Fall 2022

Bayesian Estimation of the Intensity Function of a Non-Homogeneous Poisson Process

James Jensen
jjensen@jsu.edu

Jacksonville State University
Graduate Studies
700 Pelham Rd N., 320 Bibb Graves Hall, Jacksonville, AL 36265
Phone: (256) 782-5348
http://www.jsu.edu/graduate/index.html

Thesis Approval Form

First Name: * James Last Name: * Jensen Student ID: *

Date: * 11/08/2022

Please indicate the program you are completing:

* ○ Biology
○ Criminal Justice
○ English
○ Geographic Information Science and Technology
○ History
◉ Mathematics
○ Music
○ Psychology

Thesis Title: Bayesian Estimation of the I

Will you be using additional committee members for your thesis? * ○ Yes ◉ No

Student Signature	Electronically signed by James Jensen on 11/08/2022 11:45:53 AM
Chair Signature	Electronically signed by Jason Cleveland on 11/08/2022 11:49:59 AM
Member1 Signature	Electronically signed by Falynn Turley on 11/08/2022 11:51:09 AM
Member2 Signature	Electronically signed by Jaedeok Kim on 11/08/2022 7:56:52 PM
Department Head Signature	Electronically signed by Janice Case on 11/09/2022 9:54:27 AM
Dean of Graduate Studies Signature	Electronically signed by Channing Ford on 11/15/2022 3:11:03 PM

Bayesian estimation of the intensity function of a non-homogeneous Poisson process

A Thesis Submitted to the
Graduate Faculty
of Jacksonville State University
in Partial Fulfillment of the
Requirements for the Degree of
Master of Science
with a Major in Mathematics

By

James Mitchell Jensen II

Jacksonville, Alabama

November 14, 2022

James Mitchell Jensen II Date

ABSTRACT

In this paper we explore Bayesian inference and its application to the problem of estimating the intensity function of a non-homogeneous Poisson process. These processes model the behavior of phenomena in which one or more events, known as arrivals, occur independently of one another over a certain period of time. We are concerned with the number of events occurring during particular time intervals across several realizations of the process. We show that given sufficient data, we are able to construct a piecewise-constant function which accurately estimates the mean rates on particular intervals. Further, we show that as we reduce these intervals in size, at the limit we are able to reconstruct the original intensity function.

viii., 47 pages

ACKNOWLEDGMENTS

I would like to thank the MCIS Department and Jacksonville State University for the opportunity to pursue this degree.

I would like to thank Dr. Jason Cleveland for his patience and longsuffering.

Finally, I would like to thank my mother for her encouragement and support throughout my degree.

James Mitchell Jensen II

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Preliminaries

1.1. Experiments, sample spaces, and probability measures. A **random experiment** is a repeatable procedure with a well-defined set Ω of possible *outcomes*. We call the set of outcomes Ω the **sample space** of the random experiment. (Çınlar, 1975, pp. 1-2) An example of a random experiment is flipping a particular coin. The sample space for this random experiment is the set $\Omega = \{\text{Heads}, \text{Tails}\}$. An **event** is a subset E of a sample space, such as $\{\text{Heads}\}$.

A **probability measure** $\Pr(E)$ is a function that assigns a value to some or all of the events in a sample space, with the following properties:

P1. $\Pr(\Omega) = 1$,

P2. $Pr(E) \geq 0$ for every event E, and

P3. for any countable sequence of disjoint events $E_1, E_2, \ldots$ on which the probability measure is defined,

$$\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr\left(E_i\right).$$

Together, these are known as the *axioms of probability*. Their formulation is due to Kolmogorov. (Çınlar, 1975; Hájek, 2019)

From these properties we may deduce the following additional properties of probability measures:

P4. $\Pr(E) \leq 1$ for any event E,

P5. $\Pr(\emptyset) = 0$,

P6. For any two disjoint events E_1 and E_2, $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$, and

P7. Let $p \in [0, 1]$, and let E^C denote the complement of E. Then if $\Pr(E) = p$, then $\Pr\left(E^C\right) = 1 - p$.

Property P4 is an immediate and obvious consequence of P1 and P3. Properties P5 and P6 follow from consideration of sequences consisting entirely of the empty set except for one or two other events. Note that P6 is sometimes taken as one of the axioms of probability instead of P3 (Hájek, 2019), in which case we may require probability measures to assign probabilities to all events in a sample space, a desirable restriction that unfortunately makes P3 impossible to satisfy in the case of uncountable sample spaces (Çınlar, 1975, p. 3).

For two events A and B, the **conditional probability** $P(A|B)$ is the probability of the outcome of the random experiment being an element of event A given that the outcome is an element of event B. We will usually read this simply as "the probability of A given B."

1.1.1. Conditional probabilities. A **conditional probability** is the probability of an event given that another event has occurred. For example, let $\Omega = \{1, 2, 3, 4, 5, 6\}$ be the sample space, corresponding to the roll of a fair six-sided die, let $A = \{2, 4, 6\}$ be the event in which the die lands on an even number, and let $B = \{4, 5, 6\}$ be the event in which the die lands on one of the three highest faces.

Considered individually, A and B each have a probability $1/2$ of occurring. However, if we somehow know that the event B has occurred or will occur, then the only possible values are 4, 5, or 6. In two out of the three cases, the event A *also* occurs, so the probability of A given this extra information is $2/3$, not $1/2$. We write

this as $\Pr(A|B) = 2/3$, and say that the probability of A given B equals $2/3$.

We will say more about conditional probability in section 2.

1.2. The terms "almost all," "almost certainly," and "almost surely". Suppose an event E of some sample space has $\Pr(E) = 1$. Then we say that E will *almost surely* or *almost certainly* occur. Similarly, if all outcomes in E have a certain property, then we say that *almost all* outcomes have that property.

1.3. Random variables. A **random variable** $X(\omega)$ is a function defined on a sample space Ω that assigns a real number to each outcome $\omega \in \Omega$ (Çınlar, 1975). As an example, suppose we consider the set of ordered pairs

$$\Omega = \{\langle 1,1\rangle, \langle 1,2\rangle, \ldots, \langle 5,6\rangle, \langle 6,6\rangle\}$$

of all possible rolls of two six-sided dice as a sample space. We might wish to consider only the sum of the two dice and define the random variable $X(\omega) = X(\langle a,b\rangle) = a+b$.

Now consider the probability $\Pr(\{\omega : X(\omega) = x\})$ of the event containing all outcomes in which $X(\omega)$ takes on some value x, such as probability of two fair dice summing to a certain number. For convenience, we will often neglect the arguments for random variables, writing $X(\omega)$ as X. For the same reason we will also abbreviate probability expressions involving random variables by writing only the predicate for the set. Thus the expression above would be written as simply $\Pr(X = x)$.

A random variable may take on countably-many values, in which case we call it a *discrete random variable*, or it may take on uncountably-many values, in which case we call it a *continuous random variable*. An example of a discrete random variable is the random variable X given above that sums the results of rolling two dice. An example of a continuous random variable is the random variable $Y(\langle a,b\rangle) = |a-b|$

representing the absolute difference between two randomly-chosen real numbers a, b on the interval $[0, 1]$. In this case, Y itself takes on all values in $[0, 1]$.

For discrete random variables, the probability measure $\Pr(X = x)$ is known as the **probability mass function** or PMF for the variable. For continuous random variables, the probability of individual values $\Pr(X = x)$ is not usually well-defined, due to axiom P3. Instead, a continuous random variable will have a **probability density function**, or PDF, $p(x)$ that associates nonnegative real numbers to each possible value taken on by the variable, with the additional property that $\int_{-\infty}^{\infty} p(x)\,dx = 1$. The integral of a PDF is called a **cumulative density function** or CDF.

We may thus speak of the probability of the value of the variable being included in a set of possible values, which if well-defined will be an integral of the probability density function. To return to the case of the random variable Y from before, we might ask what the probability of the distance between two points a and b being between $1/4$ and $1/2$, and this probability is the integral

$$\Pr\left(\frac{1}{4} \leq Y \leq \frac{1}{2}\right) = \int_{\frac{1}{4}}^{\frac{1}{2}} p(y)\,dy$$

where $p(y)$ is the probability density function associated with Y.

The *expected value* $\mathbb{E}[X]$ of a random variable X is defined in two ways:

- If X is a discrete random variable taking on the values in the set X', then the expected value is defined as

$$\mathbb{E}[X] = \sum_{x \in X'} \left[x \cdot \Pr(X = x)\right].$$

- If X is a continuous random variable with PDF $p(x)$, then the expected value

is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot p(x)\, dx.$$

Finally, we note that since each random variable is in fact a function of the form $X : \Omega \to \mathbb{R}$, we may combine two or more random variables in the usual ways: sum, difference, product, scalar product, quotient, and composition. Provided it is well-defined, the resulting function $X' : \Omega \to \mathbb{R}$ will then itself be a random variable. In particular, the mean of two or more random variables, being a scalar product of the sum of those variables, will be a random variable.

1.3.1. Related variables and conditional probabilities. We will call two random variables X and Y **related** if they are defined on the same sample space. For example, suppose X and Y are both defined for the results of rolling two six-sided dice, with X being the sum of two dice when rolled and Y being the higher of the two dice. We see that Y can normally take any integer value from 1 to 6, but if it is given that $X = 10$, then the only possibilities for Y will be $Y = 5$ or $Y = 6$.

Two random variables X and Y are said to be *independent* if the value of one does not affect the probability of any value for the other — that is, if $\Pr(X = x|Y = y) = \Pr(X = x)$ and $\Pr(Y = y|X = x) = \Pr(Y = y)$ for all values x and y. It is clear that random variables that are not related are always independent.

1.4. Probability distributions. The term **probability distribution** refers to either a probability mass function or a probability density function. The former is known as a *discrete probability distribution* while the latter is called a *continuous probability distribution.*

There are several named families of probability distributions. The most well-known are the normal distributions, written $\text{Normal}\,(\mu, \sigma^2)$, where the parameters μ

and σ^2 specify precisely which distribution in the family is meant. We will be dealing extensively with the gamma distribution, written $\text{Gamma}\,(\alpha, \beta)$, and the Poisson distribution, written $\text{Poisson}\,(\lambda t)$. Their parameters will be explained later.

The symbol $\sim$ is used to indicate that a random variable has a certain probability distribution as its PMF or PDF. For example, if the discrete random variable X has a Poisson distribution as its PMF, we write $X \sim \text{Poisson}\,(\lambda)$, and say that X is Poisson-distributed. If two or more variables have the same probability distribution as their PMF or PDF, we say the variables are *identically-distributed*.

1.4.1. The Poisson distribution. A Poisson distribution is specified by one parameter, λ, a rate parameter. A Poisson distribution models the number of arrivals occurring on a unit time interval. It is a discrete probability distribution and its PMF is given by

$$\Pr\,(K = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where K is a random variable representing the number of arrivals on the time interval. The expected value of K is $\mathbb{E}[K] = \lambda$.

1.4.2. The Gamma distribution. A gamma distribution is specified by two parameters, the shape α and a rate parameter β. It can be used to model the expected length of time until the k-th arrival for a mean rate of λ arrivals per unit time interval by letting $\alpha = k$ and $\beta = \lambda$.

It is a continuous probability distribution, with a PDF given by

$$p(t) = \frac{\beta^\alpha t^{\alpha-1} e^{-\beta t}}{\Gamma(\alpha)},$$

where t is a length of time. If X is a Gamma-distributed random variable representing

the length of time, the expected value of X is $\mathbb{E}[X] = \frac{\alpha}{\beta}$.

1.5. Sampling and estimation. Sampling is the act of performing a random experiment one or more times in order to generate a sequence of outcomes, or **samples**, drawn from the sample space. From these outcomes we can compute the values of one or more random variables defined on the sample space. For example, one might flip a coin repeatedly and record whether it landed heads or tails on each flip, recording 1 for heads and 0 for tails. Alternatively, one might select a sequence of leaves on a trees and record the length of each.

Once the sequence of samples has been collected, we can use the properties of the samples to estimate unknown properties of the sample space. For example, if it is unknown how fair a particular coin is, we can estimate the bias by computing the proportion of flips that landed heads. This proportion is a **point estimate** of the probability of the coin landing heads.[1] The method for producing this estimate is a function of random variables; we call such a function an **estimator**.

An estimator is called *unbiased* if its expected value is equal to the true parameter value. A paramter that is not unbiased is called *biased*. Similarly, an estimator X is called *consistent* if its expected value almost certainly converges to the true value of the parameter as the sample size tends toward infinity.

1.6. Error terms and Chebychev's inequality. An **error term** or **error** is a random variable representing the difference between a point estimate and the true value of the parameter being estimated. If θ is the true value of a parameter and $\hat{\theta}$ the point estimate of the parameter, then the error term is $\hat{\theta} - \theta$. The *absolute error* is the absolute value of the error; i.e. $\left|\hat{\theta} - \theta\right|$.

[1]More specifically, since the probability of a coin landing heads is Bernoulli-distributed, the proportion of heads in the sample is an estimate of the value of the parameter p of the Bernoulli distribution.

Chebyshev's inequality is a theorem about the relationship between a random variable and its expected value. Let X be a random variable with expected value μ and standard deviation σ, and let k be a positive real number. Then

$$\Pr\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}.$$

This theorem can be used to set an upper bound to the size of a confidence interval (or credible interval if using Bayesian inference) for an estimator.[2] If the random variable X is an unbiased estimator for some parameter θ, then by letting $k = \sqrt{20} \approx 4.447$ we see that

$$\Pr\left(|X - \theta| \geq \sqrt{20}\sigma\right) \leq \frac{1}{20} = 0.05.$$

Thus a 95% confidence interval or credible interval will be no larger than $\sqrt{20}$ standard deviations from the mean. Note that this is merely an upper bound; the actual computed interval will usually be smaller, possibly much smaller.

1.7. The law of large numbers. There are two forms of the law of large numbers, but here we are concerned only with the *strong law of large numbers*, which states that as a random experiment is repeated many times, the average of the results will converge toward the expected value. Stated formally: consider a sequence $X_1, X_2, \ldots$ of independent and identically-distributed random variables, where each variable X_i is understood as the result of the i-th repetition of the random experiment. Since the variables are identically-distributed, $\mathbb{E}[X_i] = \mathbb{E}[X_j]$ for all positive integers i and j; denote this value $\mathbb{E}[X]$. Now let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the mean of the first n random variables in the sequence. Then $\overline{X}_n \to \mathbb{E}[X]$ as $n \to \infty$. (Ross, 2006,

[2]It should be noted that since Chebyshev's inequality is a statement about the mean of a variable, not the true value of a parameter, using it in this manner properly requires the estimator to be unbiased. However, since it is an upper bound and frequently much larger than the actual value, an estimator with a small bias will usually pose no significant issues.

p. 79)

Another way of stating the result above is that

$$\Pr\left(\lim_{n\to\infty} \overline{X}_n = \mathbb{E}[X]\right) = 1.$$

1.8. Stochastic processes and Poisson processes. A **stochastic process** is a collection of related random variables, usually thought of as modeling some real-world phenomenon (Çınlar, 1975, p. 7). For example, suppose we flip a coin n times and record the results. We define the random variables $X_1, X_2, \ldots, X_n$ where each X_i is a random variable representing the result of the i-th coin flip, with $X_i = 1$ if the coin lands heads and $X_i = 0$ if the coin lands tails. Then the collection of random variables $\{X_i, 0 \leq i \leq n\}$ is a stochastic process. (In particular, it is a Bernoulli process.)

An **arrival** is a phenomenon occuring at a particular time or during a particular interval of time. A stochastic process is called an *arrival process* if it is a collection of random variables N_t, with t nonnegative, each representing the number arrivals occuring by the time t, with the following properties (Çınlar, 1975, p. 71):

1. N_t is a nonnegative integer,
2. if $t_1 < t_2$, then $N_{t_1} \leq N_{t_2}$, and
3. if $t_1 < t_2$, then $N_{t_2} - N_{t_1}$ will be the number of arrivals occurring after t_1 but no later than t_2.

A **Poisson process** is an arrival process consisting of the random variables N_t, $0 \leq t \leq S$, with the following properties (Çınlar, 1975, p. 71):

PP1. $N_0 = 0$,

PP2. if $t_1 < t_2$ then $N_{t_2} - N_{t_1} \sim \text{Poisson}\,((t_2 - t_1)\lambda_1)$, where λ_1 is the expected number of arrivals occuring on an interval of unit length.

PP3. if $t_1 < t_2 < t_3 < t_4$, then the random variables $N_{t_4} - N_{t_3}$ and $N_{t_2} - N_{t_1}$ are independent. (That is, for any two non-overlapping time intervals, the number of arrivals occurring on one does not affect the other.)

Poisson processes are used to model the occurrence of arrivals over a period of time, such as the arrival of customers at a store during operating hours. However, its constant rate coefficient λ_1 causes it to be a poor fit for modeling phenomena where arrivals are more likely to occur at some times than at other times, as in the example of a store; people are presumably more likely to go shopping outside of normal work hours for the area.

We can better model such phenomena by replacing PP2 above with the following:

PP2a. If $t_1 < t_2$, then $N_{t_2} - N_{t_1} \sim \text{Poisson}\left(\int_{t_1}^{t_2} \lambda(\tau)\, d\tau\right)$, where $\lambda(\tau)$ is an **intensity function** specifying the likelihood of arrivals for each point on the period $[0, S]$.

An arrival process meeting this new definition is called a **non-homogeneous Poisson process**, or NHPP, and it is with these sorts of processes that we will primarily concern ourselves in Section 2.

1.8.1. Realizations of a stochastic processes. Suppose we perform a random experiment with a sample space Ω, with the result that outcome $\omega \in \Omega$ occurs. Then we call ω a **realization** of the sample space. If the random variables of a stochastic process are defined on Ω, then we say that ω is a realization of the stochastic process. Repeating the random experiment will generate a finite or infinite sequence of realizations $\omega_1, \omega_2, \omega_3, \ldots$.

When discussing a realization or sequence of realizations of a stochastic process, we are often concerned with the values of the random variables of the process for each realization. Thus if X is a random variable of the process, and $\omega_1, \omega_2, \ldots, \omega_k$ are k realizations of the process, we may consider the sequence of values

$$X(\omega_1), X(\omega_2), \ldots, X(\omega_k),$$

where $X(\omega_i)$ is the value of the random variable X on the i-th realization. When we observe the values of the random variables of the process associated with a single realization, we say that we "observe a realization of the process." When we do this for a sequence of k realizations, we say we "observe k realizations of the process."

It is often more natural to think of the sequence $X(\omega_1), X(\omega_2), X(\omega_3), \ldots$ as the values of a sequence of random values $X_1, X_2, X_3, \ldots$. We may make this notion rigorous by defining each X_i as a function whose domain is the set $\Omega^* = \{\langle \omega_1, \omega_2, \omega_3, \ldots \rangle : \omega_j \in \Omega, j \geq 0\}$ of all sequences of outcomes, treating each sequence of realizations of the process as an outcome in a larger sample space. A similar definition can be made for sequences of finite length; i.e. a sequence of outcomes of length k will be treated as an outcome in the sample space $\Omega^k = \Omega \times \Omega \times \cdots \times \Omega$ (k times).

1.9. Partitions and refinements. A **partition** $\mathcal{P}$ of a closed interval $[a, b]$ is a set of n points $\{x_0 = a, x_1, x_2, \ldots, x_{n-1}, x_n = b\} \subset [a, b]$ with the property that $x_{i-1} < x_i$ for $0 < i \leq n$. These points form a sequence of n subintervals

$$[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]$$

of the interval $[a, b]$ with the properties that every point in $[a, b]$ is also in at least one subinterval, and no two subintervals share more than an endpoint in common.

The subintervals in this sequence are called the subintervals of the partition $\mathcal{P}$. The length of the largest such subinterval is called the **mesh** of $\mathcal{P}$ and is denoted $|\mathcal{P}|$. (Mattuck, 1999)

If $\mathcal{R}$ is also a partition of $[a, b]$ with the property that $\mathcal{P} \subseteq \mathcal{R}$, then $\mathcal{R}$ is a **refinement** of $\mathcal{P}$. Similarly, a **sequence of refinements** $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \ldots$ of a partition $\mathcal{P}$ is a sequence of partitions with the property that

$$\mathcal{P} = \mathcal{P}_1 \subseteq \mathcal{P}_2 \subseteq \mathcal{P}_3 \subseteq \cdots .$$

That is, each partition in the sequence is a refinement of the previous partition in the sequence. (Mattuck, 1999)

2. Bayes' rule and Bayesian inference

Bayes' rule is a theorem about conditional probability. Given two events A and B, with $\Pr(B) \neq 0$, the conditional probability of A given B can be calculated as

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}.$$

Three parts of the equation are given standard names:

- the event B is known as the **evidence**,
- the conditional probability $\Pr(B|A)$ is the **likelihood**,
- the probability $\Pr(A)$ is known as the **prior probability**, and
- the conditional probability $\Pr(A|B)$ is known as the **posterior probability**.

The probability of the evidence $\Pr(B)$ does not have a standard name, but we will call it simply the **denominator**, following Ben Lambert. (Lambert, 2018)

Most often when using Bayes' rule we are considering probability of a random variable θ having a particular value given that we have observed a certain set of data $\underline{X}$. In this case we will usually write Bayes's rule in the form

$$\Pr(\theta|\underline{X}) = \frac{\Pr(\underline{X}|\theta) \cdot \Pr(\theta)}{\Pr(\underline{X})}.$$

Bayesian inference is the use of Bayes' rule in statistical inference. This presents several challenges but can provide more robust results than conventional frequentist methods. When doing Bayesian inference, we treat the prior and posterior not as particular probabilities but as probability distributions with parameter θ, and call them the *prior probability distribution* and the *posterior probability distribution*, or simply the *prior distribution* and *posterior distribution*. The likelihood $P(\underline{X}|\theta)$ in this case will also be a function of the parameter θ and will be based on a probability distribution, but because its values will not generally sum or integrate to 1 across all values θ, we will not consider it a probability distribution and will simply call it the *likelihood function* instead.[3]

If the distribution has more than one parameter, we simply extend the notation to accommodate multiple parameters. For example, if the prior distribution has two parameters, α and β, we may write it as $\Pr(\alpha, \beta)$. Similarly, we would write a posterior distribution with those two parameters as $\Pr(\alpha, \beta|\underline{X})$ and the likelihood function as $\Pr(\underline{X}|\alpha, \beta)$.

2.1. The prior in Bayesian inference. In the ideal case, the prior probability of an event or parameter value is known in advance. Unfortunately, this is often not the case when inferring the value of a parameter. In such cases, a prior is chosen based on pre-existing information and opinion. For example, a statistician investigating the rate at which customers enter a particular store may consult security camera footage, ask the store attendants for estimates, or even observe one or more days before attempting the experiment in order to determine an appropriate prior.

The choice of prior is typically considered to be subjective, not objective.

[3]In cases where the probability distribution associated with any of the prior, posterior, or likelihood is continuous, the notation above is techincally incorrect, because in that case one or more of those terms refer not to a probability but a probability density. It is easily seen that Bayes' theorem properly needs $2^3 = 8$ forms to account for all these cases. However, as this distinction rarely matters in practice, we will retain the given form for all such cases as an abuse of notation.

This is because of the Bayesian view of probability as a subjective "degree of belief," not necessarily an objective property. For example, following this interpretation of probability we might assign probabilities to the various possible outcomes of an event that has already happened, but whose outcome is still unknown.

The classic example of this is a coin which has been flipped but not yet revealed. We might assign a probability 0.5 that the coin has landed heads, and likewise for tails, reflecting an initial assumption that the coin is fair. Or perhaps we believe we have reason to assign probability 0.75 to heads and 0.25 to tails, from some additional source of information about the coin. Regarldess, the coin has already been flipped, and the outcome of the flip has already been determined. Thus the probabilities we assign to the two outcomes cannot be facts about the event itself, but about the information available to us.

A common objection to Bayesian inference is that with the proper choice of prior, we can produce any desired posterior. Making this choice subjective then seems to be a license to bias the data as we wish. The Bayesian response to this objection is threefold. First, provided the prior is stated explicitly, it is open to inspection and criticism. Anyone who objects to a chosen prior may choose a different prior and make a new calculation of the posterior. Second, given sufficient evidence, the posteriors produced by different priors will tend to converge. (Stone, 2013, p. 126)

Finally, Bayesian statisticians have several methods at their disposal for choosing priors that emphasize the role of the data in determining the posterior. These include choosing a so-called *uninformative prior*, the method of moments, and using Bayesian hierarchical models. (Wijono, 2022)

2.2. Posteriors as future priors. Where the prior represents the information available to us before beginning a formal experiment or investigation, the posterior

represents the information available to us afterward — which continues to include the information present in the prior. Contemporary Bayesian statisticians often think of the posterior as an *update* to the prior. The posterior can then be used as the prior in any future experiments or investigations into the same subject, since it represents the state of our knowledge and belief at that time. By utilizing posteriors as priors in future inference, our understanding may be iteratively updated. (Donovan & Mickey, 2019)

2.3. Conjugate priors. In some instances when applying Bayes' rule, the prior and posterior distributions will be in the same family of distributons. For example, they might both be gamma distributions. In this case, we call the prior distribution a **conjugate prior**, and we say that prior distribution *is conjugate to* the distriubution from which the likelihood function is drawn; e.g. the Beta distribution is conjugate to the Bernoulli distribution. Conjugate priors often make the computation of posterior distributions much easier. (Winkler, 1972, pp. 147-148)

Whether a prior distribution is a conjugate prior depends on the form taken by the likelihood function. As an example, if the prior distribution is a gamma distribution and the likelihood function is drawn from a Poisson distribution, then the posterior distribution will also be gamma distribution, as we will see below. On the other hand, if the likelihood function were the PMF of a Bernoulli distribution, then the posterior distribution will *not* be a gamma distribution. We thus say that the gamma distribution is conjugate to a Poisson distribution, but not conjugate to a Bernoulli distribution.

2.4. The Gamma distribution is conjugate to a Poisson distribution. In Section 3 we will utilize a Gamma prior in order to estimate the rate parameter λ of a Poisson-distributed random variable. We choose Gamma for two reasons: first, its

mean, α/β, naturally estimates the rate of arrivals when α arrivals are observed across β unit time intervals; and secondly, because it is conjugate to a Poisson distribution. We will demonstrate the latter fact below.

Lemma 1. *A Gamma distribution is conjugate to a Poisson distribution.*

Proof. Suppose that we observe n arrivals on a Poisson process across a unit interval of time, either on a single realization of the process or multiple realizations, and we seek to estimate the rate λ of the process. We see that

$$\Pr(n|\lambda) = \text{Poisson}(\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}.$$

We choose

$$\Pr(\lambda) = \text{Gamma}(\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0} \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda}}{\Gamma(\alpha_0)}$$

as our prior probability distribution. Together $\Pr(n|\lambda)$ and $\Pr(\lambda)$ give us

$$\Pr(n|\lambda)\ \Pr(\lambda) = \frac{\lambda^n e^{-\lambda}}{n!} \times \frac{\beta_0^{\alpha_0} \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda}}{\Gamma(\alpha_0)} = \frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)} \lambda^{(\alpha_0 + n) - 1} e^{-(\beta + 1)\lambda}$$

and

$$\begin{aligned}
\Pr(n) &= \int_0^\infty \Pr\left(n|\theta\right) \Pr(\theta)\, d\theta \\
&= \int_0^\infty \frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)} \theta^{(\alpha_0+n)-1} e^{-(\beta+1)\theta}\, d\theta \\
&= \frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)} \int_0^\infty \theta^{(\alpha_0+n)-1} e^{-(\beta+1)\theta}\, d\theta \\
&= \frac{(\beta_0+1)^{\alpha_0+n}}{\Gamma(\alpha_0+n)} \times \frac{\Gamma(\alpha_0+n)}{(\beta_0+1)^{\alpha_0+n}} \times \frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)} \int_0^\infty \theta^{(\alpha_0+n)-1} e^{-(\beta+1)\theta}\, d\theta \\
&= \frac{\Gamma(\alpha_0+n)}{(\beta_0+1)^{\alpha_0+n}} \times \frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)} \int_0^\infty \frac{(\beta_0+1)^{\alpha_0+n}}{\Gamma(\alpha_0+n)} \theta^{(\alpha_0+n)-1} e^{-(\beta+1)\theta}\, d\theta \\
&= \frac{\Gamma(\alpha_0+n)}{(\beta_0+1)^{\alpha_0+n}} \times \frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)} \times \int_0^\infty \text{Gamma}\left(\alpha_0+n, \beta_0+1\right)\, d\theta \\
&= \frac{\Gamma(\alpha_0+n)}{(\beta_0+1)^{\alpha_0+n}} \times \frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)} \times 1 \\
&= \frac{\Gamma(\alpha_0+n)}{(\beta_0+1)^{\alpha_0+n}} \times \frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)}.
\end{aligned}$$

We now apply the above results to Bayes' theorem to calculate the posterior distribution:

$$\begin{aligned}
\Pr(\lambda|n) &= \frac{\Pr(n|\lambda)\ \Pr(\lambda)}{\Pr(n)} \\
&= \frac{\frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)} \lambda^{(\alpha_0+n)-1} e^{-(\beta+1)\lambda}}{\frac{\Gamma(\alpha_0+n)}{(\beta_0+1)^{\alpha_0+n}} \times \frac{\beta_0^{\alpha_0}}{n!\,\Gamma(\alpha_0)}} \\
&= \frac{(\beta_0+1)^{\alpha_0+n}}{\Gamma(\alpha_0+n)} \lambda^{(\alpha_0+n)-1} e^{-(\beta+1)\lambda} \\
&= \text{Gamma}\left(\alpha_0+n, \beta_0+1\right).
\end{aligned}$$

□

We can generalize the results of the above lemma to $k > 1$ realizations:

Corollary 1. *Let n be the total number of arrivals observed across k realizations*

of a Poisson process. Let the prior probability distribution for the rate parameter λ *of the process be* $Gamma(\alpha, \beta)$*. Then the posterior probability distribution will be* $Gamma(\alpha + n, \beta + k)$*.*

Proof. Let $n_1, n_2, \ldots, n_k$ be the number of arrivals on each of k observed realizations of the process, such that $n = \sum_{i=1}^{k} n_i$. By applying Bayes' theorem repeatedly, utilizing the posterior probability from each application as the prior probability in the succeeding application, the corollary follows. □

2.5. The role of the denominator. The denominator in Bayes' theorem plays a special role when the posterior is a probability distribution: it is the sum or integral of the numerator $\Pr\left(\overline{X}|\theta\right) \times \Pr(\theta)$ across all possible values $\theta \in \Theta$. In other words,

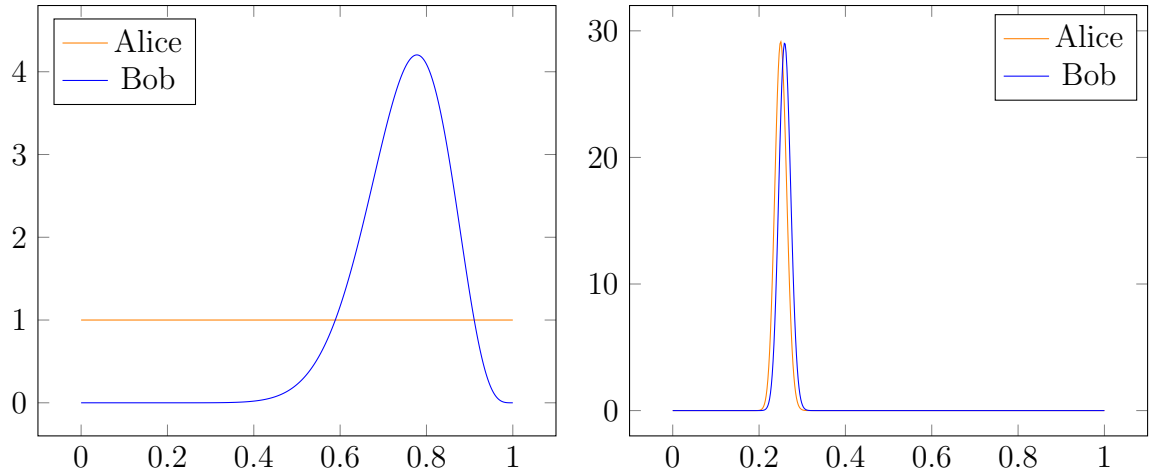
$$\Pr\left(\overline{X}\right) = \sum_{t \in \Theta} \left[\Pr\left(\overline{X}|t\right) \times \Pr(t)\right] \text{ (if the posterior is discrete)}$$

or

$$\Pr\left(\overline{X}\right) = \int_{\Theta} \Pr\left(\overline{X}|t\right) \times \Pr(t)\, dt. \text{ (if the posterior in continuous)}$$

This can be easily seen by considering that it must be a single value which causes the whole fraction to sum or integrate to 1 (and thus to be a probability distribution). (Lambert, 2018, pp. 109-110)

2.6. Calculating the posterior probability distribution. As seen above, the posterior probability distribution can be specified exactly when the prior distribution is a Gamma distribution and the likelihood function comes from a Poisson distribution. However, in some cases, finding the denominator analytically is infeasible or impossible. In these cases Markov Chain Monte Carlo techniques can be used to approximate the posterior distribution without calculating the denominator directly. Methods include the Metropolis algorithm and Gibbs sampling. (Kruschke,



(a) Priors (b) Posteriors

FIG. 1. Alice and Bob's Priors and Posteriors

2014, pp. 144-146, 162)

2.7. An example of Bayesian inference. We will illustrate the process of performing Bayesian inference with an example: Let Alice and Bob be two Bayesian statisticians both observing the results of flipping a particular coin 1,000 times. Alice knows nothing about the coin and so chooses the uninformative prior $\text{Beta}(1, 1)$. Bob, on the other hand, was given a chance to flip the coin twenty times before the beginning of the experiment, and observed that it landed on heads fifteen times and tails only five times. Bob therefore chooses the prior $\text{Beta}(15, 5)$, in effect assuming the coin is biased toward heads. These prior distributions are quite different, as can be seen in Figure 1a.

During the course of the experiment, both Alice and Bob observe the coin landing heads 250 times and tails 750 times. Applying Bayes' rule using a Bernoulli likelihood function, Alice calculates the posterior distribution as $\text{Beta}(251, 751)$, while Bob calculates the posterior distribution as $\text{Beta}(265, 755)$. These distributions are not noticeably different, as can be seen in Figure 1b.

This lack of difference can be seen another way. Before the experiment, Alice

would have estimated the probability p of the coin landing heads to be $1/2 = 0.5$, while Bob would have estimated it to be $15/20 = 0.75$. After the experiment, Alice would estimate p to be $251/1002 \approx 0.25$ while Bob would estimate it to be $265/1020 \approx 0.26$ a difference of less than 0.01. Had more data been collected this difference would likely have been even smaller.

Moreover, Alice and Bob may use their posterior distributions as the prior distributions in a future experiment on the same coin. The evidence collected will cause their new posterior distributions to converge even more closely. The ability to use the results of previous experiments to inform the results of new experiments is one of the advantages of Bayesian inference.

2.8. Credible intervals. Once the posterior distribution has been calculated, we may use it to calculate a **credible interval** for the value of the parameter. (Winkler, 1972, p. 177) Credible intervals are the Bayesian analogue of the confidence interval used in conventional statistics. Both involve computing lower and upper bounds for a parameter θ in order to in some sense capture the value of a parameter with probability γ. They differ in two ways. First, since credible intervals are computed from posterior probability distributions, they incorporate information both from the experiment and from the prior. Confidence intervals include only information from the experiment. Secondly, they differ in how the parameter and interval bounds are interpreted.

Confidence intervals treat the parameter as a fixed value and conceive of the confidence level in terms of the probability of the interval bounds capturing the value of the parameter. Credible intervals treat the parameter as a random variable with a probability distribution, and they treat the interval bounds as the endpoints of an interval with a chosen probability mass on the probability distribution. As such, it

should be noted that while students of statistics are frequently reminded that a 95% confidence intervals does not mean that there is 0.95 probability that the parameter is in the interval, a 95% credible interval *does* indicate this!

In order to calculate a credible interval with probability mass $\gamma \in (0, 1)$ for a given parameter θ, we choose real numbers $a \leq b$ such that $\Pr(a \leq \theta \leq b) \geq \gamma$. While in principle any such interval may be chosen, it is customary when working with continous random variables to choose a and b such that $\Pr(\theta < a) = \Pr(\theta > b) = (1 - \gamma)/2$. As an example, if $\gamma = 0.95$, then we would choose a and b such that $\Pr(\theta < a) = \Pr(\theta > b) = 0.025$.

In order to compute a and b using this method, we make use of the inverse cumulative distribution function for the probability distribution for the random variable θ. Suppose $F(x)$ is the CDF for θ, and let us denote its inverse as $F^{-1}(x)$. Then for any real number x, $\Pr(\theta < x) = F^{-1}(x)$. Thus we use this function to choose $a = F^{-1}(0.025)$ and $b = F^{-1}(1 - 0.025) = F^{-1}(0.975)$.

It is clear that this method depends on our ability to find or approximate the inverse CDF for a probability distribution. This is often not an easy task to perform by hand, since in many cases the CDF of a probability distribution has no closed form. Fortunately, this is usually easily done with the aid of computer software.

3. Inferring the intensity function

Consider a non-homogeneous Poisson process (NHPP) defined on the interval $0 \leq t \leq S$, with random variables N_t for each value of t and a continuous intensity function $\lambda(t)$. Let $\mathcal{P}$ be any partition of $[0, S]$. Without loss of generality, we consider only one such interval, calling it T and denoting the length of T as $|T|$. Let the random variable $N_T(\omega)$ be the number of arrivals occurring on the interval T on a realization ω.

We will utilize a Gamma prior probability distribution with parameters α and β. Such a prior has two advantages. First, it is conjugate to the Poisson distribution, greatly simplifying the calculation of the prior. Second, its expected value is α/β, which naturally estimates the mean number of arrivals on a unit-interval period of a process after observing α arrivals on β realizations.

We will use the expected value of the posterior probability distribution as our estimator for the mean value of the $\lambda(t)$ on the interval, adjusted for the length of the interval. Thus if our posterior for the interval T is Gamma$(\alpha + n, \beta + k)$, our estimate will be

$$\frac{\alpha + n}{(\beta + k)|T|}.$$

3.1. Estimating the mean value on an interval. Given the above setup, we seek to estimate the mean value of the intensity function $\lambda(t)$ on the interval T; call this value $\overline{\lambda}_T = \int_T \lambda(t)\, dt$, and denote our point estimate $\hat{\lambda}_T$. We assume that

$\overline{\lambda}_T$ is Gamma-distributed, with a prior distribution Gamma(α_0, β_0) for some positive real numbers α_0, β_0. We observe a sequence of realizations of the NHPP of length k, namely $\omega_1, \omega_2, \omega_3, \ldots, \omega_k$, giving the sequence of values $N_T(\omega_1), N_T(\omega_2), N_T(\omega_3), \ldots, N_T(\omega_k)$ for the random variable $N_T(\omega)$, each representing the numbers of arrivals occuring on the interval T on each realization.

Let $n = \sum_{i=1}^{k} N_T(\omega_i)$ be the total number of arrivals across all k realizations of the process. Then by using Bayes' Theorem, the posterior distribution

$$\Pr(\hat{\lambda}_T | n, k) = \frac{\Pr(n, k | \hat{\lambda}_T) \times \Pr(\hat{\lambda}_T)}{\Pr(n, k)}$$

is given by Gamma$(\alpha_0 + n, \beta_0 + k)$.

We begin by proving a lemma for the case when T has unit length.

Lemma 2. *If the interval T has unit length, then the estimator $\hat{\lambda}_T = \mathbb{E}[Gamma(\alpha_0 + n, \beta_0 + k)]$ is a consistent estimator of $\overline{\lambda}_T$.*

Proof. Suppose that $|T| = 1$. Then the expected value of the posterior may be used without adjustment as the estimator for the mean rate of the process, as explained above. By definition, for $\hat{\lambda}_T$ to be a consistent estimator of $\overline{\lambda}_T$, it must be the case that

$$\Pr\left(\lim_{k \to \infty} \hat{\lambda}_T = \overline{\lambda}_T\right) = 1.$$

We note that we are *not* attempting to show that $\hat{\lambda}_T$ is an unbiased estimator, as this will depend on the chosen prior. It is not too hard to see that $\hat{\lambda}_T$ is an unbiased estimator if and only if $\mathbb{E}[\Pr(\lambda)] = \overline{\lambda}_T$, that is, if the expected value of the chosen prior is already correct. Further, the amount of bias will be proportional to the difference $\left|\mathbb{E}[\Pr(\lambda)] - \overline{\lambda}_T\right|$.

We see from the definition of an NHPP that $N_T \sim \text{Poisson}\left(\overline{\lambda}_T\right)$. We wish to

estimate the value of this parameter using Bayesian inference. We assume that the value of $\overline{\lambda}_T$ is a gamma-distributed random variable with parameters α and β. We will be using Bayes' rule to calculate a posterior distribution for $\overline{\lambda}_T$, so we begin with Gamma(α, β) as our prior distribution.

We then observe k realizations of the process, yielding a sequence of values $N_T(\omega_1), N_T(\omega_2), \ldots, N_T(\omega_k)$, where each value $N_T(\omega_i)$ is the number of arrivals on the interval T on the i-th realization of the process. Let $n = \sum_{i=1}^{k} N_T(\omega_i)$ be the total number of arrivals on the interval T across all k realizations. Our posterior distribution is then Gamma $(\alpha + n, \beta + k)$ by Lemma 1, with an expected value of $\hat{\lambda}_T = \frac{\alpha+n}{\beta+k}$; we take this to be our estimator for $\overline{\lambda}_T$.

What we will show is that this method provides increasingly-accurate estimations of $\overline{\lambda}_T$ as we observe more realizations of the process. As mentioned in subsubsection 1.7.1 above, we may recharacterize the sequence of values $N_T(\omega_1), N_T(\omega_2), \ldots, N_T(\omega_k)$ as a sequence of random variables $X_1 = N_T(\omega_1), X_2 = N_T(\omega_2), \ldots, X_k = N_T(\omega_k)$. We then see that their mean $\overline{X}_k := \frac{1}{k}\sum_{i=1}^{k} X_i$ is a random variable representing the mean number of arrivals per realization on the interval. Furthermore, by the strong law of large numbers, the limit of the mean converges to the expected value with probability 1. In other words, with probability 1,

$$\overline{X} := \lim_{k\to\infty} \overline{X}_k = \mathbb{E}[N_T] = \overline{\lambda}_T \tag{1}$$

We are now in a position to recharacterize n as a function of k. Let $n(k) := k\overline{X}_k = \sum_{i=1}^{k} X_i$ and note that this is equivalent to the way we defined n above, except that it is now clearly itself a random variable, being the sum of k random variables. We now consider the estimator $\hat{\lambda}_k = \frac{\alpha+n(k)}{\beta+k}$ as k increases—that is, as we observe

increasingly-many realizations. We then have

$$\lim_{k\to\infty} \hat{\lambda}_k = \lim_{k\to\infty} \frac{\alpha + n(k)}{\beta + k} \tag{2}$$

$$= \lim_{k\to\infty} \left[\frac{\alpha}{\beta + k} + \frac{n(k)}{\beta + k} \right] \tag{3}$$

$$= \lim_{k\to\infty} \frac{n(k)}{\beta + k} \tag{4}$$

$$= \lim_{k\to\infty} \frac{k\overline{X}_k}{\beta + k} \tag{5}$$

$$= \lim_{k\to\infty} \frac{\overline{X}_k}{\frac{\beta}{k} + 1} \tag{6}$$

$$= \lim_{k\to\infty} \frac{\overline{X}_k}{0 + 1} \tag{7}$$

$$= \lim_{k\to\infty} \overline{X}_k = \overline{\lambda}_T. \qquad \text{[with probability 1]} \tag{8}$$

We thus see that regardless of the initial prior parameters α and β, with sufficient evidence, our estimator $\hat{\lambda}_T$ will almost certainly converge to the true value $\overline{\lambda}_T$. Since this is true for *any* subinterval T, it is true for every subinterval in $\mathcal{P}$. □

Having proven the lemma, the following theorem follows immediately:

Theorem 1. *The estimator* $\hat{\lambda}_T = \mathbb{E}[Gamma(\alpha_0 + n, \beta_0 + k)]|T|$ *is a consistent estimator of* $\overline{\lambda}_T$.

Having established that sufficient evidence allows the mean value $\overline{\lambda}_T = \frac{1}{|T|} \int_T \lambda(t)\, dt$ to be recovered for any subinterval T, we now consider whether we can recover the original intensity function $\lambda(t)$.

3.2. Recovering the intensity function $\lambda(t)$. Consider a sequence of refinements $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \ldots$ of the partition $\mathcal{P}$ with the property that $\lim_{i\to\infty} |\mathcal{P}_i| = 0$. Let $T_{i,j}$ denote the j-th subinterval in the partition $\mathcal{P}_i$, and let $\overline{\lambda}_{T_{i,j}}$ denote the mean

value of the intensity function $\lambda(t)$ on the subinterval $T_{i,j}$. Let

$$\hat{\lambda}_{i,k}(t) = \mathbb{E}\left[\text{Gamma}(\alpha + n_{T_{i,j}}(k), \beta + k)\right] |T_{i,j}|, \qquad t \in T_{i,j}$$

be the piecewise constant estimation of $\lambda(t)$ constructed by observing $n_{T_{i,j}}(k)$ arrivals on the subinterval $T_{i,j}$ across k realizations with prior parameters α and β. Finally, define the sequence of estimator functions $\hat{\lambda}_1(t), \hat{\lambda}_2(t), \hat{\lambda}_3(t), \ldots$ by

$$\hat{\lambda}_i(t) = \lim_{k\to\infty} \mathbb{E}\left[\text{Gamma}(\alpha + n_{T_{i,j}}(k), \beta + k)\right] |T_{i,j}|, \qquad t \in T_{i,j}, i \geq 1.$$

It follows from Theorem 1 that

$$\Pr\left(\hat{\lambda}_i(t) = \overline{\lambda}_{T_{i,j}}\right) = 1, \qquad t \in T_{i,j}.$$

Since this result is almost certain, in the following theorem we will dispense with the probability notation, treating the result as simply true, with the caveat that any result following from it is in fact only almost certain.

Theorem 2. *The sequence of estimator functions $\hat{\lambda}_1(t), \hat{\lambda}_2(t), \hat{\lambda}_3(t), \ldots$ converges pointwise to $\lambda(t)$ as $i \to \infty$.*

Proof. Consider the absolute difference $\left|\hat{\lambda}(t) - \lambda(t)\right|$ for each point t in each subinterval T of each refinement. What is needed is to show that $\left|\hat{\lambda}(t) - \lambda(t)\right| \to 0$ as $i \to \infty$ for each subinterval T.

It is clear that the difference cannot be greater than that between the supremum and infimum of $\lambda(t)$ on T: since $\lambda(t)$ is continouous, the Integral Mean Value Theorem guarantees that there is a point $t_0 \in T$ such that $\lambda(t_0) = \overline{\lambda}_T$. Since $t_0 \in T$, the value $\lambda(t_0)$ must be no less than the infimum and no greater than the supremum

of the function on T. Therefore,

$$\left|\overline{\lambda}_T - \lambda(t)\right| \leq \sup_T \lambda(t) - \inf_T \lambda(t). \tag{9}$$

But since $|\mathcal{P}_i| \to 0$ as $i \to \infty$, $\sup_T \lambda(t) - \inf_T \lambda(t) \to 0$ as $i \to \infty$. It follows that $\left|\overline{\lambda}_T - \lambda(t)\right| \to 0$ as $i \to \infty$ for each subinterval T of each refinement, which is what was needed. We conclude that the sequence $\hat{\lambda}_1(t), \hat{\lambda}_2(t), \hat{\lambda}_3(t), \ldots$ converges pointwise to $\lambda(t)$ as $i \to \infty$. □

By Theorems 1 and 2, above we see that as evidence accrues and the partition mesh shrinks, our estimation of each piece of $\lambda(t)$ approximates the original function arbitrarily closely.

4. Credible intervals for the mean rate

4.1. Adjusting credible intervals for non-unit interval lengths. After inferring the estimator $\hat{\lambda}_T$ for some interval T, we will often want to compute credible intervals for our estimate. As mentioned in Subsection 2.7, we can use the inverse CDF for the Gamma function to do this. However, the inverse CDF by itself will produce the correct results only when $|T| = 1$. To find the bounds for the credible interval for the estimated mean value of the intensity function $\lambda(t)$, we must correct the bounds by dividing by $|T|$.

4.2. An upper bound for the number of realizations needed for a desired credible interval. While we have shown that $\hat{\lambda}_T$ is a consistent estimator of $\overline{\lambda}_T$, we have not yet discussed how many realizations will need to be observed in order to produce credible intervals of a needed size. Estimating this is beyond the scope of this paper. Instead, we will use Chebyshev's inequality to produce an upper bound for the number of realizations required. As we saw in Section 1.6, Chebychev's inequality guarantees a credible interval no larger than $\sqrt{20}$ standard deviations from the mean. Since the variance of the Gamma distribution is given by $\sigma^2 = \frac{\alpha}{\beta^2} = \frac{\lambda}{\beta}$, we see that $\sigma = \frac{\sqrt{\lambda}}{\sqrt{\beta}}$.

We wish to ensure the error is less than some positive ϵ with a probability of at least 0.95. Thus we need $\epsilon \geq \left|\hat{\lambda}_T - \overline{\lambda}_T\right| \geq \sqrt{20}\sigma = \frac{\sqrt{20}\sqrt{\lambda}}{\sqrt{\beta}}$. Solving this equation

for β we get

$$\epsilon \geq \frac{\sqrt{20\lambda}}{\sqrt{\beta}} \iff \frac{\sqrt{\beta}}{\sqrt{20\lambda}} \geq \frac{1}{\epsilon} \iff \sqrt{\beta} \geq \frac{\sqrt{20\lambda}}{\epsilon} \iff \beta \geq \frac{20}{\epsilon^2}\lambda.$$

This formula gives an upper bound for the number of realizations of the process that must be observed in order to guarantee the desired level of credibility. One feature of the formula that should be noted is that the number of realizations scales linearly with the mean rate of the process on the subinterval.

However, frequently we are more concerned with the amount of error as a proportion, rather than a simple amount. For example, an error of $\left|\hat{\lambda}_T - \overline{\lambda}\right| = 1$ might be seen as miniscule, even negligble, if $\overline{\lambda}_T = 5000$ but unacceptably large if $\overline{\lambda}_T = 5$. Thus we may prefer to think in terms of the proportion of the error, making ϵ a function of λ.

If so, we happily find that the number of required realizations *decreases* as λ increases. If we set $\epsilon = \frac{\lambda}{c}$ for some $c > 1$, we have

$$\beta \geq \frac{20}{\epsilon^2}\lambda = \frac{20}{(\lambda/c)^2}\lambda = \frac{20c^2}{\lambda^2}\lambda = \frac{20c^2}{\lambda}.$$

As a concrete example, consider the case of $\epsilon = \lambda/10$, allowing a 10% error. Plugging this into the formula above, we get $\beta \geq \frac{2000}{\lambda}$. Thus for $\lambda = 1$, we may be required to observe as many as 2000 realizations to reach the desired level of credence, while for $\lambda = 10$, we will require at most 200.

5. Examples of inference

In this section we consider some examples of using the process outlined in Section 3. All examples are fictitious and used for illustrative purposes only.

5.1. Example 1. Alice's Deli is currently open Monday through Friday between the hours of 11 am and 9 pm. The owner, Alice, is considering restricting operation to lunch and dinner hours only in order to reduce operating costs. To aid in determining the optimal hours in which to be open, Alice decides to estimate the number of customers she receives during each hour of the day.

After some research, Alice decides to model the arrival of customers as a non-homogeneous Poisson process. She also decides to use Bayesian inference in order to better accommodate prior information available to her in the form of sales records, using Gamma distribution priors. Having made these decisions, she keeps careful track of the number of customers arriving over the course of a five-day work week.

Alice lets $S = 10$ be the length in hours of the relevant period (a single working day) and chooses the partition $\mathcal{P} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, with intervals corresponding to each hour the restaurant is open. (Note that for statistical purposes the time $t = 0$ corresponds to 11 am, while $t = 10$ corresponds to 9 pm.) After perusing three weeks (fifteen days) of sales records, Alice selects the prior parameters in Table 1 for each interval (note that the prior β is the same in every case because 15 days of data have been collected).

Interval	Prior α	Prior β	Prior Probability Distribution
$(0, 1]$	294	15	Gamma $(\alpha = 294, \beta = 15)$
$(1, 2]$	550	15	Gamma $(\alpha = 550, \beta = 15)$
$(2, 3]$	258	15	Gamma $(\alpha = 258, \beta = 15)$
$(3, 4]$	66	15	Gamma $(\alpha = 66, \beta = 15)$
$(4, 5]$	40	15	Gamma $(\alpha = 40, \beta = 15)$
$(5, 6]$	175	15	Gamma $(\alpha = 175, \beta = 15)$
$(6, 7]$	499	15	Gamma $(\alpha = 499, \beta = 15)$
$(7, 8]$	686	15	Gamma $(\alpha = 686, \beta = 15)$
$(8, 9]$	318	15	Gamma $(\alpha = 318, \beta = 15)$
$(9, 10]$	140	15	Gamma $(\alpha = 140, \beta = 15)$

Table 1. Example 1: Prior Probabilities

Alice's experiment resulted in the data in Table 2. Alice then uses this information to calculate the posterior according to Theorem 1 in Section 3. The results are given in Table 3 and portrayed graphically in Figure 2, with μ denoting the expected value of the prior or posterior, as appropriate. The credible intervals were computed from the CDF with R.

With this information, Alice is able to now apply various techniques to determine the expected cost and benefit of operating at certain hours of the day. If for instance, the cost of operation requires ten customers per hour on average to break even, then it would almost certainly be beneficial to close during the intervals $(3, 4]$ and $(4, 5]$, corresponding to 2pm to 3pm and 3pm to 4pm. In addition, it would probably be beneficial to close an hour earlier, during the interval $(9, 10]$ (or 8pm to 9pm), but we see that we cannot yet rule out the possibility that the rate during that interval is 10 or greater, at least at a credible level of 0.95.

5.1.1. Accuracy of the Estimates. Suppose now that the true mean rates $\overline{\lambda}$ for each interval are as in Table 4. The table also shows the error as a percent, calculated as $\overline{\lambda}/|\overline{\lambda} - \mu|$. As we have seen, if additional accuracy and confidence is needed, it could be gained through gathering more data.

Time Interval	# of Customers Mon	Tues	Wed	Thurs	Fri	Totals
$(0, 1]$	21	18	29	19	18	105
$(1, 2]$	28	43	29	32	34	106
$(2, 3]$	21	21	19	19	17	97
$(3, 4]$	4	4	2	4	8	22
$(4, 5]$	0	1	2	4	1	8
$(5, 6]$	6	10	21	10	8	55
$(6, 7]$	24	34	36	42	42	178
$(7, 8]$	48	47	42	52	48	237
$(8, 9]$	20	22	17	17	16	92
$(9, 10]$	4	6	11	9	7	37

Table 2. Example 1: Total Arrivals by Hour

Interval	n	k	Posterior	μ	Credible Interval ($p = 0.95$)
$(0, 1]$	105	5	Gamma$(399, 20)$	19.95	$(18.04, 21.95)$
$(1, 2]$	166	5	Gamma$(716, 20)$	35.80	$(33.23, 38.57)$
$(2, 3]$	97	5	Gamma$(355, 20)$	17.75	$(15.95, 19.64)$
$(3, 4]$	22	5	Gamma$(88, 20)$	4.40	$(3.53, 5.37)$
$(4, 5]$	8	5	Gamma$(48, 20)$	2.40	$(1.77, 3.13)$
$(5, 6]$	55	5	Gamma$(230, 20)$	11.50	$(10.06, 13.03)$
$(6, 7]$	178	5	Gamma$(677, 20)$	33.85	$(31.35, 36.45)$
$(7, 8]$	237	5	Gamma$(923, 20)$	46.15	$(43.22, 49.17)$
$(8, 9]$	92	5	Gamma$(410, 20)$	20.50	$(18.56, 22.53)$
$(9, 10]$	37	5	Gamma$(177, 20)$	8.85	$(7.59, 10.20)$

Table 3. Example 1: Posterior Probabilities

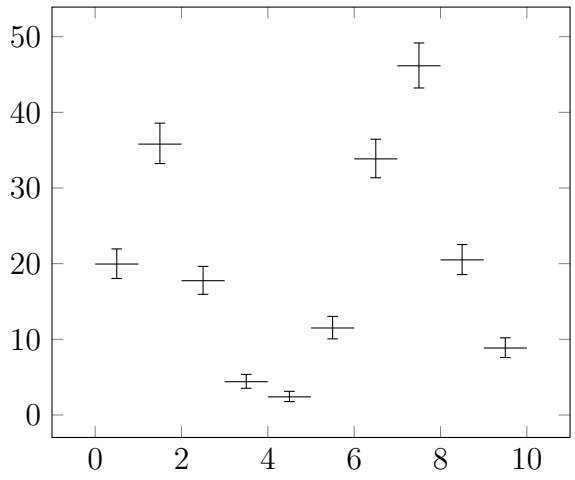


FIG. 2. Example 1: Estimated means with error bars

Interval	$[0,1)$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$	$(5,6]$	$(6,7]$	$(7,8]$	$(8,9]$	$(9,10]$
$\overline{\lambda}$	20	36	17	5	3	12	32	45	22	8
μ	19.95	35.80	17.75	4.40	2.40	11.50	33.85	46.15	20.50	8.85
$\lvert\overline{\lambda} - \mu\rvert$	0.05	0.20	0.75	0.60	0.20	0.50	1.85	1.15	1.50	0.85
% Error	0.25	0.56	4.41	12	6.67	4.17	5.78	2.56	6.82	10.63

Table 4. Example 1: Accuracy of estimates

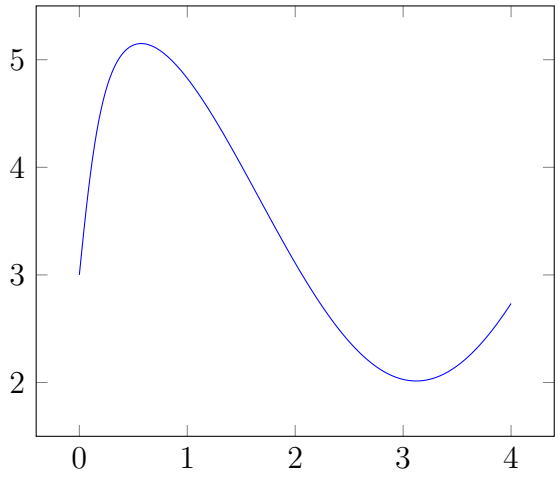


FIG. 3. Example 2: True Intensity Function

5.2. Example 2. Suppose an NHPP has a period length of 4 and the intensity function $\lambda(t) = 2\arctan(5t) + 2\cos(t) + 1$. This function is pictured in Figure 3. We observe 5 realizations of the process, yielding the set of arrivals in Table 5. We will explore multiple ways to partition the data and the results obtained.

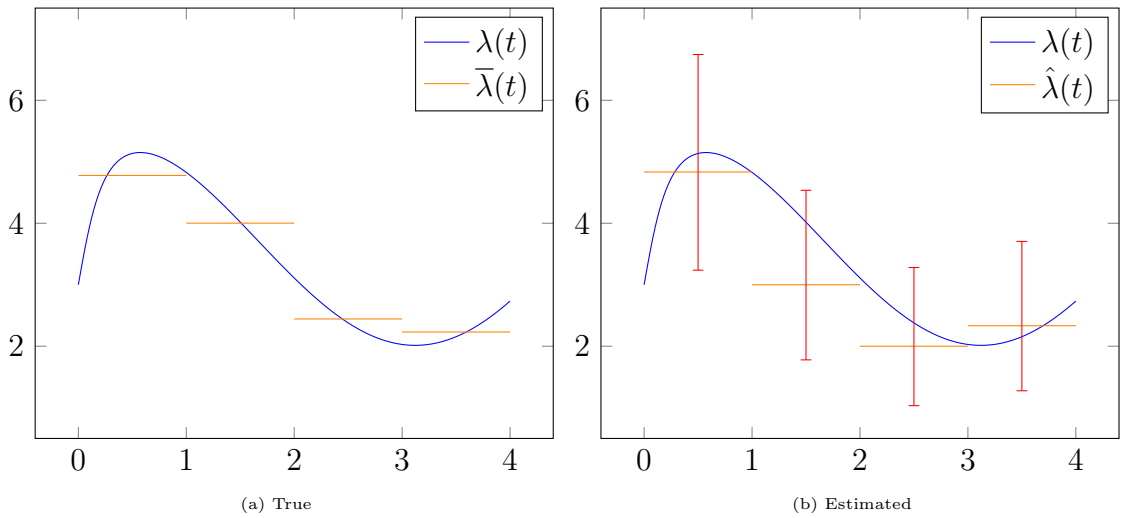
5.2.1. Hourly: $\mathcal{P}_1 = \{0,1,2,3,4\}$. Let the partition $\mathcal{P}_1 = \{0,1,2,3,4\}$. We can calculate that the true mean values for each interval are given approximately in Table 6. On the five realizations, the number of arrivals in each subinterval of the partition is shown is Table 7. Using the inference technique from Theorem 1 with a diffuse prior Gamma$(1,1)$ for each interval of the partition, we get the parameters in Table 8. Figure 4 displays and contrasts (a) the true mean function $\overline{\lambda}(t)$ with (b) the estimator function $\hat{\lambda}(t)$.

0.008316505	0.057921384	0.082521686	0.133753605	0.182678425
0.241203225	0.264257218	0.275747027	0.289507543	0.337747490
0.337959160	0.388364928	0.460051573	0.531286455	0.584262207
0.586118157	0.619313083	0.642201824	0.672972928	0.692396975
0.756105268	0.768330402	0.805124826	0.811005894	0.836802410
0.854745691	0.858150426	0.868232123	1.024238145	1.087531347
1.146256671	1.182951448	1.242949557	1.270897330	1.273996290
1.325691394	1.344078714	1.466342488	1.483802438	1.559800381
1.741699974	1.821997349	1.847895434	1.874128693	1.960376231
2.113724898	2.216344420	2.328607138	2.348401578	2.348995085
2.382247736	2.611295646	2.734642469	2.781271486	2.872947194
2.972900378	3.037728378	3.042038019	3.056092888	3.092219570
3.195661224	3.242381787	3.399784693	3.571915606	3.696918089
3.717317319	3.868671239	3.973580122	3.993450447	

Table 5. Example 2: Arrivals

Interval	$\overline{\lambda}(t)$
$(0, 1]$	4.778124
$(1, 2]$	4.001957
$(2, 3]$	2.443421
$(3, 4]$	2.230804

Table 6. Example 2: Hourly Means



(a) True (b) Estimated

FIG. 4. Example 2: Hourly Means

Interval	Arrivals
$(0, 1]$	28
$(1, 2]$	17
$(2, 3]$	11
$(3, 4]$	13

Table 7. Example 2: Hourly Arrivals

Interval	Posterior α	Posterior β	μ	Credible Interval
$(0, 1]$	29	6	4.833333	(3.236959, 6.744633)
$(1, 2]$	18	6	3.000000	(1.777990, 4.536441)
$(2, 3]$	12	6	2.000000	(1.033429, 3.280340)
$(3, 4]$	14	6	2.333333	(1.275655, 3.705066)

Table 8. Example 2: Hourly Posteriors

5.2.2. Every quarter hour. Define the partition

$$\mathcal{P}_2 = \{0.00, 0.25, 0.50, \ldots, 3.50, 3.75, 4.00\}.$$

This partitions the period into intervals of 15 minutes. Then the true mean values are given approximately in Table 9, and the number of arrivals observed in the experiment on each interval are given in Table 10. We use Bayesian inference as per Theorem 1 to estimate the means, and the estimations are given in Table 11. Finally, Figure 5 displays the true means and estimated means graphically.

5.2.3. 10000 realizations. For this section we generated 10000 realizations in order to demonstrate the effect of reducing the size of the partition. The results can be seen in Figure 6. We can see from the figure the effect of decreasing the length of the partition subintervals from (a) 1 hour each to (b) 15 minutes each, (c) 10 minutes each, and (c) 5 minutes.

Interval	$\overline{\lambda}(t)$
$(0, 0.25]$	4.018556
$(0.25, 0.50]$	4.993207
$(0.50, 0.75]$	5.132744
$(0.75, 1.00]$	4.967990
$(1.00, 1.25]$	4.648438
$(1.25, 1.50]$	4.240020
$(1.50, 1.75]$	3.788125
$(1.75, 2.00]$	3.331245
$(2.00, 2.25]$	2.903903
$(2.25, 2.50]$	2.536606
$(2.50, 2.75]$	2.254902
$(2.75, 3.00]$	2.078271
$(3.00, 3.25]$	2.019178
$(3.25, 3.50]$	2.082454
$(3.50, 3.75]$	2.265091
$(3.75, 4.00]$	2.556493

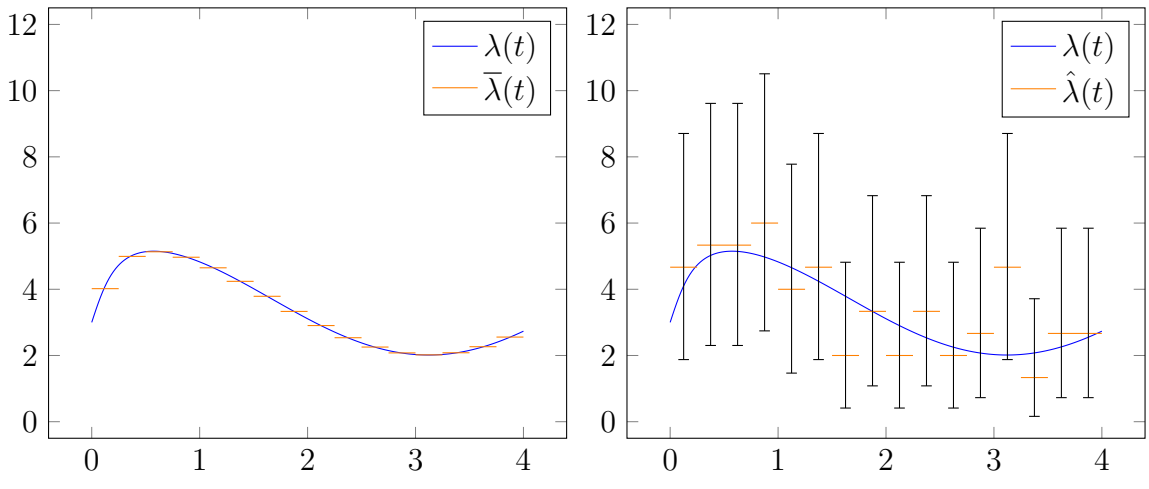
Table 9. Example 2: True means (Quarter hourly)

Interval	Arrivals
$(0, 0.25]$	6
$(0.25, 0.50]$	7
$(0.50, 0.75]$	7
$(0.75, 1.00]$	8
$(1.00, 1.25]$	5
$(1.25, 1.50]$	6
$(1.50, 1.75]$	2
$(1.75, 2.00]$	4
$(2.00, 2.25]$	2
$(2.25, 2.50]$	4
$(2.50, 2.75]$	2
$(2.75, 3.00]$	3
$(3.00, 3.25]$	6
$(3.25, 3.50]$	1
$(3.50, 3.75]$	3
$(3.75, 4.00]$	3

Table 10. Example 2: Arrivals (Quarter hourly)

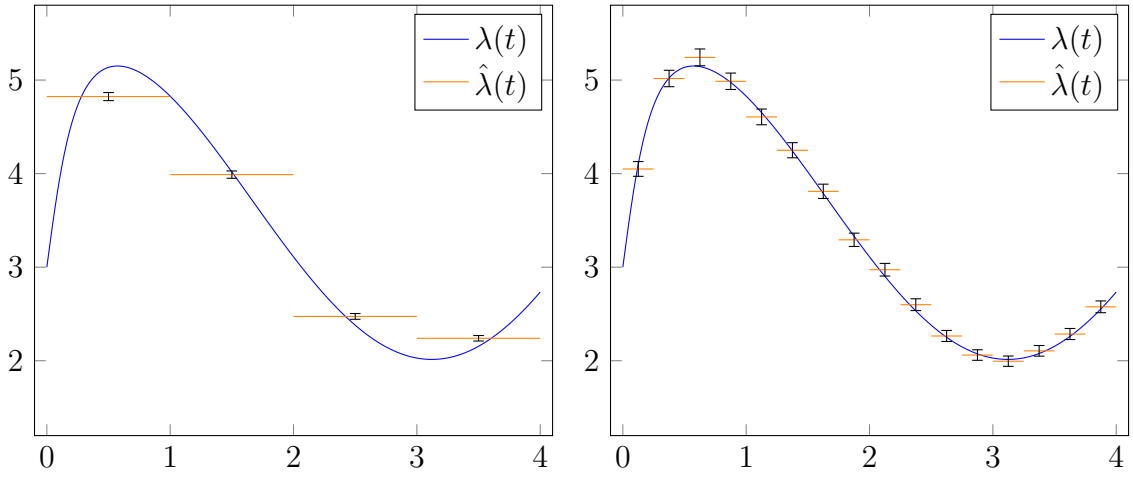
Interval	Posterior α	Posterior β	μ	Credible Interval ($p = 0.95$)
$(0.00, 0.25]$	7	6	4.6666667	$(1.8762420, 8.7063160)$
$(0.25, 0.50]$	8	6	5.3333333	$(2.3025548, 9.6151169)$
$(0.50, 0.75]$	8	6	5.3333333	$(2.3025548, 9.6151169)$
$(0.75, 1.00]$	9	6	6.0000000	$(2.7435821, 10.5087928)$
$(1.00, 1.25]$	6	6	4.0000000	$(1.4679295, 7.7788881)$
$(1.25, 1.50]$	7	6	4.6666667	$(1.8762420, 8.7063160)$
$(1.50, 1.75]$	3	6	2.0000000	$(0.4124481, 4.8164584)$
$(1.75, 2.00]$	5	6	3.3333333	$(1.0823243, 6.8277258)$
$(2.00, 2.25]$	3	6	2.0000000	$(0.4124481, 4.8164584)$
$(2.25, 2.50]$	5	6	3.3333333	$(1.0823243, 6.8277258)$
$(2.50, 2.75]$	3	6	2.0000000	$(0.4124481, 4.8164584)$
$(2.75, 3.00]$	4	6	2.6666667	$(0.7265769, 5.8448487$
$(3.00, 3.25]$	7	6	4.6666667	$(1.8762420, 8.7063160)$
$(3.25, 3.50]$	2	6	1.3333333	$(0.1614729, 3.7144289)$
$(3.50, 3.75]$	4	6	2.6666667	$(0.7265769, 5.8448487)$
$(3.75, 4.00]$	4	6	2.6666667	$(0.7265769, 5.8448487)$

Table 11. Example 2: Posteriors (Quarter hourly)



(a) True (b) Estimated

FIG. 5. Example 2: Quarter hourly means



(a) 1 hour

(b) 15 minutes

(c) 10 minutes

(d) 5 minutes

FIG. 6. Example 2: Estimates inferred from 10000 simulated realizations

6. Conclusion

In this paper we have shown that we are able to use Bayesian inference to reconstruct the intensity function for a non-homogeneous Poisson process to any desired resolution and degree of accuracy. This result requires the choice of a Gamma distribution for the prior probability distribution, but it is indifferent to the choice of initial parameters for the Gamma distribution. Using Chebyshev's inequality we set an upper bound for the number of realizations of the process we would need to observe in order to established the needed accuracy and confidence.

Appendix A. R Code

In this section we present algorithms and R code for simulating the process of observing realizations of an NHPP and demonstrate the effectiveness of the approach outlined in Section 3. The code is adapted from the algorithm in (Leemis, 1991).

Realization generation. First we will consider how to simulate the observation of a realization of an NHPP with a given intensity function $\lambda(t)$ defined on the interval $[0, S]$. Let $\Lambda(t) = \int_0^t \lambda(\tau)\, d\tau$ be the cumulative intensity function (CIF), and $\Lambda^{-1}(t)$ be the inverse of the CIF. After the algorithm is performed, N will be the number of events occuring during the realization and E_i $(1 \leq i \leq N)$ will be the time of occurence of each event (not necessarily in order).

1. $\overline{\lambda} \leftarrow \Lambda(S)$
2. $N \sim \text{Poisson}\left(\overline{\lambda}\right)$
3. For $i \leftarrow 1$ to N:
    1. $Y_i \sim \text{Uniform}\left(0, \overline{\lambda}\right)$
    2. $E_i \leftarrow \Lambda^{-1}\left(Y_i\right)$

The following code implements the above algorithm in the R language:

```
#' Generate a random realization from a given intensity
#' function
#'
#' @param intensityFn Intensity function
#' @param lengthOfRealization Length of a single period.
#'
```

```
#' @return Data frame of events
generateRealization <- function (
    intensityFn,
    lengthOfRealization,
    ...,
    integrationSubdivisions = 100L
) {
  vif <- Vectorize(intensityFn)

  # Integrate the intensity function to get the cumulative
  # intensity function
  cif <- function(t) {
    result <- integrate(vif, 0, t,
      subdivisions=integrationSubdivisions)
    return(result$value)
  }

  # Find the expected rate of events across the entire
  # realization
  rate <- cif(lengthOfRealization)

  # Take the inverse of the cumulative intensity function
  cif.inv <- Vectorize(inverse(cif, 0, rate))

  # Generate a realization
  numEvents <- rpois(1, rate)
```

```
    yValues <- runif(numEvents, 0, rate)
    events <- cif.inv(yValues)
    return(data.frame(t = events))
  }
```

Inference of intensity function. The folllowing R code implements the inference of the posterior:

```
  #' Infer the intensity function of a non-homogeneous
  #' Poisson process.
  #'
  #' @param events list of time of event occurences on the
  #'                process
  #' @param numRealizations number of realizations observed
  #' @param partition
  #' @param priorScales
  #' @param priorRates
  #' @param confidenceLevel
  #'
  #' @return data frame with estimates
  #'
  inferIntensity <- function (
      events,
      numRealizations,
      partition,
      priorScales,
      priorRates,
```

```
    ...,
    confidenceLevel = 0.05)
{
  numIntervals <- length(partition) - 1
  intervalLengths <- tail(partition, numIntervals) -
    head(partition, numIntervals)

  intervalEvents <- split(events, cut(events, partition))
  intervalEvents.counts <- lengths(intervalEvents)

  intervalScales <- intervalEvents.counts + priorScales
  intervalRates <- priorRates + numRealizations

  intervalEstimates <- ((intervalScales / intervalRates)
    / intervalLengths)

  q_lower <- qgamma(confidenceLevel / 2, intervalScales,
    intervalRates) / intervalLengths

  q_upper <- qgamma(1 - (confidenceLevel / 2),
    intervalScales, intervalRates) / intervalLengths

  intervals <- data.frame(scale = intervalScales,
    rate = intervalRates,
    mean = intervalEstimates,
    lower = q_lower,
```

```
      upper = q_upper)


  return(intervals)
}
```

References

Çınlar, E. (1975). *Introduction to stochastic processes*. Prentice-Hall.

Donovan, T., & Mickey, R. (2019). *Bayesian statistics for beginners: A step-by-step approach*. Oxford University Press.

Hájek, A. (2019). Interpretations of Probability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/`.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Elsevier Science.

Lambert, B. (2018). *A student's guide to bayesian statistics*. SAGE Publications. Retrieved from `https://books.google.com/books?id=CLZBDwAAQBAJ`

Leemis, L. M. (1991). Nonparametric estimation of the cumulative intensity function for a nonhomogeneous poisson process. *Management Science*, *37*(7), 886–900.

Mattuck, A. (1999). *Introduction to analysis*. Prentice Hall.

Ross, S. (2006). *Introduction to probability models*. Elsevier Science.

Stone, J. (2013). *Bayes' rule: A tutorial introduction to bayesian analysis*. Sebtel Press. Retrieved from `https://books.google.com/books?id=dpZRdk6-EN8C`

Wijono, W. (2022). *Don't use uniform priors.* Retrieved from `https://towardsdatascience.com/stop-using-uniform-priors-47473bdd0b8a`

Winkler, R. (1972). *An introduction to bayesian inference and decision*. Holt, Rinehart and Winston.