Methods for order restricted estimation for length biased data, with applications to crime scene prior probability estimation

Julia Wierzchoslawska

Master's thesis
2026:E15

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Methods for order restricted estimation for length biased data, with applications to crime scene prior probability estimation

Lund University

Julia Wierzchoslawska

February 27, 2026

Acknowledgements

I would like to thank my supervisor, Dragi Anevski, for guiding me through this project and making it possible. I am also grateful for the use of the real data provided by Christian Dahlman.

Thank you Wilmer, my love, for your endless support in what has been a seemingly never-ending journey.

For my Lucy and Tony.

Abstract

The objective of this thesis is to expand on previous research by implementing a new method for order-restricted estimation of length-biased data. This paper addresses the implementation and analysis of a stochastic spatial probabilistic model for characterizing the geometric relationship between perpetrator addresses and crime scene locations. The given data are length-biased observations of a decreasing density function. The approach addresses the length-bias problem using David Cox's estimator, a known solution, while handling the order restriction through the Grenander estimator, using estimation under monotonicity. To evaluate this new method for estimating prior probabilities from crime scenes, the following analysis was conducted. Simulations were performed in which length-biased order-restricted samples were processed through the new method and then compared against both a simulated ideal estimated distribution (non-length-biased) and a theoretical ideal distribution. Following this, the new method was applied directly to the real data, where we obtained a density estimate. This density and method could be used to assign prior probabilities to suspects in real-world criminal cases.

Contents

Introduction

This paper implements methods for addressing a problem presented in the two manuscripts [3] and [1]. As indicated by the title, we will be implementing methods for order-restricted estimation for length-biased data, with application to crime scene prior probability estimation. Related work includes [5] and [6], which investigate estimation of monotone densities from biased samples. Paper [6] includes simulation-based analysis, making it closely related in approach to the present work.

Currently, there are no complex models available for crime scene prior probability estimation, as described in the introduction of [3]. There is a need for a more comprehensive method to assign prior probabilities to suspects in criminal cases. The referenced manuscripts address this problem by proposing a new method, which is partially implemented in this paper.

The data used in this analysis was provided by Christian Dahlman. It consists of 100 murder crime observations with variables including the euclidean distance from victim address to perpetrator address. Additional variables are present in the data, only the distance variable is utilized in this analysis.

A central topic that we will cover is the length-biased sampling problem, which occurs when the likelihood of selecting observations is proportional to their length. Our data consists of order-restricted length-biased data, the inference problem is to recover the true uncorrupted density function f and its corresponding cumulative distribution function F. David Cox [2] was apparently the first to study estimation methods for PDF and CDF functions under length-biased sampling. The theoretical background of this problem is looked at in Chapters 2.1, 2.2.2, and 2.2.3.
The order-restricted part of this problem has been examined in [6] and [5] This requires accounting for the fact that the function of interest f is monotonically decreasing. Without this assumption, the problem would be the standard length-biased sampling problem.

The introduced method by [3] and [1] goes through the following steps: First, the length-biased data undergoes Cox's transform to obtain an empirical estimator F_n. To address the order-restriction constraint, f decreasing means that F must be concave, to ensure this criteria one must take the least concave majorant of F_n and with motivation shown in Chapters 2.2.6, 2.3.2 and in the manuscripts to obatain our estimator f_n one takes the isotonic regression of the emperical data. In application, these steps for dealing with the order-restriction problem are implemented using the Grenander estimator in R, which performs these operations.

The objective of this thesis is to implement a simulation algorithm that performs 10000 iterations of the following process: generate order-restricted length-biased data, apply Cox's transformation, and process the results through the Grenander estimator. Then to evaluate the performance of this estimator we compare it against an ideal estimate and a theoretical distribution. The second component involves applying this new method to real crime data and analyzing the resulting estimates.

We begin this paper with Chapter 2.1, which provides background and the theoretical framework from the two manuscripts [3] and [1], where the problem first presented. Chapter 2.2 covers the theoretical foundations needed for understanding this work, inlcuding the inhomogeneous spatial Poisson process, length-biased data, Cox's empirical estimate, the Grenander estimator, and performance evaluation metrics used in the analysis of this thesis. We proceed in Chapter 2.3 with the methods of estimation for $\hat{f}(0)$, which provides an a priori probability for suspects located at a distance zero from the crime scene, and $\hat{f}(r)$, giving an a priori probability for suspects at distance r from the crime scene. Chapter 2.4 covers the implementation methodology for obtaining these estimates using R, including details of the algorithms used. Chapter 3.1 presents and analyzes the simulation results, while Chapter 3.2 applies the new method to the real data, and we examaine the findings. Finally, Chapter 4 discusses the overall results and concludes with suggestions for future work.

Theory

In the first section of this chapter we will overview a manuscript by Anevski and D.Gill, cf.[3], the set up for estimating a monotone density function under length biased sampling and of spatial models that have given rise to this estimation problem. In Chapter 2.2 we will give the theory needed to understand the framework of this project. In Chapter 2.3 the method of estimation will be covered, along with the selection and validation methods.

2.1 Background

Here we will take a deeper look into the manuscripts [3] and [1], these papers introduce the problem of order restricted estimation of a density for length biased data. Furthermore, in their work regarding assigning prior probability in a criminal case, [3] and [1] suggest a new model for the spatial distribution of inhabitants as well as the spatial distribution of possible perpetrators of a given area.

Let us begin with the length biased sampling problem. Suppose we have i.i.d. data $x_1, ..., x_n$ from a density g on $\mathbb{R}_+$ and

$$g(x) = cxf(x), \tag{2.1}$$

where f is a density on $\mathbb{R}_+$ and $c = (\int xf(x))^{-1}$ is the normalizing constant. There are two inference problems arising from this, firstly, estimating f, and secondly, estimating the cumulative distribution function $F(x) = \int_0^x f(u)du$, which is directly related to the estimation of f.

David Cox [2] was apparently the first to have studied the nonparametric estimation of F, he suggested an empirical estimator which we elaborate on in section 2.2.3. Following this, Vardi [8] showed that Cox's empirical estimator was in fact the nonparametric maximum likelihood estimator(npml) of F. Next, Jones [4] suggested a kernel estimator for the density where he adapts Cox's empirical estimator to the kernel estimator approach. Jones proposes two approaches which are covered in 2.2.4.

Our goal is to estimate the density f. Suppose f is a density function which is decreasing on $\mathbb{R}_+$, from which we have length biased observations. The need to estimate this order restricted density arises in the forensic science problem which we will overview next.

2.1.1 The A Priori Probability

We begin with describing the models for the spatial distribution of inhabitants/population as well as the spatial distribution of possible perpetrators of a given geographical area $B \subset \mathbb{R}^2$.
$N_0(A)$ is a deterministic point process modeling the general population. $N_1(A)$ is a non-homogeneous Poisson process, which models the population of possible perpetrators. The processes are point processes in the plane, where $N_0(A)$ is equal to the number of inhabitants in the subset $A \subset \mathbb{R}^2$ and $N_1(A)$ is equal to the number of possible perpetrators in A. Both N_0 and N_1 are centered so that each crime scene is at the origin, making the populations addresses coordinates to each crime scene.

A model assumption for N_0 is that the intensity is constant and equal to λ_0 in B, and for a specific area A, $N_0(A)$ is equal to $\lambda_0|A|$. where $|A|$ is the area of A.

$N_1(A)$, the number of possible perpetrators in area A is modeled as an inhomogeneous Poisson process with a local intensity $\lambda_1(x, y)$. The process will be formally introduced in section 2.2.1. The model assumption for λ_1, the intensity process, is that it is a function of only the distance between the crime scene and the possible perpetrator address.

$$\lambda_1(x, y) = f((x^2 + y^2)^{\frac{1}{2}}) \tag{2.2}$$

for some function f, in which we have introduced the polar coordinates $r = (x^2 + y^2)^{\frac{1}{2}}, \theta = arctan(\frac{y}{x})$.

Assumption 1 : The possible perpetrator process N_1 is an inhomogeous Poisson process with intensity process $\lambda_1(x, y) = f((x^2 + y^2)^{\frac{1}{2}})$

Assumption 2 : The function $f : \mathbb{R} \to \mathbb{R}_+$ is decreasing.

Note, the second assumption is that f lies in the set $\mathcal{F}$ of positive functions on $[0, \infty)$ that are decreasing. Furthermore, by the first model assumption the area B must be circular, otherwise λ would not only depend on the distance.

Here the model for N_1 is modified, so that when studying different geographical areas the findings can be generalized to arbitrary geographic areas. This is done by writing the model for the perpetrator in multiplicative form, where each N_1 intensity is multiplied by its specific N_0 population intensity. The argument behind this model is that it's fair to believe that the perpetrator intensity is locally proportional to the population intensity.
The argument for the multiplicative model 2.3 is that it's reasonable to think the level of crime is directly related to the population size in each area, and that the same ratio (described by f) applies to all areas.

$$\begin{aligned} \lambda_1(x, y) &= \lambda_0(x, y) f\left((x^2 + y^2)^{\frac{1}{2}}\right) \\ &= \lambda_0(x, y) f(r) \\ &= \lambda_0 f(r) \end{aligned} \tag{2.3}$$

2.1.2 Probabilistic Model for the Data

We present an overview of the data descriptions from Anevski's two manuscripts [1, 3]. There are n geographical areas and n crimes with a guilty verdict. For each geographical area i, there are observations of the total population m_i, of the area A_i, that being, one crime that has been committed, the location of the crime scene, and the address of the perpetrator.

Once the centering of the crime scene is done we have the following point processes. Process $N_0^{(i)}$ has constant population intensity $\lambda_0^{(i)}$ and $N_1^{(i)}$ has varying perpetrator intensities $\lambda_1^{(i)} = \lambda_0^{(i)}(x,y)f\left((x^2+y^2)^{\frac{1}{2}}\right)$. We note that $f(r)$, the local probability of a chosen individual at a distance r to be a perpetrator is equal for all geographical areas i. The data can be described by a 2n point process in the plane $N_0^1, ..., N_0^n, N_1^1, ..., N_1^n$ and summarized by the sums,

$$\begin{aligned} N_0 &= \sum_{i=1}^{n} N_0^{(i)} \\ N_1 &= \sum_{i=1}^{n} N_1^{(i)}. \end{aligned} \tag{2.4}$$

We note again that N_1 is a deterministic process, and since we assume $N_1^{(i)}$ are independent Poisson then the total process N_1 is a Poisson process. Additionally we have that N_1 from which we have observations $r_1, ..., r_n$ has intensity

$$\lambda(r) = \sum_{i=1}^{n} \lambda_0^i f(r). \tag{2.5}$$

We note, $\bar{\lambda}(r) = \sum_{i=1}^{n} \lambda^{(i)}$. To make the stochastic process only depend on r, let's say $A \subset B$ is an arbitrary area, and λ denotes the intensity for the process, N_1, then the following is the cumulative intensity of N_1, indexed by sets.

$$\begin{aligned} \Lambda(A) &= \int_A \lambda(x,y)dxdy \\ &= \int_A \bar{\lambda}_0 f((x^2+y^2)^{\frac{1}{2}})dxdy. \end{aligned} \tag{2.6}$$

Now suppose that $A_r = \{(x,y) \in \mathbb{R}^2 : ||(x,y)|| \leq r\} \in B$ is a circular area centered at the origin with radius r. A change to polar coordinates in the following is performed,

$$\begin{aligned} \Lambda(A_r) &= \int_{A_r} \bar{\lambda}_0 f((x^2+y^2)^{\frac{1}{2}})dxdy \\ &= \int_0^r \int_0^{2\pi} \lambda_0 f(u)ud\theta du \\ &= 2\pi\hat{\lambda}_0 \int_0^r uf(u)du \\ &=: \tilde{\Lambda}(r). \end{aligned} \tag{2.7}$$

The last equality defines a function $\tilde{\Lambda}(r) \in \mathbb{R}_+$.

Lemma 1 shows that the process

$$\tilde{N}_1 := N_1(A_r) \tag{2.8}$$

is a stochastic process defined on $[0, \infty)$, with independent increments that are Poisson distributed and that the increment $\tilde{N}_1(r_2) - \tilde{N}_1(r_1)$ is Poisson distributed with expectation

$$E(\tilde{N}_1(r_2) - \tilde{N}_1(r_1)) = 2\pi\bar{\lambda}\int_{r1}^{r2} f(u)udu \tag{2.9}$$

Thus $\tilde{N}_1$ is a non-homogeneous Poisson process with intensity equal to

$$\tilde{\lambda}(r) = 2\pi\tilde{\lambda}_0 f(r)r. \tag{2.10}$$

From the above result, we have that our observations $r_1, .., r_n$ representing the jump times in a non-homogeneous Poisson process, conditional on the intensity $\Lambda(R) = n$ where R is our radius encompassing the area we are studying. By applying Cox and Lewis (1962) lemma we have that:

$$(r_1, r_2, ..., r_n|\Lambda(R) = n) =^d x_{(1)}, x_{(2)}..., x_{(n)} \tag{2.11}$$

Where $x_{(1)}, x_{(2)}..., x_{(n)}$ is the order statistic of $x_1, x_2, ..., x_n$ which are independaent and identically distributed with density $\hat{\lambda}(r) = 2\pi\bar{\lambda}_0 r f(r)$ This lemma allows us to view our data $r_1, ..., r_n$ as a sample from the density $\hat{\lambda}(r) = 2\pi\bar{\lambda}_0 r f(r)$. Note: f represents the unknown density function of interest, where the sampling situation in our application is identical to length-biased sampling from a density proportional to f.

Note: The lemma which characterizes the process $\tilde{N}_1$ indexed by the length measure r is given below and proven in [1].

Lemma 1

Suppose that N_1 is an inhomogeneous Poisson process in the plane with intensity $\lambda(x, y) = h(r)$, with $r = (x^2 + y^2)^{\frac{1}{2}}$. Let A_r be the circle in the plane with radius r and centre at the origin, and let $\tilde{N}_1(r) = N_1(A_r)$ be the stochastic process defined via N_1, on $[0, \infty)$. Then $\tilde{N}_1$ is an inhomogeneous Poisson process with intensity $2\pi h(r)r$.

2.1.3 Distribution Model for the Perpetrator Distance

The possible perpetrator priori distribution is modeled as a mixed distribution, it is discrete at zero and continuously decreasing at all other distances r. The crimes committed at home i.e. distance between the crime scene address and the perpetrator address is zero i.e. crimes committed at the perpetrators home are discrete. All crimes committed outside the home i.e. the perpetrator's home address is greater than zero from the crime scene address are modeled as a continuous decreasing function of the distance between the perpetrators addresses and the crime scene

addresses. Let f denote this mixed function of interest.

For the analysis of f we will break it up into two parts, the discrete part $f(0)$ and the continuous, $f(r)$.

To model the probability $f(0)$ of a perpetrator committing a crime at their home address, we calculate the ratio of home-based crimes to total crimes in the dataset: $n_{\text{home}}/n_{\text{total}}$. If additional data on the number of household members were available, this probability could be further modified by adjusting for the number of household members: $\frac{n_{\text{home}}}{/n_{\text{total}}} \cdot \frac{1}{m_{\text{household}}}$.

For $f(r)$, the possible area is treated as circular with a certain radius centered at the crime scene address. The data being used for this scenario includes r_i, the distance between the crime scene and the perpetrator address, R_i, the radius of the circular area that contains both the perpetrator address and crime scene address, and the population size, giving the population density. We will use the same R for the analysis, where R is the country radius.

For the estimation of $f(r)$ we will use the proposed method given in the manuscript [3], and introduced in Chapter 2.3.2.

2.2 Theory

2.2.1 Poisson Process

The Poisson Distribution

A Discrete random variable X is said to be Poisson distributed $X \sim Poisson(\lambda)$ with intensity parameter λ if its p.m.f is given by

$$f_x(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{2.12}$$

where k denotes the number of occurrences.

The Poisson Process in Time

Say we have a point process in time $\{N(t), t \in [0, \infty)\}$, where $B \in \mathbb{R}^2$, N(t) is called a Poisson process with intensity λ if the following conditions hold: N(0) = 0, N(t) has independent increments, and the number of arrivals at any interval length τ has $Poisson(\tau\lambda)$ distribution.

The Homogeneous Poisson Process in Space

We define a spatial point process $\{N(B), B \in \mathcal{B}(\mathbb{R}^2)\}$ on the plane, where $\mathcal{B}(\mathbb{R}^2)$ denotes the Borel sets in $\mathbb{R}^2$. For any Borel measurable set B, $N(B)$ represents the number of points (events) occurring within region B. The process N is called a homogeneous Poisson process in the plane with intensity $\lambda > 0$. For any bounded Borel set B, we have:

$$P(N(B) = n) = \frac{(\lambda|B|)^n}{n!}e^{-\lambda|B|} \tag{2.13}$$

where $|B|$ denotes the area of B, and if $N(B_1), N(B_2)$ are independant when B_1, B_2 are disjoint sets.

The Inhomogeneous Spatial Poisson Process

The inhomogeneous spatial Poisson Process is a Poisson point process where the intensity rate, the Poisson parameter is defined by a location dependent function. Rather than having a constant rate, the intensity rate varies across the space where the process is defined. For simplicity's sake we will only treat 2-dimentional space. The intensity function $\lambda : \mathbb{R}^2 \to [0.\infty)$ is dependent on the point $r = (x, y) \in \mathbb{R}^2$, the function $\lambda(r)$ gives the intensity rate at each point r in the plane.

For any bounded Borel measurable region $B \subset \mathbb{R}^2$, the expected number of points in B is given by the integral of λ over the region B.

$$\Lambda(B) = \int_B \lambda(x, y)dxdy \tag{2.14}$$

Now let's say for any collection of disjoint bounded Borel measurable sets $B_1, B_2, ..B_n \subset B$, an inhomogeneous Poisson process with intensity $\lambda(r)$ has the following distributions for the number of events in each region

$$P(N(B_i) = n_i, i = 1, .., k) = \prod_{i=1}^{k} \frac{(\Lambda(B_i))^{n_i}}{n_i!}e^{-\Lambda(B_i)} \tag{2.15}$$

$\Lambda(B_i)$ is the expected number of points in region B_i and is calculated by integrating $\lambda(x, y)$ over the area B_i. Additionally, $\Lambda(B)$ can be interpreted as being the expected number of points of the Poisson process located in the bounded area B, that is

$$\Lambda(B) = E(N(B)) \tag{2.16}$$

Where N(B) describes how many events occur within B.

2.2.2 Length-Biased Data

Suppose that f is an unknown probability density function modeling the entity of interest. However, we do not have direct observations from f. Instead, we have length-biased observations $x_1, ..., x_n$, which are i.i.d. observations from the density

$$g(x) = \frac{xf(x)}{\mu}, \tag{2.17}$$

where μ is the mean of $f(x)$ and serves as a normalizing constant such that g integrates to one.

Length-bias sampling occurs when the probability of an observation being in a sample is proportional to the length of the observation itself. By the multiplication of x and $f(x)$, the probability of an observation being in the sample is directly related to

the size of x. The function $g(x)$ is referred to as the weighted or length-biased probability density function, whereas $f(x)$ is called the unweighted probability density function.

2.2.3 Cox's Estimator

There are two fundamental inference problems when working with length-biased data, one is estimating the density $f(x)$ and the second is estimating the corresponding CDF $F(x) = \int_0^x f(u)du$. David Cox, [2], was apparently the first to study the nonparametric estimation of F based on length-biased data. He observed that in length-biased sampling, the likelihood of selecting longer observations is proportional to their length. His approach is based on this key observation.

Say we have the observations r_i which come from the length-biased distribution $g(x) = \frac{xf(x)}{\mu}$, where $\mu = E(X)$ is the mean of the original distribution.
Cox found that the empirical data assignes weights $\frac{1}{n}$ to each observation point r_i which then gives weights $\frac{1}{n}$ to $g(r_i) = \frac{r_i f(r_i)}{\mu}$. To correct the length-biased effect, Cox proposed assigning weights proportional to $\frac{1}{nr_i}$ to each observation to recover the original density $f(r_i)$.

Based on the above, Cox proposed the following empirical estimator for the CDF:

$$F_n(x) = \frac{1}{n\hat{\mu}} \sum_{i=1}^{n} \frac{1}{r_i} 1\{r_i \leq x\}. \tag{2.18}$$

The estimate $\hat{\mu}$ is used as a normalizing constant and has been shown by Cox to be

$$\hat{\mu} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{r_i} \right)^{-1}. \tag{2.19}$$

2.2.4 Jones's Estimator

Jones [4] proposed two distinct methods of estimation using kernel estimators that build upon Cox's empirical estimator approach.

The Direct approach: Jones' first estimator directly estimates the density f by correcting length bias as in Cox's approach.

$$\hat{f}(t) = \frac{1}{nh\hat{\mu}} \sum_{i=1}^{n} \frac{1}{x_i} k(\frac{t - x_i}{h}), \tag{2.20}$$

where $\hat{\mu}$ is defined in the section above as Cox's normalizing constant, t is the point at which we would like to estimate the density, $h > 0$ is the bandwidth parameter that controls the smoothness, and $k(\cdot)$, the kernel function which determines the shape around each data point. In this approach, like in Cox's, $\frac{1}{x_i}$ is incorporated to counteract the length-biased effect.

The two-step approach: Jones' second estimator begins by first estimating the length biased density g with a standard kernel estimator and then using the relationship $g(x) = xf(x)/\mu$ to obtain the original density f, as shown;

$$\hat{f}(t) = \frac{1}{tnh\hat{\mu}} \sum_{i=1}^{n} k(\frac{t - x_i}{h}). \tag{2.21}$$

The $1/t$ factor provides additional weighting after the kernel smoothing to transform the length-biased density estimate to the original form.

2.2.5 Linear Isotonic Regression

Linear Isotonic regression allows for fitting a piece-wise linear model with monotonic restraint. It comes in handy when the relationship between variables is believed to be monotonic, but the exact form is unknown. Given the obsered data points $(x_1, y_1, w_1), ..., (x_n, y_n, w_n)$, where $x_1 \leq x_2, ..., \leq x_n$ and $w_i > 0$, the isotonic regression problem aims to find estimates $\hat{y}_1, .., \hat{y}_n$ that minimizes the weighted sum of squared deviations,

$$min \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2. \tag{2.22}$$

Since we are specifically looking at the case with a monotonically decreasing constraint, for our minimization above, we have that $\hat{y}_1 \geq \hat{y}_2, ..., \geq \hat{y}_n$.

2.2.6 The Grenander Estimator

The Grenander estimator is used to estimate a non-parametric, non-increasing density function under monotonicity constraints. Grenander showed that this estimator is in fact the non-parametric maximum likelihood estimator (NPMLE) under the constraint that the density is non-increasing, as reviewed in [7].

Say we have independent observations for our data $X_1, ...X_n$ and we believe that our data comes from a density $f(x)$ that is decreasing, i.e. $x_1 < x_2$, then $f(x_1) \geq f(x_2)$ our goal is to then estimate $f(x)$ without assuming a specific parametric form, while keeping the monotonicity constraint.

The problem can be formulated as finding the non-parametric maximum likelihood estimator under the monotonicity constraint.

So given the observations/order statistic $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ we want to estimate f by solving,

$$\hat{f}_n = argmax_{g \in \mathcal{F}} \prod_{i=1}^{n} g(X_i) \tag{2.23}$$

where $\mathcal{F}$ is the set of non-increasing densities on R_+. We can equivalently maximize $\hat{f}_n = argmax_{g \in \mathcal{F}} \sum_{i=1}^{n} log(g(X_i))$.

To compute this we begin by obtaining the empirical distribution function from our order statistic,

$$\hat{F}_n = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\} \tag{2.24}$$

$\hat{F}_n$ is a step function that jumps by $1/n$ at each observed data point X_i.

Since f is decreasing, F must be concave. The maximum likelihood principle then selects the concave function that best fits the data while remaining above the empirical CDF. We can take the least concave majorant $S(F_n)$ of F_n. The least concave majorant is the smallest possible concave function that lies above $\hat{F}_n$ at all points.

$$S(F) = inf\{G : G \geq F,\ G \text{ concave}\} \tag{2.25}$$

The Grenander estimator is then obtained by taking the left derivative of $S(\hat{F}_n)$,

$$\hat{f}_n = \frac{d}{dt} S(\hat{F}_n) \tag{2.26}$$

An important connection exists between shape-constrained estimation and isotonic regression, where the goal is to find he best monotone approximation to observed data. The least concave majorant of an empirical CDF can be obtained through isotonic regression applied to the empirical data. This result, which we state below, has been established in various contexts including length-biased data [1].

Theorem 1 *The estimator $\hat{f}_n$ defined in (2.26) is the isotonic regression of the empirical data*

$$\hat{f}_n = argmin_G \sum_{i=1}^{n} (G(X_i) - F_n(X_i))^2 \tag{2.27}$$

where G is non-decreasing, concave, and where $G(x) \geq F_n(x)$ for all x.

2.2.7 The MISE Test

The Mean Integrated Squared Error (MISE) test is used to determine the global goodness of fit of an estimated probability density function. It assesses the difference between an estimated density and the true underlying density function.

For an estimated density $\hat{f}(x)$ and the true density function $f(x)$, the MISE is defined by:

$$MISE(\hat{f}) = E\left[\int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 dx\right], \tag{2.28}$$

where E is the average value that would be taken across all possible samples.

The MISE is only defined when both densities belong to L_2, meaning $\int f^2(x)dx < \infty$ and $\int \hat{f}^2(x)dx < \infty$. For densities that do not fall into this criterion of belonging to L_2, alternative methods for measuring distance are needed. The best possible outcome for a MISE value is represented by a 0, meaning a perfect estimation. MISE values are always non-negative, and a lower value indicates that you have a good performing estimator. However, the magnitude of the result depends on the size of the data set and the specific problem being addressed.

2.2.8 The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) test is a nonparametric test used to determine if a sample comes from a specific theoretical distribution, or to determine if two samples come from the same underlying distribution. We will only consider the single-sample case, where we test the goodness of fit of a sample to a theoretical distribution. The K-S test assesses the maximum difference between empirical distribution functions.

Let $F_n(x)$ be the empirical distribution for a sample of size n, and let $F(x)$ be the theoretical cumulative distribution function. The K-S statistic is defined as:

$$D_n = \sup_x |F_n(x) - F(x)| \tag{2.29}$$

where $\sup_x$ is the supremum of the set of distances. D_n, the K-S statistic, takes the largest absolute difference between the two distribution functions across all x values.

To interpret the results of D_n, the null hypothesis is used. H_0: the sample comes from the specified theoretical distribution vs H_1: the sample does not come from the specified theoretical distribution.

The K-S statistic D_n value ranges between 0 and 1, where a value of 0 indicates a perfect fit between the samples and the theoretical distribution, while 1 indicates a complete difference between the empirical and theoretical distributions.
D_n's interpretation depends on the chosen significance level α, with common choices being $\alpha = 0.05, 0.01$. D_n is used to calculate a p-value. If the p-value is less than α we reject H_0, and conclude that the sample does not follow the theoretical distribution.

2.3 Estimation Methods

In this chapter we will cover the estimation methods for the two different scenarios that have been briefly discussed in Chapter 2.1.2. Firstly, the estimation of $f(0)$, and secondly, of $f(r)$.

2.3.1 Estimation for Distance Zero

The mixture proportion $\hat{p}$ is a simple estimate calculating the fraction of crimes in the given data set that have occurred at distance zero (i.e., at the victim's residence), and is given by the following:

$$\hat{p} = \frac{\sum_{i=1}^{n} 1\{r_i = 0\}}{n}. \tag{2.30}$$

From our data set, we have 33 out of $n = 100$ crimes that occurred at the victim's residence, so $\hat{p} = 33/100 = 0.33$.

To calculate the a priori probability that a specific household member is the perpetrator, one would need additional data: the number of people residing in the victim's household.
We assume each household member has an equal chance of being the perpetrator, we

divide the overall probability by the number of potential suspects in the household. Here is an example: Say a murder occurs in a household of 4 people (not including the victim); each individual in that household would have an a priori probability of being the perpetrator equal to $1/4 \cdot 0.33 = 0.0825$.

2.3.2 Estimation for Distances Greater than Zero

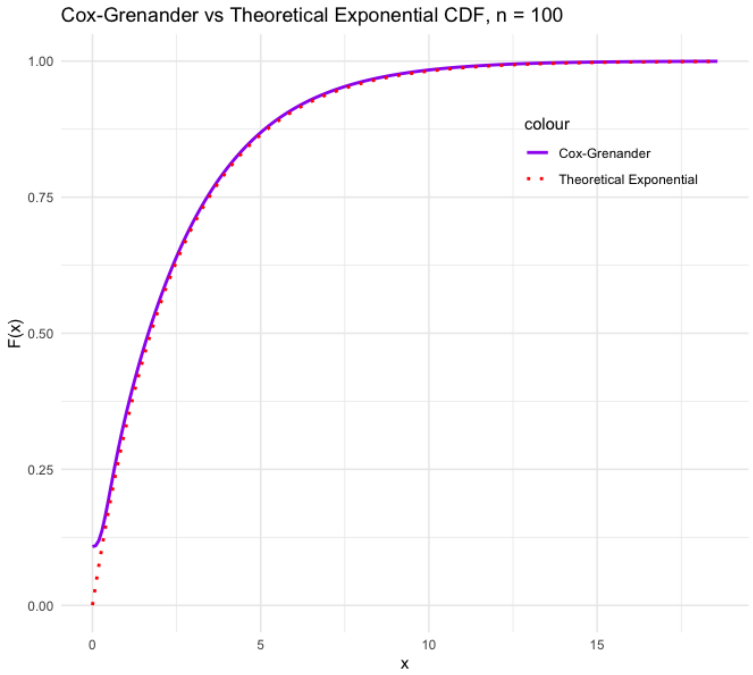

Figure 2.1: Plot of average Cox-Grenander cdf vs Theoretical Exponential cdf for 10000 simulations of 100 data points.

The data that we have can be viewed as follows, $r_1, ...r_n$ are i.i.d. from the density g on $\mathbb{R}_+$, where $g(x) = cxf(x)$, here f is a density on $\mathbb{R}_+$ and $c = (\int xf(x))^{-1}$ is the normalizing constant. The goal is to estimate the unkonwn density f based on the data $r_1, ..., r_n$ under the assumption that f is decreasing. The estimation of f without having an order restriction upon f is known as the standard length biased sampling problem which has been studied. The problem that we are looking at has only been studies by [1] and is called the order restricted length biased sampling problem. Let us begin by the standard length biased sampling problem where we estimate f by estimating its corresponding cumulative distribution function $F(x) = \int_0^x f(u)du$. The first to have studied the estimation of F is David Cox who suggested the following empirical estimator F_n of F as,

$$F_n(x) = \frac{1}{n\hat{\mu}} \sum_{i=1}^{n} \frac{1}{r_i} 1\{r_i \leq x\}. \tag{2.31}$$

The estimate $\hat{\mu}$ is used as a normalizing constant and has been shown by Cox to be $\hat{\mu} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{r_i}\right)^{-1}$. This estimator is obtained by giving each observation, point r_i, the emperical weight $1/n$, it then follows that the emperical weight $1/n$ should be given to the estimator of $g(r_i) = cr_if(r_i)$ so that they should give an emperical

estimate weight proportional to $1/nr_i$ to the estimate of $f(r_i)$. Next, [8] showed that Cox's F_n is in fact the nonparametric maximum likelihood estimator of F.

Cox's emperical estimator F_n satisfies the standard length biased sampling problem but it does not satisfy the order restricted length biased sampling problem where we assume F emerges from. In short, f decreasing implies that F must be concave. A modification that can be done to ensure that F be a concave distribution function is to take the least concave majorant $S(F_n)$ of F_n. [1] have proposed an estimator of f, which is the derivative of $S(F_n)$,

$$\hat{f}_n(t) = \frac{d}{dt} S(F_n)(t). \tag{2.32}$$

Here S is the least concave majorant defined as $S(z) = inf\{u : u \geq z, u \text{ concave}\}$, and $\frac{d}{dt}$ denotes the left hand derivative. This will result in $\hat{f}_n$ to be left-continuous. As mentioned in Chapter 2.2.6, the least concave majorant of an empirical CDF can be obtained through isotonic regression applied to the empirical data, and is shown below as in [1].

Theorem *The estimator f_n defined in 2.29 is the isotonic regression of the empirical data $\{(r_i, 1/nr_i)\}_{i=1}^n$,*

$$(f_n(r_1), ..., f_n(r_n)) = argmin_{\theta \in \Theta} \sum_{i=1}^{n} \left(\theta_i - \frac{1}{n\hat{\mu}(r_n - r_{n-1})} \right)^2, \tag{2.33}$$

where $\Theta = \{\theta \in \mathbb{R}^n : \theta_1 \geq ... \geq \theta_n\}$.

In practise to assign a priori probability to a suspect at a certain distance zero one would take the distance of the suspect at question, find that distance on the pdf plot made by the model

2.4 Implementation Method

In this chapter we will overview the structure of code used in R to implement the estimation methods for $f(0)$ and $f(r)$, which were theoretically described in sections 2.3.1 and 2.3.2.

2.4.1 Implementation in R

Implemenation of f(0)

Since our data from previous murder cases has limited information, we do not know the number of household members for each crime that occurred at the victim's residence. So, in our case, we have to simply assign the same a priori probability to all suspects at distance zero. To do so, we calculated the proportion of crimes in our dataset that occurred at distance zero (the victim's residence), and found that 33 out of 100 crimes occurred at this location, giving us: $\hat{p} = \sum_{i=1}^{n} 1\{r_i = 0\}/n = 33/100 = 0.33$.

As a result, any suspect at distance zero receives an a priori probability of 0.33 based on our data.

If we had the additional information, the number of people residing in each victim's household, we could adjust this approach. We would assume that each household member has an equal chance of being the perpetrator. The a priori probability for any specific household member would then be calculated by:

$$\hat{p}_{individual} = \frac{1}{\#\text{ of household members}} \cdot 0.33 \tag{2.34}$$

This would give a more comprehensive probability estimate to account for the set of potential suspects within each household. For both of these situations above, complex computing in R is not necessary.

Implementation of f(r)

The implementation of $f(r)$ in R consists of two main components. First, we simulate length-biased data to test and compare the performance of our new estimation method. Second, we apply the new estimation method to our existing length-biased dataset as one would in practice.

We begin with the first part.

Simulation Implementation in R

The analysis utilized several key packages: in2extRemes, Iso, fdrtool, ggplot2, dplyr, and ks. We set some simulation parameters to ensure reproducibility, a random seed was set at the beginning of all simulations. The parameters in the pseudo code below were used for this analysis. We will cover in this section our first case where our sample size is 100 and threshold is chosen to be at 0.9, the choice of this parameter is explained in Chapter 3.1. For the length-biased exponential distribution, we use the theoretical result that length-biased sampling from an exponential distribution is equivalent to sampling from a Gamma distribution with shape parameter 2, $\alpha = 2$. We show this in the following equation where $f_{LBE}(x)$ is the length-biased density function and $f_G(x)$ is the Gamma density with shape parameter 2. In fact we have

$$\begin{aligned} f_{LBE}(x) &= \frac{x f_E(x)}{E(X)} \\ &= \frac{x\lambda e^{-\lambda x}}{1/\lambda} \\ &= \lambda^2 x e^{-\lambda x} \end{aligned}$$

and

$$\begin{aligned} f_G(x) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{(\alpha-1)} e^{-\lambda x} \\ &= \frac{\lambda^2}{(2-1)!} x^{(2-1)} e^{-\lambda x} \\ &= \lambda^2 x e^{-\lambda x}. \end{aligned}$$

This relationship allows us to generate length-biased exponential observations by sampling from $Gamma(2, \lambda)$.

```
set.seed(123)
x_start <- 0.90
n_samples <- 100
n_simulations <- 10000
lambda <- 0.4
shape <- 2
exp_rate <- 0.4
```

Next, storage structures are initialized to save the results from each of the 10000 simulations. This includes matrices for storing probability density function and cumulative distribution function values at 200 evaluation points spanning the distribution range. Additionally, vectors are initialized for recording performance measures such as integrated squared errors and maximum distances between estimated and true distributions, which are used for calculating the MISE test and comparing estimated and theoretical distributions.

The simulation consists of a forloop that repeats the following 4-step process 10000 times:

STEP 1: Simulate Gamma samples, Cox transform, and prepare F_{ncox} for the Grenander estimator.

We begin by simulating Gamma samples and sorting them.

```
for (i in 1:n_simulations) {
gamma_samples <- rgamma(n_samples, shape = shape, rate = lambda)
gamma_samples <- sort(gamma_samples)
```

We define normalizing constant, and perform Cox's transform on Gamma samples.

```
mu <- sum(1/gamma_samples)/n_samples
Fn_cox <- cumsum(1/gamma_samples)/n_samples/mu
```

We make F_{ncox} an ecdf object to be able to use in the Grenander function.

```
x_values <- gamma_samples
Fn_cox_ecdf <- stepfun(x_values, c(0, Fn_cox))
class(Fn_cox_ecdf) <- c("ecdf", "stepfun", class(Fn_cox_ecdf))
attr(Fn_cox_ecdf, "call") <- call("ecdf", quote(gamma_samples))
```

We Apply Gernander estimator to F_{ncox}.

```
result_cox <- grenander(Fn_cox_ecdf, type="decreasing")
```

STEP 2: Obtain Grenander and Exponential CDFs, calculate distances.

We create 200 equally spaced evaluation points (0 to max value in Gamma data), make grid of x-values to evaluate CDFs for comparison, and remove evaluation points close to .

```
eval_points <- seq(0, max(gamma_samples), length.out = 200)
eval_points <- eval_points[eval_points > 0.001]
```

We obtain the Grenander estimators CDF values at evaluation points, linear interpolation between Grenander knot points (x-values and CDF values).

```
grenander_cdf <- approx(result_cox$x.knots, result_cox$F.knots,
    xout = eval_points, method = "linear", rule = 2)$y
```

We calculate the exponential CDF, then calculate distances between the Grenander CDF and the exponential CDF.

```
exp_cdf <- 1 - exp(-exp_rate * eval_points)
max_distances[i] <- max(abs(grenander_cdf - exp_cdf))
integrated_distances[i] < sum(abs(grenander_cdf - exp_cdf))
    *(eval_points[2] - eval_points[1])
```

STEP 3: Obtain PDF values from Cox results and exponential.

We obtain Grenander PDF, calculate exponential PDF, and Calculate CDFs at common evaluation points.

```
grenander_pdf <- approx(result_cox$x.knots, result_cox$f.knots,
    xout = eval_points_common, method = "linear", rule = 2)$y
exp_pdf <- exp_rate * exp(-exp_rate * eval_points_common)
grenander_cdf <- approx(result_cox$x.knots, result_cox$F.knots,
    xout = eval_points_common, method = "linear", rule = 2)$y
exp_cdf <- 1 - exp(-exp_rate * eval_points_common)
```

STEP 4: Calculate squared difference, ISE, and store results.

We keep everything larger than x_{start}.

```
keep <- eval_points_common >= x_start
```

We perform square difference calculation, to then calculate the Integrated Square Error (ISE).

```
sim_squared_diff <- (grenander_pdf[keep] - exp_pdf[keep])^2
integrated_squared_errors[i] <- sum(sim_squared_diff) * (eval_points_common[2]
      - eval_points_common[1])
```

We store all PDF and CDF results.

```
 grenander_pdf_values[i, ] <- grenander_pdf
  exp_pdf_values[i, ] <- exp_pdf
  grenander_cdf_values[i, ] <- grenander_cdf
  exp_cdf_values[i, ] <- exp_cdf

}
```

Step 1: The Gamma distribution samples are generated and sorted; this is our length biased data. Then the Gamma distributed samples go through Cox's transformation. We modify our cdf results to be able to plug them into the Grenander estimator to estimate our decreasing density.

Step 2: Evaluation points are created across the range of the simulated data. Values close to zero are excluded to avoid issues. Linear interpolation at the evaluation points are used to obtain the Grenander cdf estimates and compare them to the theoretical exponential cdf. Then the Kolmogorov-Smirnov distance(maximum absolute difference) is calculated along with the L1 distance (integrated absolute error).

Step 3: The pdf and cdf values are extracted from the Grenander results, and the theoretical exponential pdf and cdf are calculated. This is all done at a common set of evaluation points to ensure valid comparability between the estimated and true distributions.

Step 4: We filter the domain so that only values above a certain threshold are evaluated. Next, the squared differences between the Grenander and Exponential pdfs are computed, as well as the Integrated Squared Error. These will be used as measures of estimation accuracy. Lastly, we store all pdf and cdf values for further analysis.

Outside of the for loop we we calculate the average pdfs across all simulations and store the information in a data frame for plotting.

We then conducted a second simulation to see the effectiveness of our Cox-Grenander approach. In this simulation, we generated exponential data and applied the Grenander estimator directly, without any transformation(not length-biased data). By comparing the performance of the direct Grenander method against our Cox-Grenander approach, we can assess how well the Cox empirical estimator handles length-biased data. The MISE test and the K-S test are also used as evaluation tools.

In the following chapter, the results of these comparisons are presented. We plot the Cox-Grenander estimates against the Theoretical-Exponential distribution for different sample sizes, and compare these with plots of the direct Exponential-Grenander estimates against the Theoretical -Exponential distribution across the same sample sizes.

Estimation on Real Data in R

We begin by loading the data from the Excel file that is provided. The data is then filtered to only include records where the radius is greater than 0. Some variables are renamed for ease of use. The variable radius is converted to a double-precision number for numerical analysis and sorted. We then do the following:

We extract unique radius values.

```
radius.u <- unique(radius.sorted)
n.u <- length(radius.u)
```

We apply Cox's transform to the unique radius values.

```
mu <- sum(1/radius.u)/n.u
Fn_cox <- cumsum(1/radius.u)/n.u/mu
```

We make F_{ncox} and ecdf object for the Grenander estimator.

```
x_values <- radius.u
Fn_cox_ecdf <- stepfun(x_values, c(0, Fn_cox))
class(Fn_cox_ecdf) <- c("ecdf", "stepfun", class(Fn_cox_ecdf))
    attr(Fn_cox_ecdf, "call") <- call("ecdf", quote(radius.u))
```

We apply the Grenander estimator.

```
result_cox <- grenander(Fn_cox_ecdf, type="decreasing")
```

We extract the pdf and cdf coordinates from the results. An original cdf is created for comparison with our estimates, and two data frames are set up for plotting: one for our Grenander estimates and another for the original empirical Cox transformation. The results are presented in Chapter 3.2.

Results

In this chapter the results are presented. In Chapter 3.1 we present the simulation results, and in Chapter 3.2 the results from our real data.

3.1 Simulation Results

Boundary calculation

With no set boundary limit around 0 we get the following MISE values for 10000 simulations on 100 simulated data points.
Cox-Grenander vs. Exponential-Theoretical: 3377.203
Grenander-Exponential vs. Exponential-Theoretical: 861.0109

These results are not good, and the reason for this is the inconsistency of the Grenander estimator between $[0, \epsilon)$. To resolve this problem, we would like to find a point $\epsilon(n_1)$, which will be the lower boundary of the area where we would like to calculate, $[\epsilon, \infty)$.

We choose the boundary $\epsilon(n_1)$ by looking at the PDF of our first simulation of 10000 runs of 100 simulated data points, shown in the top left plot of Figure 3.1. From a close-up of this plot, we choose $\epsilon(n_1)$ to be 0.90 since that is where the Cox-Grenander PDF no longer exhibits inconsistent behavior.
Now, to find appropriate boundaries for simulations with 1000 and 10000 data points we need to perform proper proportional boundary scaling. We have that $n_1 = 100$, $n_2 = 1000$, $n_3 = 10000$, and $\epsilon(n_1) = 0.90$. To obtain $\epsilon(n_2)$, and $\epsilon(n_3)$ we perform the following calculations,

$$\begin{aligned} \epsilon(n_2) &= \left(\frac{n_2}{n_1}\right)^{-\frac{1}{3}} \epsilon(n_1) = \left(\frac{1000}{100}\right)^{-\frac{1}{3}} 0.9 = 0.42 \\ \epsilon(n_3) &= \left(\frac{n_3}{n_2}\right)^{-\frac{1}{3}} \epsilon(n_2) = \left(\frac{10000}{1000}\right)^{-\frac{1}{3}} 0.42 = 0.19. \end{aligned} \tag{3.1}$$

Having these boundary values and implementing them, we perform the simulations again and obtain nicer results that are given in the following table:

Table 3.1: MISE Simulation results for different sample sizes with a constant 10000 simulations and the boundary for each sample size. Comparing Cox-Grenander vs. Theoretical-Exponential and Exponential-Grenander vs.Theoretical-Exponential

n	**Boundary**	**MISE Cox-Grenander vs Theoretical-Exponential** (Length-Biased)	**MISE Exponential-Grenander vs Theoretical-Exponential** (Not Length-Biased)
100	$[0.90, \infty)$	0.02096812	0.005565331
1000	$[0.42, \infty)$	0.002815102	0.001528452
10000	$[0.19, \infty)$	0.0007916372	0.000382476

We assessed the performance of our Cox-Grenander estimator by computing MISE values between the Cox-Grenander and the Theoretical-Exponential PDFs. The results are given in the third column of Table 3.1 and corresponding PDF plots in the three left subplots of Figure 3.1.

We compute the MISE values between the Exponential-Grenander estimator and the Theoretical-Exponential PDF to establish a performance baseline. These ideal results are shown in the fourth column of Table 3.1, with three corresponding PDF plots on the right in Figure 3.1.

Our MISE comparison demonstrates that the Cox-Grenander estimator's performance improves with larger sample sizes up until a certain point before stabilizing. For $n = 100$, the Cox-Grenander MISE is approximately 3.8 times larger than the Exponential-Grenander baseline. This gap decreases as the sample size increases: for $n = 1000$, the ratio decreases to less than 2, but for $n = 10000$ it increases slightly and approaches 2. This suggests that the Cox-Grenander estimator converges toward baseline performance as the sample size increases, but stabilizes beyond a certain threshold with no further gain in performance.

Figure 3.1 displays the plots of the average PDFs for different sample sizes, $n = 100, 1000, \&10000$, with each estimate based on 10,000 simulation runs. The left side plots compare Cox-Grenander estimates against the Theoretical-Exponential PDF, while the right side plots show Exponential-Grenander estimates against the Theoretical-Exponential PDF.

The plot comparison in Figure 3.1 shows the convergence of both estimators. We clearly see that as the sample sizes increase, both the Cox-Grenander and Exponential-Grenander estimates gradually approach the Theoretical-Exponential distribution. We see a notable deviation when the sample size is $n = 100$ for both estimates, and when $n = 10000$, Cox-Grenander has almost no deviation, and Exponential-Grenander even less. This gives a clear indication that as the sample size increases, both estimators converge to the theoretical distribution.

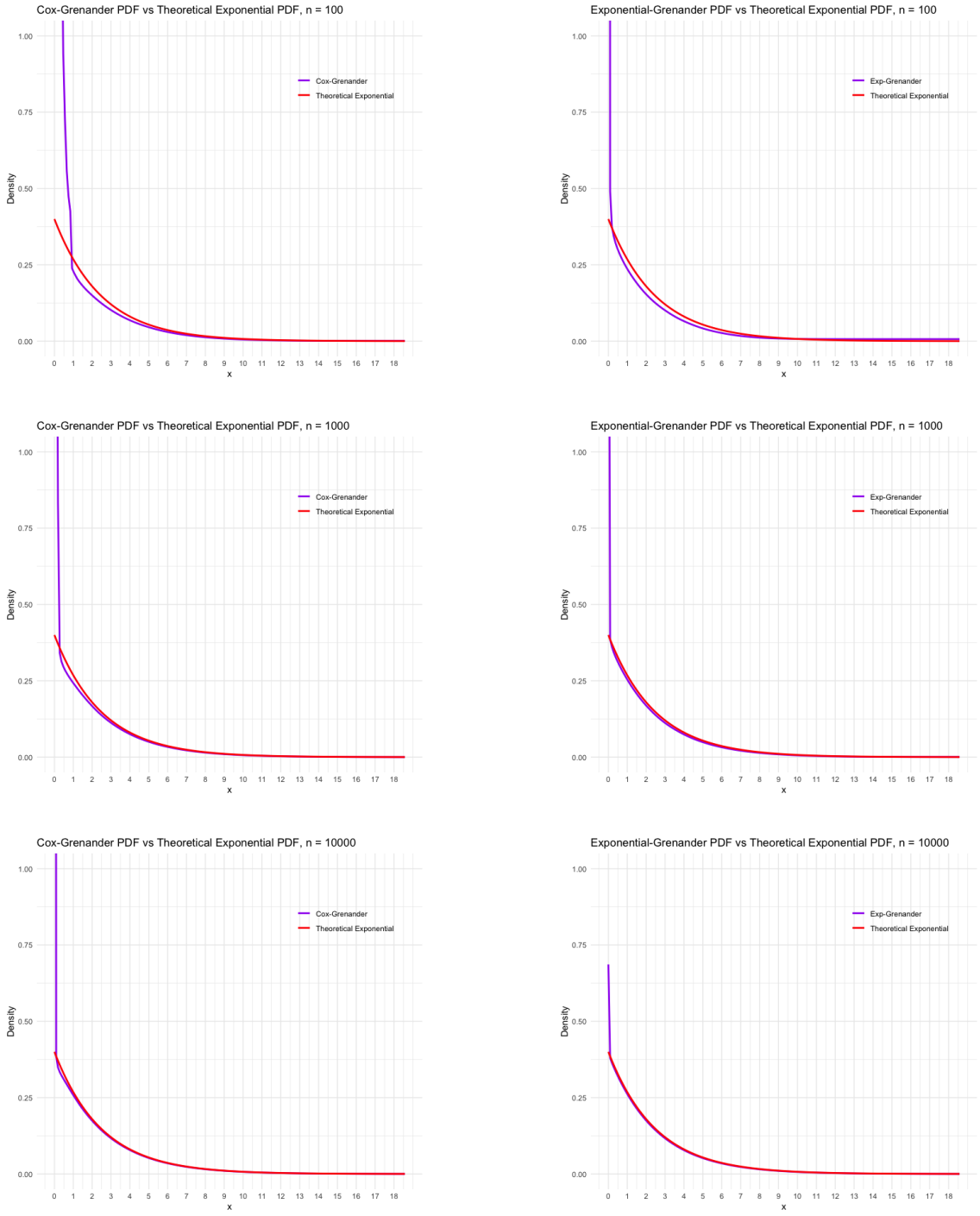

Figure 3.1: Plots on the left side are the average PDFs of Cox-Grenader vs Theoretical exponential PDFs for 100, 1000, and 10000 simulated data points. Plots on the right side are the average PDFs of Exponential-Grenander vs Theoretical-Exponential PDFs for 100, 1000, and 10000 simulated data points.

Figure 3.2 presents three plots comparing the average PDFs of Cox-Grenander and Exponential-Grenander estimators for $n = 100, 1000, \&10000$ with each result based on 10000 simulations. These comparisons support our previous results, showing that the two estimators converge toward each other and that their PDFs become

progressively similar as n increases.

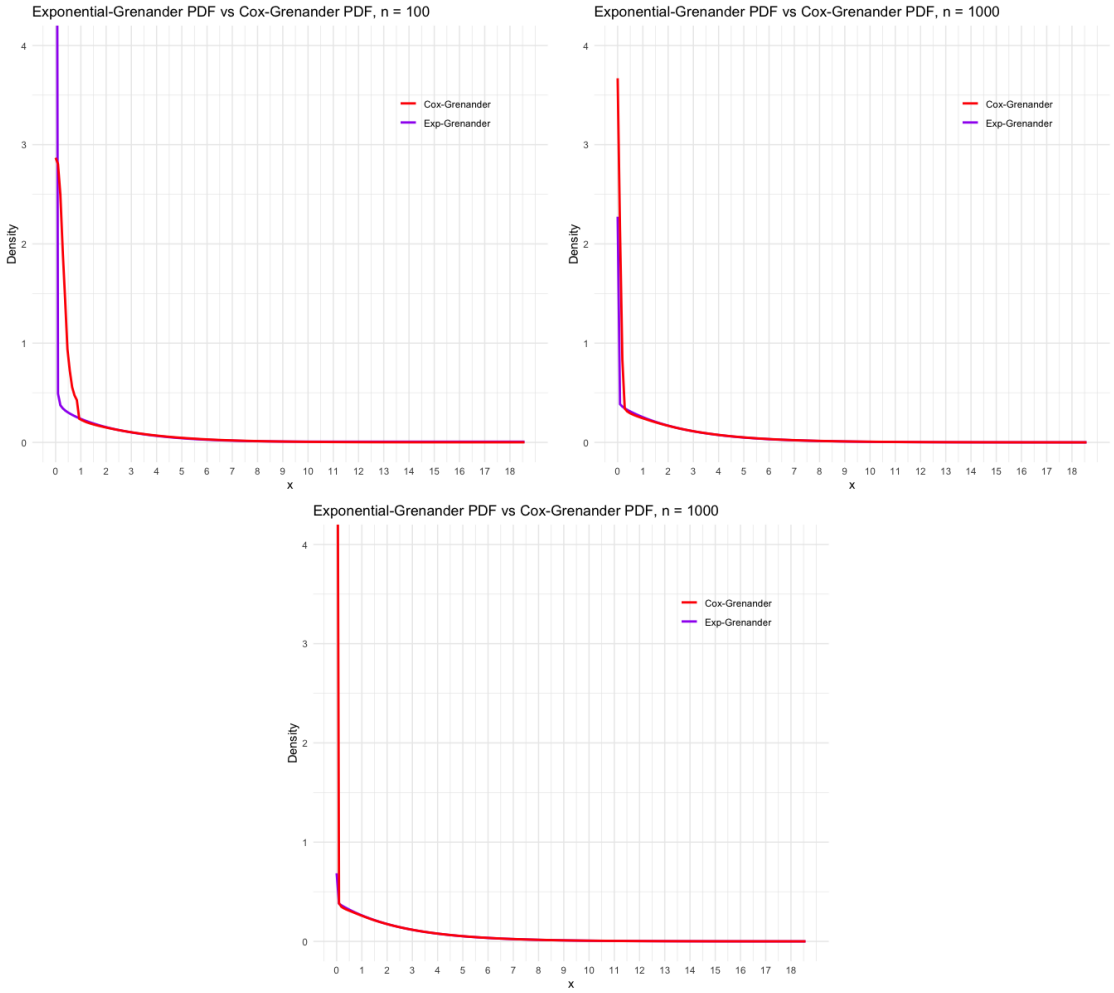

Figure 3.2: Plots of average PDFs of Exponential-Grenander vs Cox-Grenader PDFs for 100, 1000, and 10000 simulated data points.

3.2 Results from Real Data

In this section, we present results for both the $f(0)$ estimate calculation and the $f(r)$ estimate obtained by applying the Cox-Grenander method to the real data.

We begin with $f(0)$, as discussed in previous chapters, specifically Chapters 2.4.1 and 2.3.1, where it is covered how to obtain the a priori probability. Given our limited data, specifically the absence of information regarding the number of people in victims' households, we use a simple approach. We calculate the ratio of murders where the perpetrator shared a home with the victim (distance 0) to the total number of murders in our dataset (including both perpetrators at distance 0 and perpetrators at distances r). This simple ratio calculation using our given data gives $f(0) = 33/100 = 0.33$.

Next, we review the results for $f(r)$. Figure 3.3 displays the distribution of crime radius across our 67 data points(we only included the data points where $r > 0$), showing the frequency of crimes by distance from the perpetrator's residence. The plot reveals a trend: that the crime frequency is highest at distances closer to zero and decreases progressively as distance increases.

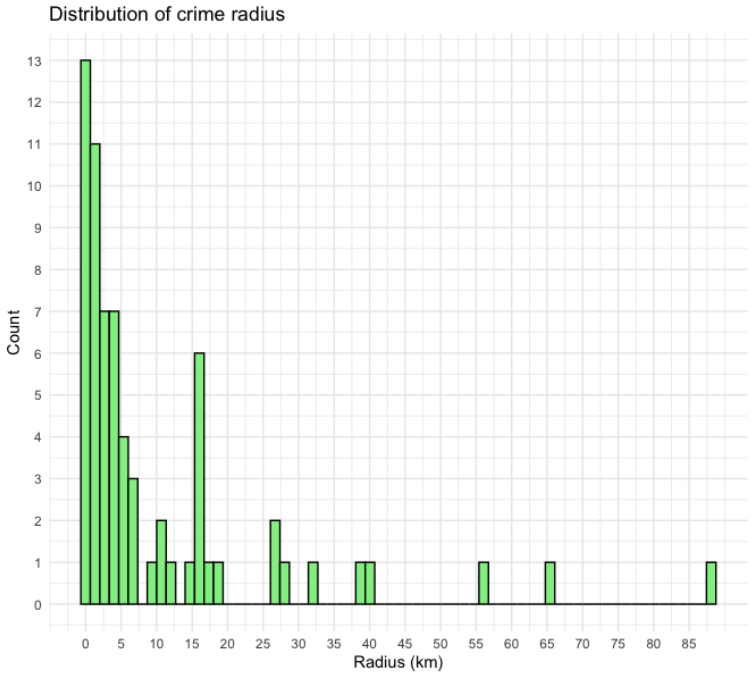

Figure 3.3: Distribution of crime radius showing the frequency of crimes by distance.

In the next plot, Figure 3.4, we show the Cox-Grenander CDF estimate against the empirical Cox estimator, both derived from the real data. The results demonstrate that the estimated Cox-Grenander distribution closely follows the empirical Cox CDF and looks as it should.

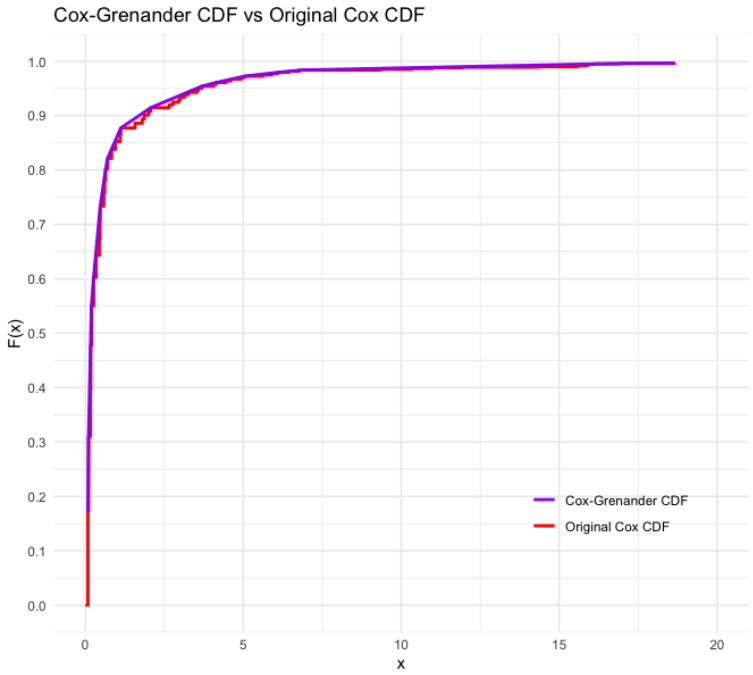

Figure 3.4: Plot of the Cox-Grenander estimate vs Cox CDf using the real data.

Finally, we have Figure 3.5, which presents the estimated Cox-Grenander PDF derived from real data, representing our $f(r)$. Further normalization and interpretation would need to be done to use this in practice.

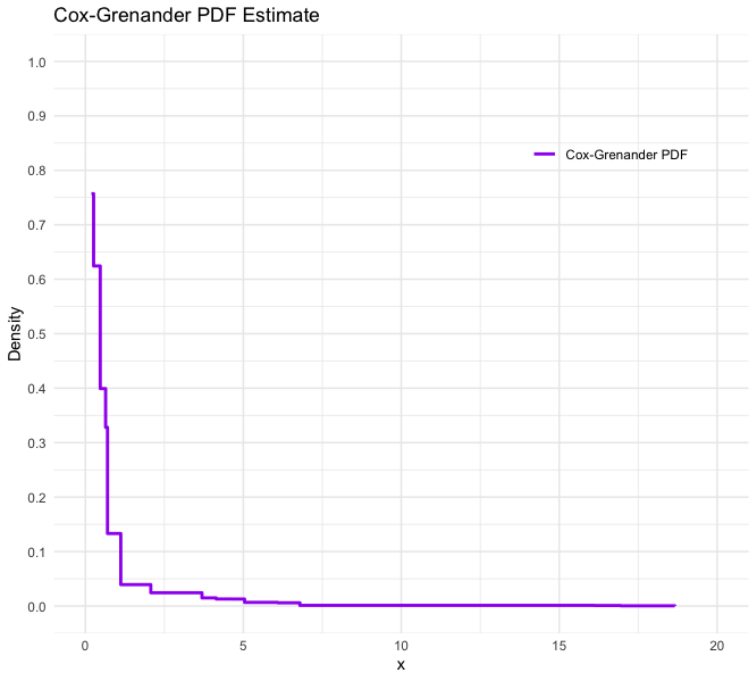

Figure 3.5: Plot of the Cox-Grenander PDF estimate using the real data.

Conclusion and Discussion

In this final chapter, we summarize the work presented in this thesis and discuss potential avenues for future research on this topic.

4.1 Conclusion

In this thesis we have presented the implementation and analysis of a developed method described in the manuscripts [3] and [1]. The objective was to implement a stochastic spatial probabilistic model for characterizing the geometric relationship between perpetrator addresses and crime scene locations. The specific crimes in question are murder cases. The model is described as an inhomogeneous spatial Poisson process with monotonically decreasing intensity on $\mathbb{R}_+$ that is a function only of the euclidean distance between perpetrator address and crime scene address. A principle insight from the referred manuscripts is that the observed data can be interpreted as length-biased observations of an unknown decreasing density function. The estimation of such densities has apparently been unexplored in previous literature.

We began by providing background and theoretical information from [3] and [1], where the set-up of the problem was presented. Chapter 2.2 covers helpful theory needed for understanding this work. In Chapter 2.3 we covered the methods of estimation for $f(0)$ and $f(r)$, and in Chapter 2.4 we covered the implementation method in R, going over the used algorithm. Finally, in Chapter 3.1 the results are reviewed, while Chapter 2.4 details the implementation approach in R, including the algorithms employed. Finally, Chapter 3.1 presents the results.

An algorithm was implemented in R, performing 10000 simulations across different data sizes. The algorithm generates n observations from a $Gamma(2, \lambda)$ distribution, representing our length-biased data. The length-biased data is then put through the Cox transformation to obtain an empirical CDF, which is in turn put through the Grenander estimator/function to obtain an estimate of $f(0)$, our monotonically decreasing function of interest. In the algorithm, theoretical exponential PDF and CDF are computed to serve as a reference model. A second algorithm was implemented to simulate unbiased data by generating exponential observations, which are directly put through the Grenander estimator/function. The theoretical exponential PDF and CDF are again computed for comparative analysis. The comparison between the Cox-Grenander estimates and the Exponential-Grenander estimates evaluates how effectively the Cox transformation removes bias from the length-biased data to obtain the desired underlying density.

The simulation results were evaluated using the MISE test to assess the differences between the estimated densities and the underlying theoretical density. Each estimator, Cox-Grenander and Exponential-Grenander, was compared against the Theoretical-Exponential distribution and then compared with each other. The Exponential-Grenander acts as the ideal target for the Cox-Grenander estimator. The results show us that as n, the number of observations per simulation, increases, performance for both estimators improves, with MISE values decreasing. However, the ratio of performance difference between the Cox-Grenander and Exponential-Grenander estimators stabilizes as n increases beyond a certain point. Specifically, for $n = 1000$ and $n = 10000$, the performance ratio remains approximately 2, whereas for $n = 100$ this ratio is closer to 4.

For the analysis of the real data results, we have that 33 observations occurred at distance zero, giving an estimate of $\hat{f}(0) = 33/100 = 0.33$. To obtain $\hat{f}(r)$, the remaining 67 observations were extracted and put through the Cox transformation, followed by the Grenander estimator/function. This gave us the density function given in Figure 3.5.

With the data that we have, to be able to apply this in a real world situation if we have multiple suspects and we want to use this data to obtain an a priori porbability on the suspect, say there is a suspect at distance 0 and at multiple distances r. Firstly we would say that the suspect 0 has a priori probability of 0.33. For the suspects at different distances we would first need to normalize the density shown in Figure 3.5. This is something that needs to be looked at in further work.

Several areas require further investigation in this project. First, normalizing the obtained density function remains critical for real-world implementation. Second, obtaining a larger dataset that includes the number of individuals per household as a variable would perhaps make our a priori probability for zero distance more comprehensive. Being aware of potential biases, such as transportation distances being longer than Euclidean distances, and the possibility that perpetrators living closer to the victim may be easier to identify. Additionally, increasing the number of observations would enable a more thorough study suitable for practical use. The goal is to develop this method into a statistically reliable system for real-world legal applications, providing lawyers with a statistical tool for assigning a priori probabilities to suspects.

References

[1] D. Anevski and R. D. Gill. *Estimating a monotone density under length biased sampling*. 2025 manuscript.

[2] D. R. Cox. Some sampling problems in technology. *In New Developments in Survey Sampling Johnson, N. L. and Smith, H. Jr. (eds.). Wiley-Interscience, New York, pp*, pages 506–527, 1969.

[3] R. D. G. Dragi Anevski, Christian Dahlman and G. Sveréu. The continuous decreasing opportunity model for prior probability in criminal cases - estimating the geographic crime baseline probability under order restriction. 2025 manuscript.

[4] M. C. Jones. Kernel density estimation for length biased data. *Biometrika*, pages 511–519, 1991.

[5] T. S. Kwun Chuen Gary Chan, Hok Kan Ling and S. C. P. Yam. Estimation of a monotone density in s-sample biased sampling models. *The Annals of Statistics, Vol. 46, No. 5*, page 2125–2152, 2018.

[6] B. E. H. . N. I. Paul. The montone density estimators from selection biased samples. *Journal of Statistical Computation and Simulation, 67:3,*, pages 203–217, 2000.

[7] P.Groeneboom and H.P.Lopuhaa. Isotonic estimators of monotone densities and distribution functions: basic facts. *Statistica Neerlandica, Vol.47, nr.3*, pages 175–183, 1993.

[8] Y.Vardi. Nonparametric estimation in the presence of length bias,. *The Annals of Statistics*, page Vol 10 no 2 616–620, 1982.