x)) {
    FGT[k, i] <- FGT[k - 1, i] + rgenpois(1, lambda1 = r, lambda2 = aG[i]) # BT cdf for x>r
  }
}

# D^*(a,x) and other variables
Dstar <- matrix(0, length(x), pmax)
Dtail <- matrix(0, length(x), pmax)
mEB <- matrix(0, length(seeds), length(x)) # Monotonized EB estimates
```

```r
# Main loop for monotonization
for (j in 1:length(seeds)) {
  for (i in 1:pmax) {
    Dstar[1, i] <- ifelse(alpha[j, i] > FGT[1, i], 1, alpha[j, i] / FGT[1, i])
    for (k in 2:length(x)) {
      if (FGT[k - 1, i] > alpha[j, i]) {
        Dstar[k, i] <- 0
      } else if (FGT[k, i] <= alpha[j, i]) {
        Dstar[k, i] <- 1
      } else {
        Dstar[k, i] <- (alpha[j, i] - FGT[k - 1, i]) / (FGT[k, i] - FGT[k - 1, i])
      }
    }
  }

  # Calculation of Dtail and tempmEB
  for (k in 1:length(x)) {
    Dtail[k, ] <- 1 - Dstar[k, ]
    tempmEB <- sum(Dtail[k, ]) / pmax # define monotonized Empirical estimate
    mEB[j, k] <- tempmEB
  }
}

# Print the final monotonized estimates
for (j in 1:length(seeds)) {
  cat("Monotonized estimates for seed", seeds[j], ":", mEB[j,], "\n")
}


# Correctly defining x_values based on your provided range
x_values <- seq(5, Xmax)  # Assuming Xmax is 20 as mentioned


# Parameters
tau <- 5 # Define the value of τ based on your model's specification
x_max <- 25
m <- 10  # Number of realizations


# Risk for Empirical Bayes Estimator
intEBE <- function(theta, x_i, mEB, lambda1) {
  f_x <- dgenpois(x_i, lambda1, lambda2 = theta)
  Cx <- ifelse(x_i >= 3, (3 / x_i) * (x_i^(x_i - 3) / factorial(x_i - 3)), 1)
  integrand <- (mEB- theta)^2 * (f_x / Cx)
  return(integrand)
}
```

```r

# Set parameters for intEBE function
lambda1 <- 5
a <- 0.5
b <- 0.8

# Initialize REB vector
REB <- numeric(length(x))

# Compute REB values
for (i in 1:length(x)) {
  Cx <- ifelse(x[i] >= 3, (3 / x[i]) * (x[i]^(x[i] - 3) / factorial(x[i] - 3)), 1)
  REB[i] <- (Cx / (b - a)) * integrate(intEBE, lower = a, upper = b, x_i = x[i], mEB= mEB[i], lambda1 = lambda1)$value
}

# Print the calculated REB values
cat("Risk Empirical Bayes (REB) values for each x in 0:20:\n")
print(REB)

sum(REB)/length(seeds)

# Plotting the first set of monotonized EB estimates for the first seed
plot(x_values, mEB[1, ], type = 'b',
     ylim = c(0, max(mEB[1, ], na.rm = TRUE)),
     xlab = 'Current Outbreak Size',
     ylab = 'Monotonized Empirical Bayes Estimates',
     main = sprintf("Monotonized EB Estimates for Seed %d", seeds[1]),
     col = 'blue', pch = 19, lty = 1)


##########################################################################################################

# Plotting
plot(x, results_n[1, ], type = 'b', col = 'red', pch = 16, lty = 1, ylim = c(0,1),
     main = "Comparison of Estimates",
     xlab = "Current Outbreak Size", ylab = "Reproduction Number")
lines(x, iso_results_n[1, ], type = 'b', col = 'blue', pch = 16, lty = 1)
lines(x, mEB[1, ], type = 'b', col = 'green', pch = 16, lty = 1)

# Adding a legend with a smaller size
legend("bottomright", legend = c("Empirical Bayes", "Isotonic Regression", "Van Houwelingen EB"),
       col = c("red", "blue", "green"), pch = c(16, 16, 16), lty = c(1, 1, 1), cex = 0.75)

```

```
#########################################################################################################################

# All Plots (Bayes Estimates, MLE Estimates, Empirical Bayes, Isotonic Regression MEB, Van Houwelingen MEB)
plot(x_values, theta_G_estimates, type="b", ylim=c(0,1), col="blue", pch=16,
     xlab="Current Outbreak Size", ylab="Reproduction Number",
     main="Comparison of Estimation Methods")

# Add MLE estimates
points(x_values, theta_MLE_estimates, type="b", col="red", pch=16, lty=1)

# Add Empirical Bayes Estimates
points(x_values, results_n[1, ], type="b", col="green", pch=16, lty=1)

# Add Isotonic Regression Estimates
points(x_values, iso_results_n[1, ], type="b", col="purple", pch=16, lty=1)

# Add Monotonized EB Estimates
points(x_values, mEB[1, ], type="b", col="orange", pch=16, lty=1)

# Add a legend
legend("bottomright", legend=c("Bayes Estimates", "MLE Estimates", "Empirical Bayes", "Isotonic Regression MEB", "Van Houwelingen MEB"),
       col=c("blue", "red", "green", "purple", "orange"), pch=c(16, 16, 16, 16, 16), lty=c(1, 1, 1, 1, 1), cex = 0.55)
```

APPENDIX B

APPENDIX B

Table B1: References on notation

Notation	**Description**
Θ	rv parametrizing X; the reproduction number
θ	a realization of the reproduction parameter Θ
$\hat{\theta}$	refers to any estimator
θ_G	Bayes estimator
θ_{mle}	Maximum likelihood estimator for GPD distribution
θ_n	Empirical Bayes estimator for GPD
θ_n^*	Monotonized EB estimator for GPD based on Van Houwelingen
θ_n^{**}	Monotonized EB estimator for GPD based Isotonic Regression
$Poi(\lambda)$	Poisson distribution with parameter λ
$R(G,\hat{\theta})$	Bayes risk for estimator $\hat{\theta}$ under $G-$prior
$R(\hat{\theta})$	Regret risk for estimator $\hat{\theta}$
$\hat{S}(\hat{\theta})$	Average regret risk for estimator $\hat{\theta}$
$Uni(a,b)$	Uniform distribution with parameters (a,b)
GPD	Generalized Poisson distribution
cdf	cummulative distribution function
EB	Empirical Bayes
GW	Galton–Watson also known as Bienaymé–Galton–Watson
iid	independent identically distributed
MLE	maximum likelihood estimator
MLR	monotone likelihood ratio
pmf	probability mass function
rv	random variable
MEBE	monotone empirical Bayes Estimator
MEB	monotone empirical Bayes

ALGORITHM 1: Bayes Estimate and Minimum Bayes Risk

/* Parameters are set to $\tau = 5$, $max\_x = 25$, $a = 0.5$, $b = 0.8$. */
Generate $x\_values = 5, 6, \ldots, max\_x$ /* Vector of current outbreak sizes */
Initialize *theta_G_estimates* as an empty vector of length *x_values* /* For storing Bayes estimates */
for *x in x_values* **do**
    Compute integral numerator as $\int_a^b \theta \cdot (\tau + \theta \cdot x)^{x-1} \cdot e^{-(\tau+\theta \cdot x)} d\theta$ /* Function *psi_G* for numerator */
    Compute integral denominator as $\int_a^b (\tau + \theta \cdot x)^{x-1} \cdot e^{-(\tau+\theta \cdot x)} d\theta$ /* Function *q_G* for denominator */
    Compute $\theta\_G(x)$ as the ratio of integral numerator to integral denominator /* Bayes estimate for single *x* */
    Store $\theta\_G(x)$ in *theta_G_estimates* corresponding to *x*
end
Initialize $min\_risk \leftarrow 0$ /* Accumulate minimum Bayes risk */
for *x in x_values* **do**
    Compute *risk_single_x* for each *x* using the *risk_single_x* function /* Compute risk for single *x* */
    $min\_risk \leftarrow min\_risk + risk\_single\_x$ /* Accumulate risk */
end
$min\_risk \leftarrow min\_risk/(b-a)$ /* Average minimum Bayes risk */
Print "Minimum Bayes risk:", *min_risk* /* Output minimum Bayes risk */

ALGORITHM 2: MLE Estimate and it Minimum Risk

/* Parameters are set to $\tau = 5$, $max\_x = 25$, $a = 0.5$, $b = 0.8$. */
Generate $x\_values = 5, 6, \ldots, max\_x$ /* Vector of current outbreak sizes */
Initialize *theta_MLE_estimates* as an empty vector of length *x_values* /* For storing MLE estimates */
for *x in x_values* **do**
    if $x > \tau$ **then**
        $theta\_MLE(x) \leftarrow \max(0, (x - \tau - 1)/x)$ /* MLE estimate for $x > \tau$ */
    end
    else
        $theta\_MLE(x) \leftarrow 0$ /* Ensures *theta_MLE* does not return negative values */
    end
    Store $theta\_MLE(x)$ in *theta_MLE_estimates* corresponding to *x* Print "Theta_MLE for $x =$", *x*, "is", $theta\_MLE(x)$ /* Output MLE estimate */
end
Initialize $min\_risk\_mle \leftarrow 0$ /* To accumulate minimum risk for MLE */
for *x in x_values* **do**
    Compute *risk_single_x* for each *x* using the *risk_single_x* function /* Compute risk for single *x* */
    $min\_risk\_mle \leftarrow min\_risk\_mle + risk\_single\_x$ /* Accumulate risk */
end
Print "Minimum MLE risk:", *min_risk_mle* /* Output minimum MLE risk */
Calculate Regret risk $S(\theta_{mle})$

ALGORITHM 3: Empirical Bayes Estimator: Simulation Procedure

```
/* Simulation of EB estimates, conditional EB risks, overall Bayes risk, and estimated regret
   risk for different sample sizes.                                                          */
for sample size n in {20,40,60,80} do
    Generate n independent copies {X_i, Z_i, Θ_i}, where Θ is Uni(0.5,0.8) and given Θ, X follows GPD
      and Z follows Poi(3)                                  /* Setup simulation environment */
    Produce m = 10 sets of the n triples                    /* For each sample size */
    for j = 1 to 10 do
        Calculate EB estimates θ_n^(j)(x) for x = 5,6,...,25   /* Compute EB estimates for each set */
    end
    for j = 1 to 10 do
        Calculate conditional EB risk R̃(θ, θ_n^(j)) using the given formula for x = 5,6,...,25 /* Compute
          conditional EB risks */
    end
    Estimate overall Bayes risk R̂(θ, θ_n) as the average of the 10 conditional EB risks  /* Aggregate to
      overall Bayes risk */
    Calculate estimated regret risk Ŝ(θ_n) as R̂(θ, θ_n) − R(θ, θ_G)          /* Compute regret risk */
    Print estimated overall Bayes risk and estimated regret risk for current n     /* Output results for
      current sample size */
end
```

ALGORITHM 4: Empirical Bayes Estimator θ_n

```
Input: n, τ, x_min, x_max, m, seeds
Output: Results matrix results_n with EB estimates for each x and realization j
for x ← x_min to x_max do
    for j ← 1 to m do
        Set seed to seeds[j]
        Z^(j) ← Generate Poisson distributed values with mean τ
        θ^(j) ← Generate uniform values between 0.5 and 0.8
        Initialize X^(j) as numeric vector of length n
        for i ← 1 to n do
            X_i^(j) ← Sample from distribution with parameters θ_i^(j) and τ
            // Assuming rbort is a placeholder for the actual distribution sampling function
            Define c1, cZx, and cZX functions with appropriate error handling
            c1(x, X_i^(j)) ← Compute based on X_i^(j) and x
            cZx(x, Z_i^(j)) ← Compute based on Z_i^(j) and x
            cZX(X_i^(j), Z_i^(j)) ← Compute based on X_i^(j) and Z_i^(j)
        end
        // Compute ψ_nj and qnj for each x and aggregate
        ψ_n(x) ← Sum of ψ_nj(x, X^(j), Z^(j)) over i divided by n
        qn(x) ← Sum of qnj(x, X^(j), Z^(j)) over i divided by n
        // Calculate θ_n(x) using aggregated ψ_n and qn
        θ_n(x) ← Compute EB estimate from ψ_n(x) and qn(x)
        Store θ_n(x) in results matrix results_n[j, as.character(x)]
    end
end
// Output the results matrix
Print results_n
```

ALGORITHM 5: Monotonized EB Estimator θ_n^*

for *j in 1:m* **do**
  for *i in 1:na* **do**
    for *x in 1:xmax* **do**
      if $\theta_n^{(j)}(x) < a_i$ **then**
        $\alpha^{(j)}(a_i) = \alpha^{(j)}(a_i) + \sum_{i=1}^{na} p_r(x \mid a_i)$ /* Construct D and calculate α from (??) */
      end
    end
  end
end
Initiate $F_{x_{max} \times na}(x \mid a_i)$ as zero matrix /* Construct BT cdf */
for *i in 1:na* **do**
  $F(r \mid a_i) = p_r(r \mid a_i)$
  for *x in r+1:xmax* **do**
    $F(x \mid a_i) = F(x-1 \mid a_i) + p_r(x \mid a_i))$
  end
end
j=1 /* Construct D^* from (??) */
while *j<=m* **do**
  for *i in 1:na* **do**
    if $\alpha^{(j)}(a_i) > F(r \mid a_i))$ **then**
      /* case: $x = r$ */
      $D^{*(j)}(a_i \mid r) = 1$
    else
      $D^{*(j)}(a_i \mid r) = \frac{\alpha^{(j)}(a_i)}{F(r \mid a_i)}$
    end
    /* case: $x > r$ */
    for *x in r+1:xmax* **do**
      if $F(x-1 \mid a_i) > \alpha^{(j)}(a_i)$ **then**
        $D^{*(j)}(a_i \mid x) = 0$
      else
        if $F(x \mid a_i) < \alpha^{(j)}(a_i)$ **then**
          $D^{*(j)}(a_i \mid x) = 1$
        else
          $D^{*(j)}(a_i \mid x) = \frac{\alpha^{(j)}(a_i) - F(x-1 \mid a_i)}{F(x \mid a_i) - F(x-1 \mid a_i)}$
        end
      end
    end
  end
  x=r /* Construct θ_n^* from (3.4) */
  while *x<=xmax* **do**
    for *i in 1:na* **do**
      $tail_i(x) = 1 - D^{*(j)}(a_i \mid x)$
      $\theta_n^{*(j)}(x) = \frac{1}{na} \sum_{i=1}^{na} tail(x)$
    end
    x=x+1 /* Update of current outbreak size x */
  end
  j=j+1 /* Update of data set j */
end

ALGORITHM 6: Monotonized EB Estimator θ_n^{**}

Input: Results matrix *results_n*, vector x
Output: Isotonic regression-adjusted results matrix *iso_results_n*
// Initialize matrix to store isotonic regression results
Initialize *iso_results_n* as a matrix with the same dimensions as *results_n*
// Loop through each set of estimates
for $j \leftarrow 1$ ***to*** m **do**
    // Apply isotonic regression to the j-th set of estimates
    $iso\_fit \leftarrow$ Apply isotonic regression to x and $results\_n[j,]$
    // Store the fitted values
    $iso\_results\_n[j,] \leftarrow iso\_fit$'s fitted values
end
// Output the isotonic regression-adjusted results
Print *iso_results_n*

VITA

Alberta Araba Johnson is a first-generation college graduate, who has demonstrated excellence in mathematical biology, statistics, and data science. She completed her Master's degree in Applied Statistics and Data Science at the University of Texas Rio Grande Valley (UTRGV) in May 2024, where she acquired in-depth skills in modeling, simulations, and data analysis. Alberta was awarded the College of Science Dean's Fellowship from August 2022 to May 2024. Prior to that, Alberta earned a Bachelor of Science in Actuarial Science degree from the University of Cape Coast (UCC), Ghana, where she emerged as the Best Graduating Female Student in the Actuarial Science program.

Alberta's research interests encompass survival analysis, Bayesian methods, disease modeling, and machine learning. She looks forward to extending her expertise to making meaningful contributions to public health research.

Feel free to send her an email at the following address: arabaa.johnson@gmail.com