IMPERIAL

MSci Research Project, Department of Mathematics

Thinned Spatial Point Processes for Confidence Interval Formation

by

Lucinda Khalil
01717289

2024

Declaration

This project report is submitted in partial fulfilment of the requirements for the degree of MSci Mathematics, Imperial College London. I declare that this thesis is my original work, except where stated otherwise. This thesis has never been submitted for any degree or examination to any other University.

Lucinda Khalil
01717289

This thesis was conducted under the supervision of Dr. Cohen.

Abstract

Throughout this project, we present a novel resampling method, called thinning. We use this method to form confidence intervals for statistics estimated from a set of observations, and focus on spatial point processes. A thinned sample is a sample where each point in the observed sample is retained with a predetermined probability, p, and otherwise discarded. We prove that for some desired statistic, θ, with estimate $\hat{\theta}$ and thinned replication $\hat{\theta}_s$, that the normalised distribution of $\hat{\theta}_s$ approaches that of $\hat{\theta}$ assuming the underlying data is i.i.d and θ is a linear statistical functional. This means that with a significance level of α, the studentized confidence interval contains the true value about $(1-\alpha)\%$ of the time. We demonstrate this result numerically in the setting of probability distributions. We see that the theory extends to the spatial setting, in the case that the statistic is a homogeneous or inhomogeneous intensity function of a Poisson point process. However, the desired result does not hold when estimating Ripley's K, which is a measure of the amount of clustering in a process. We conclude that similarly to the jackknife and bootstrap, thinning does not adequately preserve the interactions between dependent points, and therefore is not useful for all spatial statistics, although it does provide accurate results for some.

The code used to carry out experiments throughout this research can be found here.

Contents

List of Figures

Chapter 1

Introduction

The aim of this project is to present and explore a novel approach to confidence interval formation, focusing on the spatial setting. The estimate of a statistic is only useful if we can assign an uncertainty or confidence to it, which of course becomes challenging when we do not know the true underlying distribution of the statistic, and have limited observed data. When working with one-dimensional probability distributions, the problem of forming confidence intervals has been explored extensively in the literature, with various resampling techniques being presented with many adaptations. The most popular of these include the jackknife (Quenouille, 1949, 1956; Tukey, 1958; Efron and Tibshirani, 1993; Wu, 1986; Berger and Skinner, 2004; Hinkley, 1977), and more successfully, the bootstrap (Efron, 1979, 2012; Künsch, 1989; Wang, 1995).

Spatial statistics is the study of multidimensional data. In particular, a spatial point pattern is a set of points within a region which arises from an underlying stochastic mechanism. The methods used for probability distributions cannot be applied analogously to spatial data, with the most prominent reason for this being that we can no longer assume that the data is independently and identically distributed in the same way that we can for probability distributions (Cressie, 1991; Singh, 1981). The addition of interactions between points in spatial data means that the classical resampling techniques result in a change in the underlying process since they involve the addition and/or removal of pairs of points.

There have been various adaptations of bootstrapping to the higher dimensional setting which attempt to preserve the structure of the interactions. Of these, the block bootstrap and its counterparts are the most popular (Loh and Stein, 2004; Hall, 1985a, 1988; Mattfeldt et al., 2013; Borrajo et al., 2020). However, these methods are shown to have flaws and rarely produce confidence intervals with the desired confidence levels, specifically in the case of second-order statistics such as Ripley's K as detailed in Loh and Stein (2004). Intuitively, Ripley's K is a measure of the amount of clustering or regularity in a point process, and therefore is highly affected if the interactions within the data change. For this reason, it is particularly difficult to form useful confidence intervals on this statistic. Solving this problem would be very valuable, since Ripley's K is used in a wide range of fields including biology, archaeology, and geography, among many more (Kiskowski et al., 2009; Jafari Mamaghani

et al., 2010; Hamer et al., 2019; Negre et al., 2018; Ge et al., 2019). Although an estimator of the variance of K exists for the most simple point process, analogous expressions are very difficult or even impossible to obtain in more complex settings, meaning a resampling procedure must be used (Loh, 2010).

We propose an alternative method, thinning. Thinning consists of generating samples from observed data by choosing to retain or discard each point with some predetermined probability, p, which we will call the thinning parameter. This method differs from the block bootstrap in the fact that it does not introduce new interactions by placing pairs of points together. We will provide an analytical and numerical exploration of the efficacy of thinning, on both probability distributions, and on spatial point processes.

We will begin by discussing preliminary theory, consisting mostly of spatial statistics in Chapter 2. In Chapter 3 we will discuss the current bootstrapping methods in greater detail, and formally present thinning. In Chapter 4, we will prove that thinned samples are useful for obtaining confidence intervals on distributions and we will demonstrate this numerically. We will discuss and compare various ways to formulate confidence intervals in Chapter 5 and apply these to Ripley's K and inhomogenous intensities in Chapters 6 and 7 respectively.

Contributions

The overall aim of this work is to present an analytic and numerical exploration of thinning. The resulting contributions consist of the following:

(1) Introducing the novel resampling approach of thinning for the purpose of confidence interval formation.

(2) Proving that the asymptotic distribution of the normalised thinned replications approaches that of the desired estimator under certain conditions, and discuss which of these conditions are necessary.

(3) Explicitly showing this result holds in the higher dimensional spatial setting with respect to an arbitrary homogeneous intensity of a Poisson point process.

(4) Demonstrating the asymptotic distribution of the thinned replications in various settings including for probability distributions, and spatial point processes.

(5) Proposing and proving that thinning is equivalent to Efron's bootstrap in certain settings.

(6) Exploring various types of confidence interval formations, analytically and numerically demonstrating that they do/do not produce the desired result.

(7) Applying the novel method to inhomogeneous intensity and Ripley's K estimates of spatial point processes.

Chapter 2

Background Theory

In this chapter, we discuss preliminary material which will be useful for understanding the content of this project. This mainly consists of spatial point processes, and some statistics and estimators of these.

As described in Diggle (2013), 'a spatial point pattern is a set of locations, irregularly distributed withing a designated region and presumed to have been generated by some form of stochastic mechanism.' We will focus in particular on the intensity, meaning the number of points we expect to see within some subregion, and the amount of clustering of these point processes.

The study of spatial point processes is referred to in the literature as a subset of spatial statistics. Spatial statistics is a relatively young field aimed at summarising and analysing data which has a position assigned to each data point. Throughout this project, we consider data which is assumed to be generated from point processes in two dimensions, although there are extensions to this such as spatial processes indexed over continuous space, called spatial processes, or over lattices, called lattice processes (Cressie, 1991). Note that throughout the project we also only consider *simple* point processes where the probability of more than one event occurring at a given position is almost surely zero.

Point processes emerged due to a need to model observed data in areas such as biology, physics, geography and geology among others (Ripley, 1981). Some examples of point processes which are important for this research are the Neyman-Scott process, as described in Neyman and Scott (1958); Stoyan et al. (1995), and the Soft Core process described in Stoyan (1987).

2.1 Spatial Point Processes

2.1.1 Definition of Point Processes

Point processes arise over some set X, in our case some subset of $\mathbb{R}^2$. We can extend this idea to a marked point process. This is a process where each point $x \in \mathbb{R}^2$ is assigned a mark $z \in Y$, then we can think of this new marked point process as a point process on the

product space $X = \mathbb{R}^2 \times Y$ (Cressie, 1991). To define a well-behaved point process, we first introduce *locally finite point configurations*:

Definition 1. *(van Lieshout, 2019) A family $N^{lf}(\mathbb{R}^d)$ is a family of* **locally finite point configurations** *in $\mathbb{R}^d$ if it consists entirely of subsets $x \subset \mathbb{R}^d$ which place finitely many points in every bounded Borel set $A \subset \mathbb{R}^d$.*

This ensures that each Borel set contains at most a countably infinite number of points. We can now define a point process:

Definition 2. *(van Lieshout, 2019) A* **point process***, $X \in N^{lf}(\mathbb{R}^d)$, on $\mathbb{R}^d$ is a random locally finite point configuration, such that for all bounded Borel sets $A \subset \mathbb{R}^d$, the number of points of X that fall in A is a finite random variable which is denoted as $N_X(A)$.*

2.1.2 Properties

We now define some important properties of spatial point processes.

Definition 3. *(van Lieshout, 2019; Ibe, 2013) A* **homogeneous** *(or analogously stationary) spatial point process is one whose distribution is invariant under translation. Given the bounded Borel sets $B_1, B_2, ..., B_k \in \mathbb{R}^d$ for some $k \in \mathbb{Z}_+$, the joint distribution of $N(B_1), N(B_2), ..., N(B_k)$ is equal to the joint distribution of $N(B_1+y), N(B_2+y), ..., N(B_k+y)$ for an arbitrary y:*

$$P(N(B_1 + y) \leq n_1, ..., N(B_k + y) \leq n_k) = P(N(B_1) \leq n_1, ..., N(B_k) \leq n_k).$$

Definition 4. *An* **isotropic** *spatial point process is one whose distribution is invariant under rotation.*

Definition 5. *(Cressie, 1991) A point process exhibits* **complete spatial randomness** *when, conditioned on the number of points in the bounded region A, $N(A)$, the points are independently and uniformly distributed over the region. This means that if $N(A) = n$, then the ordered tuple of points $(s_1, s_2, ..., s_n)$ satisfies*

$$P(s_1 \in B_1, s_2 \in B_2, ..., s_n \in B_n) = \prod_{i=1}^{n} \frac{|B_i|}{|A|}, \qquad B_1, B_2, ..., B_n \subset A$$

where $|B| = \int_B ds$.

Complete spatial randomness (CSR) is present in a process when points are equally likely to occur anywhere within a given bounded set and the points do not interact repulsively or attractively. A homogeneous Poisson process exhibits complete spatial randomness. One statistic useful in measuring CSR is Ripley's K which is discussed further in Section 2.2 (Cressie, 1991; Shekhar et al., 2005).

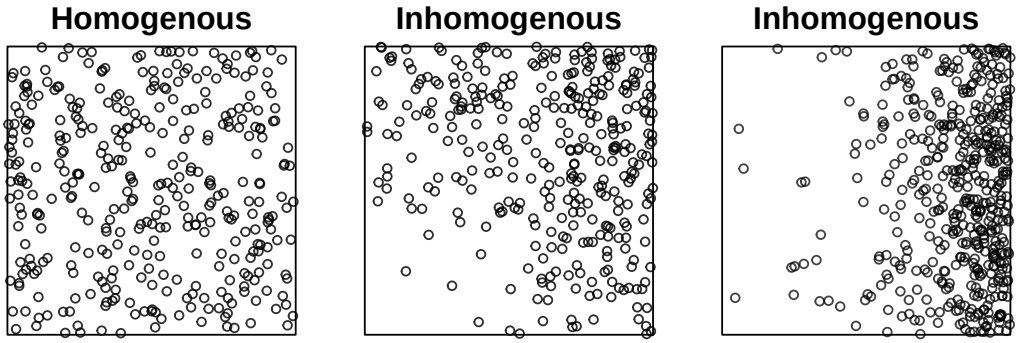

Figure 2.1: Plots showing Poisson point processes on $\mathbb{R}^2$ simulated on a unit square, with various intensities. The first plot is a homogeneous process with an intensity of $\lambda(x, y) = 350$. The second and third plot have inhomogenous intensities of $\lambda(x, y) = 500\sqrt{x^2 + y^2}$ and $\lambda(x, y) = 15\exp(5x)$ respectively.

2.1.3 Poisson Point Process

A homogeneous Poisson point process is a type of point process where each point is generated independently of the others, and there is no interaction between them. It is considered to be equivalent to complete spatial randomness as introduced in Definition 5, and is the 'white noise' of spatial point processes (Cressie, 1991). These point processes are very useful in modelling events we see in the real world since they can be described as the limit of a binomial point processes as the limit of the number trials goes to infinity (van Lieshout, 2019). A homogeneous Poisson point process is defined in van Lieshout (2019) in the following way:

Definition 6. *(van Lieshout, 2019) A* **homogenous Poisson point process**, *X, is a point process on $\mathbb{R}^d$ with intensity $\lambda > 0$ if*

- *$N_X(A)$ is Poisson distributed with mean $\lambda|A|$ for every bounded Borel set $A \subset \mathbb{R}^d$;*
- *for any k disjoint bounded Borel sets $A_1, A_2, ..., A_k$, $k \in \mathbb{N}$, the random variables $N_X(A_1), ..., N_X(A_k)$ are independent.*

This can be extended to an **inhomogenous Poisson point process** by allowing the intensity, λ, to vary over A, and therefore the mean will take the form

$$E\left(N_X(A)\right) = \int_A \lambda(x)dx$$

for some function $\lambda : \mathbb{R}^d \to \mathbb{R}^+$ which is integrable on bounded sets.

Figure 2.1 shows an example of a homogenous Poisson point process simulated on a unit square, along with two examples is inhomogenous processes, where the intensity is a non-constant function of the location.

2.1.4 Cluster Point Processes

Cluster point processes have a wide range of practical applications such as modelling failures of complex systems such as computers, ecology, and teletraffic to a server (Lewis, 1964;

Bartlett, 1964; Aryal and Jones, 2021; Faÿ et al., 2006). A cluster process, is itself a combination of two processes. The first of these is the parent process, N_p, which determines the positions of the cluster centres, and the second is the daughter process, N_d, which determines the positions of points in each cluster. Each point of N_p initiates a process N_d around it independently to generate the members of the cluster, therefore the clusters are each independent and identically distributed point processes. The final process is a superposition of all these resulting clusters, and the parent points can either be included or discarded (Westcott, 1971).

Neyman-Scott processes are a class of point processes which result in clustered spatial data. They were first developed to model the distribution of galaxies on the universe, something which had been done deterministically beforehand (Neyman and Scott, 1958). They are a special case of cluster processes where the parent distribution is homogenous Poisson of intensity λ_p, and the number of daughter points of each cluster is found randomly according to some distribution and are scattered independently and with identical spatial probability density function f about the parent. Also, the parent points are discarded. If f is radially symmetric, then the process is isotropic. Let $\bar{c}$ be the mean number of daughter points per parent, then the overall intensity, λ, of a Neyman-Scott process is

$$\lambda = \lambda_p \bar{c}.$$

Matérn Cluster Process

The Matérn cluster process, as described in Stoyan et al. (1995), is a specific case of the Neyman-Scott class of point processes. The positions of the parent points are of course generated with a Poission point process of intensity λ_p. Each daughter cluster has a number of points determined by a Poisson distribution with intensity λ_d, which are independently and uniformly distributed on on the ball $B(\mathbf{x}, R)$ centered around the respective parent, $\mathbf{x}$. R is a radius parameter which is determined. Therefore, the intensity of the overall process is

$$\lambda = \lambda_p \lambda_d.$$

An example of a Matérn processes is shown in Figure 2.2.

2.1.5 Regular Point Processes

Regular point processes are those which deter points from being within some distance of eachother, which in turn reduces clustering. They are useful in modelling distributed wireless networks and in forestry (ElSawy and Hossain, 2013; Stoyan and Penttinen, 2000).

Soft Core Process

One example of these processes is the Soft Core process which is a more general form of Matern's second hard-core process, see Matern (1986) for more details on this. The Soft

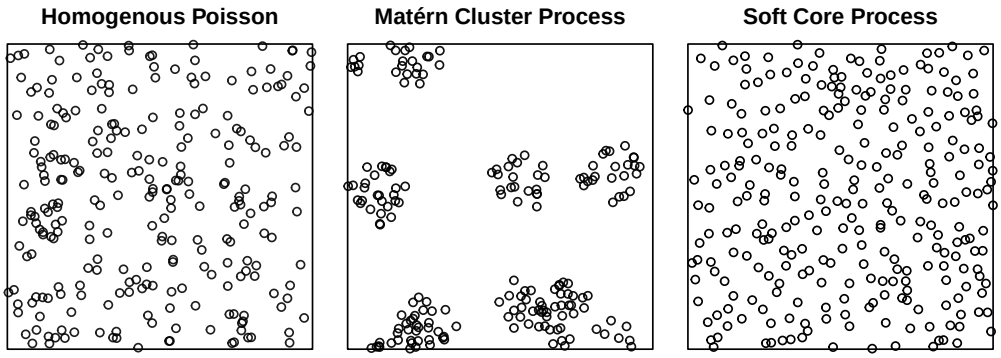

Figure 2.2: This Figure shows examples of various point processes discussed in this section, all of which are shown on a unit square. The first subplot shows a homogenous Poisson point process with an intensity of $\lambda = 250$. The second subplot shows a Matérn cluster process which is a type of Neyman-Scott process. This has a parent and daughter intensities of $\lambda_p = 10$ and $\lambda_d = 25$ respectively, and a cluster radius of $R = 0.1$. The final subplot is a regular Soft Core point process with $\lambda_b = 500$, and the radii, ρ, are assigned with a probability density function of $f(\rho) = 800\rho$ where $0 < \rho < 0.05$.

Core process, as introduced in Stoyan (1987), is simulated by first generating a stationary Poisson process of intensity λ_b. Each point is then assigned a radius according to some continuous distribution function, F. Then, each point is then considered is either kept or removed according to some criteria. Consider an arbitrary point x with radius $r(x)$. This point survives if and only if there is no other point y with radius $r(y)$ such that such that x is in a circle of radius $r(y)$ centred at y, and $r(y) \geq r(x)$. Alternatively, a mark can also be randomly generated for each point alongside the radius according to a uniform distribution on $[0, 1]$. Then, the point x survives iff there is no point y with radius y and mark $m(y)$ such that x is in a circle of radius $r(y)$ centred at y, and $m(y) \geq m(x)$. Other slight variations to this process can be seen in Stoyan (1987).

In all cases, the resulting process is stationary, isotropic and has an intensity, λ, of:

$$\lambda = \lambda_b p_s$$

where p_s is the probability of an arbitrary point from the initial Poisson process surviving. This probability is given by:

$$p_s = \int_0^1 \exp\left(-\lambda_b V(u)\right) du$$

where $V(u)$ is the volume of the set $M(o, u)$ which is three dimensional. For $x = (x_1, x_2)$ we have:

$$M(x, u) = \{(\xi, \eta, \zeta) : \sqrt{(\xi - x_1)^2 + (\eta - x_2)^2} \leq F^{-1}(\zeta), u \leq \zeta \leq 1\}.$$

An example of a Soft Core process is shown in Figure 2.2.

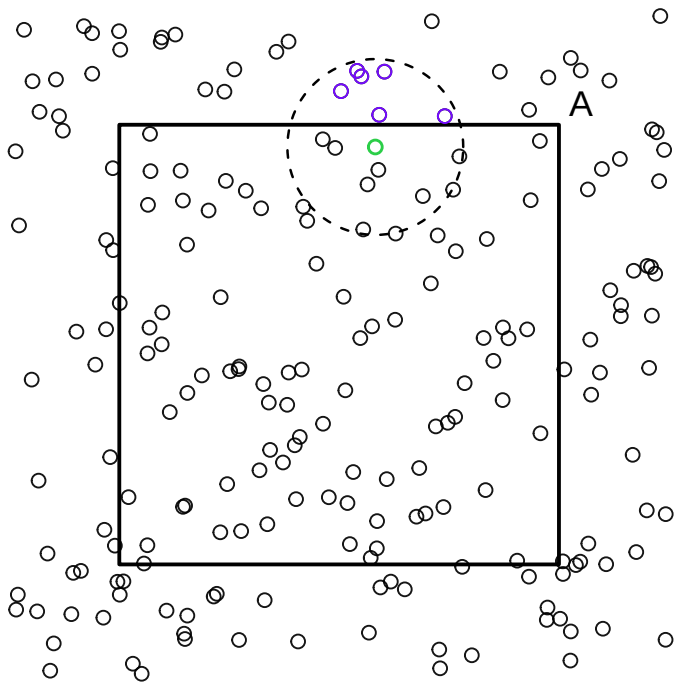

Figure 2.3: This Figure shows the flaw in the naive estimator of Ripley's K, 2.2. Assume we have observed the points in the unit square region A. We sum over each of the points inside A, counting how many other points are within a distance R. Consider the green point inside the region, which happens to be located at $(1.08, 1.45)$. If we set $h = 0.2$, then clearly, the circular region extends beyond the observed region, and the purple points which are outside of A but within the radius will not be included in the sum. Therefore, this estimator is biased for a bounded observation window.

2.2 Ripley's K

Ripley's K, sometimes known as the reduced second moment measure, is a distance based statistic which 'captures the spatial dependence between different regions of the point process'. Given some distance $h > 0$, Ripley's K is given as:

$$K(h) = \frac{E(C(x,h)|\text{point at } x)}{\lambda} \tag{2.1}$$

where $C(x,h)$ is the count of the number of points within a radius h of an arbitrary point x, not counting x itself. In other words, Ripley's K is the expected number of points within a radius h of an arbitrary point in the process, normalised by the intensity, λ (Cressie, 1991). Note that we are assuming that the intensity, λ, is a constant, and therefore that the point process is homogeneous.

2.2.1 Isotropic Estimator

Ripley's K of a point process can be estimated using an empirical average to replace the expectation. Let $(x_1, ..., x_n)$ be the locations of the n observed points over some region

$A \subset \mathbb{R}^d$. The simplest estimator then takes the following form:

$$\hat{K}(h) = \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{I(||x_i - x_j|| \leq h)}{n} \tag{2.2}$$

where $\hat{\lambda}$ is an estimate of the intensity. This estimator is unbiased if we assume all points which are within a distance h of the observed points are also observed. If this is not the case, then we have $E(\hat{K}(h)) < K(h)$, and therefore this estimator is biased (Ripley, 1988). This is illustrated in Figure 2.3.

Denote the circular region of radius R centered at an arbitrary point x as $B(x, R)$, and its boundary as $\partial B(x, R)$. The isotropic correction described in Ripley (1988) compensates for the issue presented for the estimator in Equation 2.2 by taking advantage of the isotropy of the point process, and counting pairs more than once where required. Consider a pair of points (x_i, x_j) for some $i, j \in \{1, ..., n\}$, where the boundary $\partial B(x_i, ||x_i - x_j||)$ is partially outside of the observation window. There could be unobserved points a distance of $||x_i - x_j||$ from x_i, and so to compensate for this we count the pair (x_i, x_j) more than once in the summation. We do this by multiplying each summand by a weight $w(x_i, x_j)$:

$$w(x_i, x_j) = \frac{|\partial B(x_i, ||x_i - x_j||)|}{|\partial B(x_i, ||x_i - x_j||) \cap A|}$$

which is essentially the inverse of the proportion of the boundary which lies within the observed region, A. Using these weights, an unbiased estimate can be formed

$$\hat{K}_{iso}(h) = \frac{|A|}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} I(||x_i - x_j|| \leq h) w(x_i, x_j) \tag{2.3}$$

where $|A|$ is the volume of the observed region.

The variance of Ripley's K with the isotropic correction is derived in closed form in Ripley (1988) as:

$$var(\hat{K}_{iso}(h)) = \frac{2|A|^2 \beta(h)}{n^2} (1 + 0.305\gamma(h) + \beta(h)(-1 + 0.0132n\gamma(h))) \tag{2.4}$$

where $\beta(h) = \frac{\pi h^2}{|A|}$ and $\gamma(h) = \frac{|\partial A| h}{|A|}$.

For the remainder of this report, we will discard the subscript and $\hat{K}(h)$ will refer to the Ripley's K estimate with an isotropic correction.

2.2.2 Ripley's K Under CSR

Assuming that a point process is completely spatially random in $\mathbb{R}^2$, such as a homogenous Poisson point process, we know that $K(h) = \pi h^2$ (Cressie, 1991). Intuitively, this is because the numerator of Equation 2.1 is simply the area of the circular region we are considering,

multiplied by the intensity:

$$K(h) = \frac{E(C(x,h)|\text{point at } x)}{\lambda} = \frac{\lambda\pi h^2}{\lambda} = \pi h^2.$$

For clustered processes, Ripley's K is less than πh^2, and for regular processes it is greater than πh^2.

2.2.3 Inhomogenous Ripley's K

Now we consider the case where we no longer assume the point process to be homogenous, but we instead assume that the intensity at a point, x, if determined by the function $\lambda(x)$. The isotropic estimate is extended to this case in Baddeley et al. (2000) in the following way:

$$\hat{K}_{inhom}(h) = \frac{1}{|A|}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{I(||x_i - x_j|| \leq h)w(x_i, x_j)}{\lambda(x_i)\lambda(x_j)}.$$

2.3 Inhomogenous Intensity Estimates

Another property which will be estimated throughout this project in the intensity of an inhomogenous spatial point process. This can be done both parametrically and non-parametrically, both of which will be outlined here.

2.3.1 Parametric Estimation

The problem of parametric inhomogeneous intensity estimation is called 'estimation with incomplete covariate data' in Waagepetersen (2008). When using this method, we will consider an inhomogeneous Poisson point process, whose intensity function is dependent on some parameters β, and of course the spatial location/covariates.

Log-linear Model

A popular model discussed in Waagepetersen (2008) is the log-linear model for a Poisson point process, which is also sometimes called 'a modulated Poisson process'. Let $z(x)$ be a vector of the spatial covariates at some location $x \in A \subset \mathbb{R}^d$. The intensity function following this model is then

$$\lambda(x;\beta) = \exp\left(z(x)^T\beta\right), \tag{2.5}$$

where A is the region of interest. The reason this function is used is that is guarantees non negativity of the intensity at all positions (Rathbun et al., 2007).

Maximum Likelihood Estimator

In order to estimate the parameters β, we will use a variant of the maximum likleihood estimator. The loglikelihood of an inhomogenous Poisson point process with a log-linear

intensity, X, observed within the bounded observation window A, is

$$\mathcal{L}_A(\beta) = \beta^T \sum_{x \in X} z(x) - \int_A \exp\left(\beta^T z(x)\right) dx$$

as seen in Rathbun et al. (2007).

The issue which arises here is that in order to obtain the optimal β through maximising this likelihood, we require the values of $z(\cdot)$ to be known within the entire region, A, however, these are usually only observed at a finite subset of locations which are the events of the point process X. One way to overcome this is to approximate the integral above term above. Firstly, note that the MLE we are aiming to obtain, $\hat{\beta}$ satisfies $\psi_A(\hat{\beta}) = 0$, where

$$\psi_A(\beta) = \frac{1}{|A|}\left(\sum_{x \in X} z(x) - \int_A z(x) \exp\left(\beta^T z(x)\right) dx\right) \tag{2.6}$$

where the extra $1/|A|$ multiplier is added here for consistency with Rathbun et al. (2007), in which it is a technical requirement for the derivation of large-sample properties of $\hat{\beta}$. This quantity is also sometimes called the score function, such as in Waagepetersen (2008).

We now want to approximate the integral

$$Y(A;\beta) = \int_A z(x) \exp\left(\beta^T z(x)\right) dx.$$

Since we do not have the values of the covariates for all $x \in A$, we use a 'design-unbiased' estimator for $Y(A;\beta)$ using the observed covariates which are obtained from a probability based sample of sites $u_1, u_2, ..., u_m \in A$. These are called 'covariate sample sites', and are sampled from a known joint probability distribution. Note that these are different to the event of the point process. We then define the inclusion density,

$$\pi(x) = \sum_{i=1}^{m} f_i(x),$$

where $f_i(\cdot)$ is the known marginal probability density function for $u_1, u_2, ..., u_m$. Assuming that $\pi(x) > 0$ for all $x \in A$, then an unbiased estimator for $Y(A;\beta)$ using the Horvitz-Thomson estimator is

$$\hat{Y}(A;\beta) = \sum_{i=1}^{m} \frac{z(u_i) \exp\left(\beta^T z(u_i)\right)}{\pi(u_i)},$$

(Rathbun et al., 2007). Using this estimator, Equation 2.6 becomes

$$\hat{\psi}_A(\beta) = \frac{1}{|A|}\left(\sum_{x \in X} z(x_i) - \sum_{i=1}^{m} \frac{z(u_i) \exp\left(\beta^T z(u_i)\right)}{\pi(u_i)}\right), \tag{2.7}$$

and we can solve for $\tilde{\beta}$ which satisfies $\hat{\psi}_A(\tilde{\beta}) = 0$.

Asymptotic Properties of $\hat{\beta}$

Rathbun and Cressie (1994) show that the estimator $\hat{\beta}$ satisfying $\psi_A(\hat{\beta}) = 0$ as in 2.6, has various asymptotic properties.

It is shown that $\hat{\beta}$ is consistent, and asymptotically Gaussian with a covariance matrix equal to the inverse Fisher information, $var\left(\hat{\beta}\right) = I(\beta)^{-1}$ where

$$I(\beta) = \int_A z(x)z^T(x)\exp\left(\beta^T z(x)\right)dx.$$

As well as this, properties of $\tilde{\beta}$ are derived in Rathbun et al. (2007), including property that it is also asymptotically Gaussian.

2.3.2 Kernel Density Estimation

An alternative to the parametric intensity estimate discussed in the previous subsection is the non-parametric kernel estimator. The edge corrected kernel estimator is presented in Cressie (1991) as:

$$\hat{\lambda}_h(x) = \frac{1}{p_h(x)}\sum_{i=1}^{n} h^{-d}\kappa\left(\frac{x - x_i}{h}\right),$$

where we sum over all events within an observation window $A \subset \mathbb{R}^d$. The parameter $h > 0$ determines the amount of smoothing. The kernel function, κ, is any probability density function which is symmetric about the origin. The most commonly used kernel, and the one which we will use for experiments here is the Gaussian kernel:

$$\kappa(u) = (2\pi)^{-d/2}\exp\left(-\frac{u^T u}{2}\right).$$

Also, $p_h(x)$ denotes the edge correction introduced in Diggle (1985):

$$p_h(x) = \int_A \kappa_h(x - y)dy.$$

Chapter 3

Existing Bootstrapping Methods

In this chapter we will review the current methods found in the literature for forming confidence intervals, particularly in the spatial setting. We will discuss various benefits and flaws of these, and finally, introduce the alternative method of thinning, which is the focus of this project.

In summary, bootstrapping is a method used to overcome the issue faced when estimating a statistic, θ, with limited observations (Efron and Tibshirani, 1993). In order for the estimate to be useful we require a standard error, or some knowledge of its distribution, to be able to understand its accuracy. This can then be used to form a confidence interval. Unless we know the analytic form of the distribution, which is typically not the case, it is difficult to form a confidence interval and gauge the accuracy. Over the past several decades, many methods have been developed to remedy this, using only the given limited number of observations. We will discuss some of the relevant methods here.

3.1 Bootstrapping Methods for Distributions

3.1.1 Efron's Bootstrap

Efron's bootstrap was first introduced in Efron (1979) and later summarised in Efron and Tibshirani (1993) and is still one of, if not the most popular method used today. It is typically described in the literature simply as the bootstrap.

We formalise the idea of bootstrapping by considering the problem of estimating the distribution of some statistic of interest, $\theta = \theta(F)$, where F is an unknown underlying probability distribution, given an observed random sample $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ which is independently and identically distributed. In this Section we use $\mathbf{x}$ to denote a set of points. Let F_n be the empirical distribution estimating F which places a probability of $1/n$ on each of the observed values $x_1, x_2, ..., x_n$, and a probability of zero otherwise. The estimate of the statistic is denoted by $\hat{\theta} = \hat{\theta}(F_n) = \hat{\theta}(\mathbf{x})$.

We can then define a bootstrap sample, which is a sample of size n drawn from F_n, denoted by $\mathbf{x}^* = \{x_1^*, x_2^*, ..., x_n^*\}$. These n points are drawn using sampling with replacement, such that points in $\mathbf{x}$ may appear more than once, or not at all in $\mathbf{x}^*$. The bootstrap replication

of $\hat{\theta}$ is $\hat{\theta}^* = \hat{\theta}^*(\mathbf{x}^*)$. We repeat this bootstrapping process a large number of times, B, to obtain multiple bootstrap replications $\hat{\theta}^*_1, \hat{\theta}^*_2, ..., \hat{\theta}^*_B$. The standard error of $\hat{\theta}$ is then estimated using the sample variance of these replications.

This method has proven hugely successful, partly due to its simplicity and ease of application. Bootstrapping can be applied to very straightforward and very complicated statistics in the same way, and the computing power required is relatively low.

A significant issue arises however when bootstrapping higher dimensional spatial data. The bootstrap requires that the observed data is independently and identically distributed, which is typically a reasonable assumption when sampling from probability distributions, however this is not the case for spatial point processes. In the spatial setting, there are usually interactions between points, and it is often these interactions which we are looking to understand better through various statistics and their confidence intervals.

The bootstrap has been adapted in various ways to account for this problem and preserve the interactions between points. We will explore a few of these methods in the next Section.

3.1.2 Quenouille-Tukey Jackknife

The Quenouille-Tukey jackknife (Quenouille, 1949, 1956; Tukey, 1958) was a predecessor to the bootstrap, which was found to in fact be an approximation to the bootstrap (Efron and Tibshirani, 1993).

Again, this method is used to estimate the standard error or distribution of $\hat{\theta}$ using an observed random i.i.d sample $\mathbf{x} = \{x_1, x_2, ..., x_n\}$. The jackknife samples are a set of samples, each of which include $n-1$ of the observed points, with one point left out:

$$\mathbf{x}_{(i)} = \{x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n\}$$

for $i = 1, 2, ..., n$. Each of these can be used to form a jackknife replication, $\hat{\theta}_{(i)} = \hat{\theta}_{(i)}\left(\mathbf{x}_{(i)}\right)$. Analogously to the bootstrap, the standard error is obtained using the sample variance of these replications, however we now require an additional multiplicative 'inflation factor':

$$\widehat{SE} = \left(\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2\right)^{1/2}$$

where

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}.$$

The reason for this adaptation of the sample variance is that the jackknife deviations $\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}$ are generally smaller than the bootstrap deviations $\hat{\theta}^*_i - \hat{\theta}^*_{(\cdot)}$ where

$$\hat{\theta}^*_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{B}\hat{\theta}^*_i.$$

We will not discuss the derivation of this inflation factor here, but this idea will become very important in the later chapters when obtaining confidence intervals using thinning.

The main issue with the jackknife which paved the way for the bootstrap is that it becomes inefficient for nonlinear statistics. A linear statistic is a statistic which can be written in the form

$$\hat{\theta}(\mathbf{x}) = \mu + \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) \tag{3.1}$$

where μ is constant and $\phi(\cdot)$ is a function of one point only. No further conditions are required on the function. Of course, many desirable statistics do not satisfy this condition, so the bootstrap is used instead, since it has been shown that the jackknife is a linear approximation to the bootstrap.

An extension to this method introduced by Wu (1986), is the delete-d jackknife. In this case, each jackknife sample has d points left out, rather than just one. This method has been shown to work better than the jackknife for non-smooth statistics, however, does not solve the linearity constraint.

Algorithm 1: Confidence Interval Cover

set R, n, B, α, θ
initialise count $= 0$
for $i = 1, 2, ..., R$ **do**
  generate 'observed' sample $x_1, x_2, ..., x_n$
  for *j=1,2,...,B* **do**
    generate subsample $\mathbf{x}'_{(j)}$
    calculate replication $\hat{\theta}'_{(j)}$
  end
  calculate and store $(1-\alpha)\%$ confidence interval (L, U)
  if $L \leq \theta \leq U$ **then**
    count += 1
  else
    do nothing
  end
end
cover $= \frac{\text{count}}{R}$
return *cover*

Algorithm 2: Thinning

Data: Observed data $x_1, x_2, ..., x_n$
Result: Approximation of $\hat{\sigma}^2$
set $B, p, \gamma(\cdot)$
for $i = 1, 2, ..., B$ **do**
  initialize thinned sample $\mathbf{x}_s^{(i)}$
  for x_i *in* $x_1, x_2, ..., x_n$ **do**
    generate $u_i \sim Unif(0, 1)$
    if $u_i < p$ **then**
      add x_i to $\mathbf{x}_s^{(i)}$
    else
      discard x_i
    end
  end
  calculate and store $\hat{\theta}_s^{(i)}\left(\mathbf{x}_s^{(i)}\right)$
end
set $\bar{\theta}_s = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_s^{(i)}\left(\mathbf{x}_s^{(i)}\right)$
set $\hat{\sigma}_s^2 = \frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_s^{(i)}\left(\mathbf{x}_s^{(i)}\right) - \bar{\theta}_s\right)^2$
return $\hat{\sigma}^2 \approx \gamma(p)\hat{\sigma}_s^2$

3.2 Confidence Interval Cover

Before exploring various methods of confidence interval formation in the spatial setting, we will begin by explaining the idea of a confidence interval cover, which is used to determine the accuracy of these methods. In order to calculate a cover, we must know the true value of the statistic of interest, θ, and be able to simulate a large amount of 'observed' data.

We begin by simulating a set of observed data of size n, $\mathbf{x} = \{x_1, x_2, ..., x_n\}$. We then obtain B subsamples, $\mathbf{x}'_{(1)}, \mathbf{x}'_{(2)}, ..., \mathbf{x}'_{(B)}$, and their corresponding replications $\hat{\theta}'_{(1)}, \hat{\theta}'_{(2)}, ..., \hat{\theta}'_{(B)}$. We use a dash here to denote an arbitrary resampling method. These replications are then used to form a $(1-\alpha)\%$ confidence interval for the value of θ, where α is the confidence level. This entire process is repeated R times, such that we obtain R confidence intervals. The cover is then simply the proportion of confidence intervals for which the true value falls within the bounds. Clearly, we would like this value to be close to $1-\alpha$. Algorithm 1 outlines this procedure.

3.3 Bootstrapping Spatial Point Processes

In the spatial setting, these classical methods must be adapted. One reason for example why Efron's bootstrap should not be applied to spatial point processes is that the sampling with replacement will result in a process with points which appear twice, meaning it is no longer simple (Mattfeldt et al., 2013). The most prominent reason however is that the assumption that points are generated independently typically no longer holds, except in the case of stationary Poisson (Cressie, 1991).

The most prominent of these adaptations is the block bootstrap, which instead focuses on the resampling of groups of points, rather than individuals, to preserve the dependence. (Künsch, 1989; Mattfeldt et al., 2013; Hall, 1985b; Lahiri, 1993; Politis and Romano, 1991; Liu, 1992) have carried out experiments using variants of the block bootstrap which attempt to preserve dependence structure. These experiments mainly consider first-order statistics.

Loh and Stein (2004) explore the effectiveness of four bootstrapping methods to find confidence intervals for estimates of Ripley's K statistic for various spatial point processes. This is a more difficult task than resampling a point process to estimate a first-order statistic, since it is important to avoid producing pairs of points which have a different clustering structure to the original process. In particular, they examine the performance of the splitting, tiling, subsets, and marked point methods, which we will discuss in this chapter.

3.3.1 Methods

Splitting Method

The splitting method is the most simple of the four. It entails splitting the region over which the point process has been observed into N congruent subregions. We assume that the statistics calculated on each subregion are independent with a Gaussian distribution, which

becomes a more sensible assumption at N increases (as a result of the central limit theorem, (Billingsley, 1986)). We also assume that the variance of the statistic in a subregion is N times greater than the statistic observed over the entire observed region. To obtain a confidence interval the variance of the statistic is estimated by the sample variance of the set of N statistics estimated from each of the subregions, more precisely, for an arbitrarily statistic θ, we obtain an $(1-\alpha)\%$ confidence interval as:

$$\left(\hat{\theta} - t_{N-1,\alpha/2}\sqrt{\frac{\widehat{Var}(\hat{\theta}_i)}{N}}, \hat{\theta} + t_{N-1,\alpha/2}\sqrt{\frac{\widehat{Var}(\hat{\theta}_i)}{N}}\right), \qquad \widehat{Var}(\hat{\theta}) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_i - \hat{\theta}\right)^2$$

where $\hat{\theta}_i$ is the statistic estimated from the ith subregion, $\hat{\theta}$ is the statistic estimated from the entire observed region, and $t_{N-1,\alpha/2}$ is the $(1-\alpha/2)$th percentile of the t-distribution with $N-1$ degrees of freedom. This is the studentized confidence interval.

Tiling Method and Subsets Method

The tiling method is also sometimes known as the block bootstrap and was introduced by Hall (1985b), Künsch (1993) and Liu (1992) independently. This method is carried out by randomly sampling N subregions of area a/N where a is the area of the entire observed window. The patterns within each of these subregions are then rearranged in a predetermined way into a new point process. This new process has $\hat{n}$ points and an estimate of the statistic is found from it. This process is repeated B times, and using the basic bootstrap method, a $(1-\alpha)\%$ confidence interval is given by

$$[2\hat{\theta} - \tilde{\theta}_{(B+1)(1-\alpha/2)}, 2\hat{\theta} - \tilde{\theta}_{(B+1)\alpha/2}] \tag{3.2}$$

where $\tilde{\theta}_{(B+1)(1-\alpha/2)}$ and $\tilde{\theta}_{(B+1)\alpha/2}$ are the $(B+1)(1-\alpha/2)$th and $(B+1)\alpha/2$th orderd values of the B bootstrapped estimates. This confidence interval is called the basic bootstrap interval and was introduced by Davison and Hinkley (1997). For this method, we consider a toroidal wrapping as described in Politis and Romano (1991); Loh and Stein (2008), where we treat the observed window as being wrapped around a torus. Figure 3.1 demonstrates how this method is carried out when $N = 4$. The left subfigure is the original point process, in the case a Poisson process with intensity of $\lambda = 50$. The overlaid boxes are the 4 randomly selected subregions. The boxes shown in red are ones where toroidal wrapping is applied since the subregions extend beyond one observed region. The right subfigure shows the selected subregions arranged in a pre-determined way into the new bootstrapped process.

This also shows a key limitation of the tiling method, in that when artificially placing resampled tiles next to each other can introduce new point pairs which effect the second-order characteristics. This is most prominent in hard core processes where points cannot be within a distance of r_0, but this may not be true for the bootstrapped process. It is shown in Lahiri (1993) that this method of bootstrapping does not provide an accurate distibution of the desired statistic when the data exhibits long range dependence, meaning that the

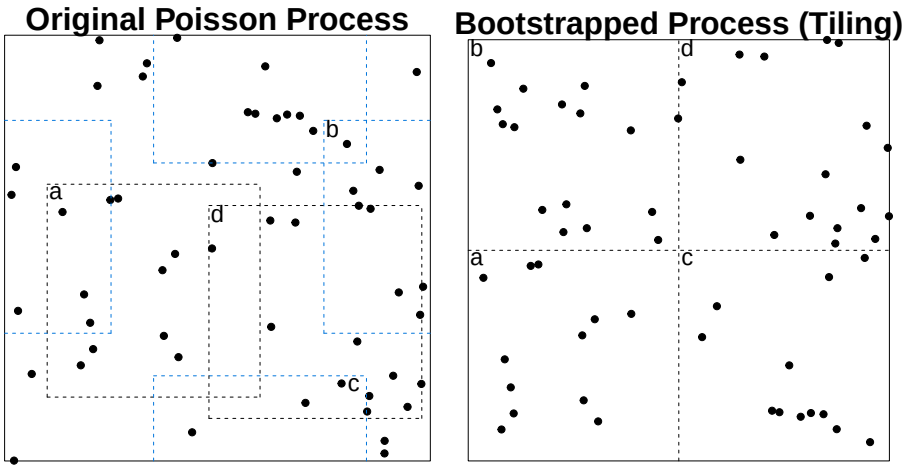

Figure 3.1: Demonstration of bootstrapping via the tiling method with $N = 4$. The left subfigure shows the randomly selected subregions. Those outlined in red demonstrate toroidal wrapping. The right subfigure shows the new bootstrapped process.

covariance between points persists over long distances Cressie (1991).

The subsets method is similar to the tiling method in that it requires selecting subregions of the entire observed region, but in this case they are not rearranged into a new process, but the statistic is caluclated by only considering the effects between points in the same subregion. This method therefore does not have the problem of introducing correlation which did not exist in the original process. Note however that use of toroidal wrapping results in some artificially produced pairs, but much fewer than the tiling method. This method is similar to subsampling introduced in Politis and Romano (1994).

Marked Point Method

The marked point method was also developed in Branu and Kulperger (1998) to estimate the second-order intensity of a point process in $\mathbb{R}$ and is adapted in Loh and Stein (2004) to spatial point processes, in particular when the desired statistic is Ripley's K with an isotropic edge correction. This is done by assigning each observed point, x, a mark, m_x which is equal to the sum of the weights of all points within a distance r of it. Note that this is the weight applied during the isotropic edge correction, as discussed in Section 2.2.1. More precisely,

$$m_x = \sum_{y:y\neq x} 1\{|y-x| \leq r\} w_A(x,y)$$

where $w_A(x,y)$ is the weight arising from the isotropic edge correction. We now again resample the points, along with their marks, by selecting randomly placed subregions. Note that no new marks are calculated. Then, the bootstrapped Ripley's K is calculated by finding the sum of these marks, and normalising them:

$$\tilde{K}(r) = \frac{a}{\sum n_i \left(\sum n_i - 1\right)} \sum_{i=1}^{N} \sum_{j=1}^{n_i} m_{ij}$$

where a is the area of the entire observed region, and m_{ij} is the mark corresponding to the jth point of the ith sampled subregion.

Again, this method does not artificially introduce point pairs, and it is useful since the mark assigned to each point encodes information of the point pairs beyond the subregion window.

Resampling Method

The resampling method is carried out by first splitting the observation window into N subregions, by partitioning it such that the regions do not overlap. These blocks are then resampled with replacement N times, and the samples are rearranged in a predetermined way to form a new observation.

3.3.2 Performance

Mattfeldt et al. (2013) and Loh and Stein (2004) present results for the covers of confidence intervals computed for the estimation of intensity and Ripley's K for various processes. Both of these consider completely spatially random, clustered and regular processes. Mattfeldt et al. (2013) uses tiling and resampling methods, and Loh and Stein (2004) uses splitting, tiling, subsets and marked point methods.

Something that is observed consistently throughout is that the covers, which we could ideally like to be close to $1-\alpha$, vary with the subregion size for the same underlying process and bootstrapping method. In fact, Mattfeldt et al. (2013) shows a consistent positive correlation between number of subregions and cover. This effect is less prominent when the statistic being estimated is Ripley's K, especially with an underlying clustered process. We see also that for the marked point method, the differences in coverage arising from changing the size of the subregions are small relative to other methods.

The reason for this dependence is that for most methods, if the subregion is too small, such that the number of subregions is high, it does not adequately capture the clustering structure. However, if we have a small number of subregions which are very large, then there is very little variability between the samples. This dependency is a problem as it means that different statistics and underlying distributions will require different subregion sizes, however, calibrating this parameter would require knowledge of the true process, which of course makes the estimation and confidence interval formation redundant. The marked point method seems to vary the least with respect to the subregion size, and therefore this dependence is less of an issue.

The marked point process, which performs relatively well in Loh and Stein (2004), is discussed further in Loh (2010) with a focus on estimating the inhomogeneous intensity instead. It is shown that this produces a consistent variance estimator, however this only holds under some strong conditions. It is shown that with the use of an 'adjustment factor', the coverage is consistently close to $1-\alpha$, and the method performs well in this case. This factor is only dependent on the size of the observed region, and the number of points, so does

not require additional information. The main downside of this method is the computational cost.

Additionally, Mattfeldt et al. (2013) shows by carrying out the same experiment on Matérn processes with different amount of clustering, that the subregion size is also dependent on the amount of clustering. Clustering is a very important property which is often unknown and the aim of the analysis, which clearly presents problems.

The covers for the intensity estimates are relatively close to the desired 0.95, will all the methods producing covers between 0.81 and 0.96. The results for Ripley's K however are much more varied and unreliable, sometimes reaching extremes of 0 or 1. Confidence interval formation for second-order statistics is a difficult problem, since resampling may produce new pairs of points with new properties, or conversely discard important properties.

3.4 Proposed Method: Thinning

3.4.1 Method

The aim of this project is to explore an alternative method of confidence interval formation in the spatial setting, which we call thinning. This method involves forming thinned subsamples from the observed data by choosing to independently retain or discard each point with a predetermined probability. This process is of course repeated and the estimates formed using each of these subsamples are aggregated to approximate the distribution of the desired statistic.

We can approach this process more formally in terms of distributions using similar reasoning to Section 3.1.1. Let F be an underlying probability distribution, and $\theta = \theta(F)$ be the statistic of interest. We have an observed sample of n points, $\mathbf{x} = \{x_1, x_2, ..., x_n\}$. Let F_n denote the empirical distribution corresponding to F which places a probability of $1/n$ at each of the observed points $x_1, x_2, ..., x_n$, and zero otherwise. The estimate of the statistic using all of the observed data is denoted by $\hat{\theta} = \hat{\theta}(F_n) = \hat{\theta}(\mathbf{x})$.

We now define the thinned sample, $\mathbf{x}_s$. Let p denote the thinning probability, such that $\mathbf{x}_s$ is a subsample of $\mathbf{x}$ where each point is retained with probability p, and otherwise discarded. Therefore, the expected size of $\mathbf{x}_s$ is $r = np$, and the expected number of points discarded is $d = n - np = n(1-p)$. The thinning replication is then $\hat{\theta}_s = \hat{\theta}_s(\mathbf{x}_s)$. This process is carried out B times, for a large B, and the estimates are used together to approximate the distribution of $\hat{\theta}$.

We will see in the later chapters that the quantity of interest required for forming confidence intervals is the sample variance of $\hat{\theta} - \theta$, which we denote by $\hat{\sigma}^2$. We will approximate this using the sample variance of $\hat{\theta}_s - \hat{\theta}$, denoted by $\hat{\sigma}_s^2$. It will be shown in the next chapter that we require a multiplicative constant, similar to the 'inflation factor' for the jackknife to carry out this approximation. This constant will be a function of the thinning parameter, p. For now, denote this as $\gamma(p)$, such that $\hat{\sigma}^2 \approx \gamma(p)\hat{\sigma}_s^2$. Note that given one observed sample, $\hat{\theta}$ can be treated as a constant, s.t $var\left(\hat{\theta}_s - \hat{\theta}\right) = var\left(\hat{\theta}_s\right)$, so we will use the sample

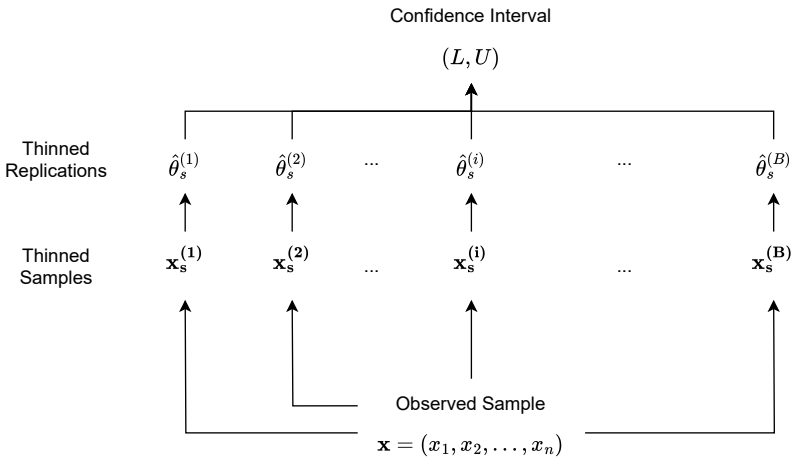

Figure 3.2: Diagram showing process of using thinning method to obtain a confidence interval.

variance of the thinning replications to obtain $\hat{\sigma}_s^2$. Figure 3.2 demonstrates this method.

Algorithm 2 shows the process of approximating $\hat{\sigma}^2$, which can then be used to form a confidence interval using the method described in Chapter 5. Note that this algorithm forms the thinned samples by looping through all observations and using a uniform random variable to determine whether it is retained or discarded. However, in reality this is implemented with vectorisation, which significantly speeds up the process.

3.4.2 Potential Benefits and Issues

The reason for presenting this new method is that spatial statistics is a relatively new field, and the amount of research carried out on confidence interval formation is not extensive, especially for second-order statistics. The methods that have been explored are mostly types of block bootstrap, with the restriction that the cover is dependent on the subregion size. The most successful method where this issue is less prominent is the marked point method, however this introduces a large computational load. This proposed method is a very different approach. It of course does not involve selecting subregions, so the size of these does not matter. A new thinning parameter is introduced which may present similar issues, but we will see in the new chapter that we can remove dependence on this parameter in a lot of cases. Therefore, there is potential for the thinning method to solve this issue in the existing methods.

Also, thinning does not involve moving points around, and therefore does not introduce new pairs of points and new interactions. This could allow it to preserve the dependence structure of the underlying process, unlike the majority of the existing methods. However, the observed window is thinned uniformly, so of course many interactions are discarded, which may present issues. This is also the major flaw which does not allow Efron's bootstrap and the Quenouille-Tukey jackknife to be applied to the spatial setting.

Chapter 4

Analytic Results

4.1 Asymptotics of Thinning

4.1.1 Expansion of Statistical Functional

In this chapter, we aim to obtain the distribution of a statistical functional, $\theta : \mathcal{F} \to \mathbb{R}$, when estimates are obtained from thinned samples and as the number of observations increases $n \to \infty$. $\mathcal{F}$ is the space of distributions, including the true distribution F and the empirical distribution F_n. Also, we require $\mathcal{F}$ to be convex, such that for any $F, G \in \mathcal{F}$, the linear combination $(1-\lambda)F + \lambda G$ is also in $\mathcal{F}$.

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be the observed data, which are independent and identically distributed according to some unknown distribution, F, with finite variance. Let $\theta = \theta(F)$ be the true statistical functional which is a function of the true distribution, and $\hat{\theta} = \hat{\theta}(F_n) = \hat{\theta}(\mathbf{x})$ be an estimator of θ which is a function of the empirical distribution.

Firstly, we aim to understand the distribution of $\hat{\theta} - \theta$ for an arbitrary statistical functional. von Mises (1947) describes the Taylor expansion for a statistical functional:

$$\hat{\theta}(F_n) = \theta(F) + d_1\theta(F; F_n - F) + \frac{1}{2!}d_2\theta(F; F_n - F) + ...$$

There are two possible formulations of this notion of a derivative of a statistical functional, the Gâteaux derivative, and the Fréchet derivative. The first Gâteaux derivative of θ at F in the direction of G is defined as

$$d_1\theta(F; G - F) = \lim_{\lambda \to 0^+} \frac{\theta((1-\lambda)F + \lambda G) - \theta(F)}{\lambda}$$

where $G \in \mathcal{F}$. More generally, the kth derivative is given by

$$d_k\theta(F; F_n - F) = \frac{d^k}{d\lambda^k}\theta((1-\lambda)F + \lambda G)\Big|_{\lambda=0^+}$$

given that the limit exists. We can then set $G = F_n$ in order to formulate the remainder

term as

$$R_n = \hat{\theta}(F_n) - \theta(F) - \sum_{k=1}^{m} \frac{1}{k!} d_k\theta(F; F_n - F)$$

for some $m \in \mathbb{N}$. For the asymptotic result which will be derived in Section 4.1.3, we will take $m = 1$.

In general, we are considering estimators which can be expanded in the following form:

$$\hat{\theta}(F_n) = \theta(F) + \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) + R_n, \tag{4.1}$$

where ϕ is measurable in x with $E(\phi(X)) = 0$, $0 < E(\phi^2(X)) = \sigma^2 < \infty$. The condition that ϕ should be measurable ensures that the resulting $\phi(x_i)$ terms are all also random variables, since random variables are defined to be measurable functions (Fristedt and Gray, 1997), and the composition of measurable functions is also measurable (Strichartz, 2010). We then see that this form of the estimator splits it into a sum of functions of each observed point, and the remainder which captures the error arising from having limited observations. We can therefore describe estimators satisfying Equation 4.1 as having a linear part which is $O(n^{-1/2})$ and a remainder term with a lower order.

Note that the summation in Equation 4.1 matches up to the first derivative in the Taylor expansion:

$$d_1\theta(F; F_n - F) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i).$$

This happens via the von Mises derivative (Fernholz, 1983). θ is von Mises differentiable if its Gâteaux derivative, $d_1\theta(F; G - F)$, and can be expressed as an integral of a measurable function ϕ, over the difference of the distributions:

$$d_1\theta(F; F_n - F) = \frac{d}{d\lambda}\theta((1-\lambda)F + \lambda G)\Big|_{\lambda=0^+} = \int \phi(x) d(G - F)(x).$$

If this holds, we call ϕ the influence function of θ at F. As explained in Fernholz (1983), this function applied to some point x can be written as

$$\phi(x) = \frac{d}{d\lambda}\theta((1-\lambda)F + \lambda\delta_x)\Big|_{\lambda=0^+}$$

where δ_x is the Dirac measure at x, which assigns a measure of 1 to any set containing x and a measure of 0 otherwise. Also, note that the influence function is only unique up to an additive constant, so we can choose it such that

$$\int \phi(x) dF(x) = 0.$$

Therefore, given that θ is von Mises differentiable, the expansion in Equation 4.1 can be

obtained by truncating the Taylor series of the statistical functional in the following way:

$$\begin{aligned}\hat{\theta}(F_n) &= \theta(F) + d_1\theta(F; F_n - F) + \frac{1}{2!}d_2\theta(F; F_n - F) + \ldots \\ &= \theta(F) + d_1\theta(F; F_n - F) + R_n \\ &= \theta(F) + \int \phi_F(x)d(F_n - F)(x) + R_n \\ &= \theta(F) + \int \phi_F(x)dF_n(x) - 0 + R_n \\ &= \theta(F) + \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) + R_n\end{aligned}$$

4.1.2 Convergence of Remainder Term

In the case of $m = 1$, it is desirable for the remainder term to satisfy

$$\sqrt{n}R_n \xrightarrow{p} 0. \tag{4.2}$$

which also shows that the σ^2 seen in Equation 4.1 is simply the limiting variance of $\sqrt{n}\left(\hat{\theta}(F_n) - \theta(F)\right)$. We can now also interpret this expansion of $\hat{\theta}(F_n)$ as allowing us to write it as a linear statistic, with a lower order approximation error. The convergence seen in Equation 4.2 can be guaranteed with the condition that

$$\sqrt{n}\sup_{0\leq\lambda\leq 1}\left|\frac{d^2}{d\lambda^2}\theta((1-\lambda)F + \lambda F_n)\right| \xrightarrow{p} 0,$$

but this is difficult to analyse as it requires a derivative of higher order than necessary. Numerical explorations of this are less complicated and produce useful results. Another way to ensure 4.2 holds is via the Fréchet derivative.

Let $F, G \in \mathcal{F}$, and $\mathcal{D}$ be the linear space generated by differences $F - G$. Let $||\cdot||$ be a norm on $\mathcal{D}$, then the functional $\theta : \mathcal{F} \to \mathbb{R}$ is said to have a Fréchet derivative at $F \in \mathcal{F}$ with respect to norm $||\cdot||$ if there exists a functional $\theta(F; \Delta) : \mathcal{D} \to \mathbb{R}$ and linear in the argument Δ such that

$$\theta(G) - \theta(F) - \theta(F; G - F) = O(||G - F||)$$

as $||G - F|| \to 0$. $\theta(F; \Delta)$ is called the Fréchet derivative (Serfling, 1980). Although Fréchet differentiability is a stronger condition than Gâteaux differentiability, they both provide approximations to the desired

$$\hat{\theta}(F_n) - \theta(F)$$

as seen in Equation 4.1. This term is approximated using the random variable $\theta(F; F_n - F)$ and $d_1\theta(F; F_n - F)$ respectively. The following lemma, shows that these approaches are equivalent, and that the Fréchet derivative is stronger than the Gâteaux derivative.

Lemma 1. *If θ has a Fréchet derivative at F with respect to $||\cdot||$, then for any $G \in \mathcal{F}$,*

$d_1\theta(F; G - F)$ exists and

$$d_1\theta(F; G - F) = \theta(F; G - F)$$

This lemma is proven on pg. 218 of Serfling (1980). The following lemma shows that Fréchet differentiability is sufficient to satisfy the condition in Equation 4.2.

Lemma 2. *Assume θ has a Fréchet derivative at F with respect to $||\cdot||$. Let $(x_1, ..., x_n)$ be observations from F such that $\sqrt{n}||F_n - F|| = O(1)$. Then*

$$\sqrt{n}R_n \xrightarrow{p} 0.$$

Again, this proof can be found on pg. 218 of Serfling (1980). The requirement $\sqrt{n}||F_n - F|| = O(1)$ is established in $\mathbb{R}$ by the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956), which states

$$P\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \epsilon\right) < C\exp(-2n\epsilon^2)$$

where $\epsilon > 0$ and C is some constant.

Many statistics satisfy the condition 4.2, including L statistics, maximum likelihood estimates, least-squares estimates, and U statistics (Serfling, 1980; Wu, 1990).

4.1.3 Derivation of Distribution of Thinned Estimates

Consider $\mathbf{x}_s$ to be a subsample of $\mathbf{x}$ which is obtained by thinning, (sometimes described as Poisson subsampling as in Hájek (1960)). This is done by selecting each data point in $\mathbf{x}$ with probability p. Here we will consider the case where the probability assigned to each x_i is the same, this method is sometimes called Bernoulli subsampling (Wang and Zou, 2021). Thinning is equivalent to simple random sampling without replacement where the size of the subsample, r, is a random variable with a binomial distribution. In order to obtain an estimate of the distribution of θ, the statistic is calculated on B thinned samples where B is large, so this size of each thinned subsample is conditionally distributed as $r_s|N = n \sim Bin(n, p)$ for $s = 1, 2, ..., B$. Also set $d_s = n - r_s$ to be the number of deleted samples. Note that if r_s is sampled to be 0 or n, we either have an empty subsample, or one equivalent to the original observed sample, so we discard this and resample r_s. Both of these cases are unlikely when p is not close to 0 or 1, and the probability of this happening converges to 0 as $n \to \infty$. This result is proven on page 151 of Feller (1968) and will be discussed in more detail later in this proof. We will assume that $\hat{\theta}$ obeys the central limit theorem and is asymptotically normal with limiting variance σ^2,

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N(0, \sigma^2) \tag{4.3}$$

so we need to show that the distribution of the thinned $\hat{\theta}_s$ is also asymptotically normal. Note that thinning can also be reframed as a delete-d Jackknife where d, the number of

x_i deleted for each subsample is instead a binomial random variable. The following proof is based on the asymptotic normality of the distribution of the Jackknife estimates presented in Theorem 2 (iii) of Wu (1990).

Here we define the Thinning histogram analogously to the 'Jackknife histogram' in Wu (1990)

$$H(t) = P\left\{\left(\frac{nr_s}{d_s}\right)^{1/2}\frac{\hat{\theta}_s - \hat{\theta}}{\hat{\sigma}} \leq t\right\}, \tag{4.4}$$

where $\hat{\sigma}^2$ is a consistent estimator of the limiting variance σ^2. Since $\hat{\theta}$ is assumed to be normal, the aim is to prove that $\hat{\theta}_s$ is also $\forall s \in 1, 2, ..., B$.

Theorem 3. *If $d_s/n \geq \lambda$ for some $\lambda > 0$ and $r_s \to \infty$ for all $s = 1, 2, ..., B$, then*

$$\sup_t |H(t) - \Phi(t)| \xrightarrow{p} 0 \tag{4.5}$$

For this proof, we will use the assumed form of the estimator seen in Equation 4.1 to obtain:

$$\hat{\theta}_s = \theta + \frac{1}{r_s}\sum_{i \in x_s} \phi(x_i) + R_{n,s},$$

$$\hat{\theta}_s - \hat{\theta} = \underbrace{\frac{1}{r_s}\sum_{i \in x_s} \phi(x_i) - \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)}_{A} + \underbrace{R_{n,s} - R_n}_{B}.$$

Firstly, we will show in Theorem 4 that the term linear in ϕ, A, converges asymptotically to a normal distribution, and then in Theorem 5 that the remainder term, B, is asymptotically negligible. To show the convergence of the linear part, define $\hat{\theta}_s^L$ and $\hat{\theta}^L$ to be the truncated versions of $\hat{\theta}_s$ and $\hat{\theta}$ respectively by removing the remainders:

$$\hat{\theta}_s^L = \theta + \frac{1}{r_s}\sum_{i \in x_s} \phi(x_i),$$

$$\hat{\theta}^L = \theta + \frac{1}{n}\sum_{i=1}^{n} \phi(x_i).$$

Let the corresponding Thinning histogram be:

$$H_L(t) = P\left\{\left(\frac{nr_s}{d_s}\right)^{1/2}\frac{\hat{\theta}_s^L - \hat{\theta}_s}{\hat{\sigma}} \leq t\right\}.$$

Theorem 4. *If $0 < E[(X - \mu)^2] < \infty$, $r_s \to \infty$ and $d_s \to \infty$ for all $s = 1, 2, ..., B$, then*

$$\sup_t |H_L(t) - \Phi(t)| \xrightarrow{a.s.} 0,$$

where Φ is the standard normal cdf.

Proof. We will begin by showing that the following is satisfied for a thinned set $\mathbf{x}_s$ of size r_s:

$$\lim_{n\to\infty} \frac{1}{(n-1)\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \hat{\theta})^2 \, 1\left(|x_i - \hat{\theta}| \geq \tau \sqrt{\frac{r_s d_s}{n}} \hat{\sigma}\right) = 0, \qquad \tau > 0. \tag{4.6}$$

Note that $\hat{\sigma}^2 r_s d_s / n$ is the variance of $\hat{\theta}_s$ as seen in Hájek (1960). Since we have assumed $\hat{\theta} \to \theta$ and $\hat{\sigma} \to \sigma < \infty$ almost surely, Equation 4.6 becomes:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (x_i - \theta)^2 \, 1\left(|x_i - \theta| \geq \tau \sqrt{\frac{r_s d_s}{n}}\right), \tag{4.7}$$

where the coefficient of the sum has been simplified but remains $O(1/n)$ and the σ inside the indicator is taken into the constant τ.

Now we bound $r_s d_s / n$ above and below. Without loss of generality, let $\min(r_s, d_s) = d_s$, such that $\hat{p} = r_s/n \geq \frac{1}{2}$, then

$$\frac{r_s d_s}{n} = n\hat{p}(1-\hat{p}) = d_s \hat{p} > \frac{1}{2} \min(r_s, d_s)$$

$$\frac{r_s d_s}{n} = \frac{(n - d_s) d_s}{n} = d_s - \frac{d_s^2}{n} < d = \min(r_s, d_s)$$

$$\implies \frac{1}{2} \min(r_s, d_s) < \frac{r_s d_s}{n} < \min(r_s, d_s)$$

which shows $\min(r_s, d_s) \to \infty \iff r_s d_s / n \to \infty$ as $n \to \infty$. This means that given any $k > 0$, we can choose an m large enough such that for $n \geq m$, $r_s d_s / n > k^2$, then 4.7 can be bounded above:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta})^2 \, 1\left(|x_i - \theta| \geq \tau k\right),$$

which is arbitrarily small since k is arbitrarily large. So we have shown that Equation 4.6 holds, then clearly

$$\lim_{n\to\infty} \frac{1}{(n-1)\hat{\sigma}^2} \sum_{i \in x_s} (x_i - \hat{\theta})^2 \, 1\left(|x_i - \hat{\theta}| \geq \tau \sqrt{\frac{r_s d_s}{n}} \hat{\sigma}\right) = 0 \qquad \tau > 0$$

since we have only removed d_s terms from the sum here. By Theorem 3.1 in Hájek (1960), this implies that $\hat{\theta}_s^L$ has an asymptotically normal distribution with mean θ and variance $\hat{\sigma}^2 r_s d_s / n$. Then Pólya (1920) implies the initial claim:

$$\sup_t |H_L(t) - \Phi(t)| \xrightarrow{p} 0.$$

Therefore the claim holds. □

We now show that the remainder term, B, is asymptotically negligible.

Theorem 5. *If* $\sqrt{n}R_n \xrightarrow{p} 0$, $\frac{d_s}{n} > \lambda$ *for some* $\lambda > 0$ *for all* $s = 1, 2, ..., B$ *and* $r_s \to \infty$, *then*

$$Q(\epsilon) = P\left\{\left(\frac{nr_s}{d_s}\right)^{1/2} |R_{n,s} - R_n| \leq \epsilon\right\} \xrightarrow{p} 0.$$

Proof. $Q(\epsilon)$ is clearly non-negative, and we can use the constant λ to bound it above:

$$\left(\frac{nr_s}{d_s}\right)^{1/2} |R_{n,s} - R_n| \leq \left(\frac{r_s}{\lambda}\right)^{1/2} |R_{n,s} - R_n|.$$

Then using the assumption that $\sqrt{n}R_n \xrightarrow{p} 0$:

$$\lim_{n\to\infty} \left(\frac{r_s}{\lambda}\right)^{1/2} |R_{n,s} - R_n| = \lim_{n\to\infty} |\sqrt{r_s}R_{n,s} - \sqrt{r_s}R_n|,$$

where the first term clearly converges to 0 since $R_{n,s}$ is of size r_s, and the second term is bounded above by $\sqrt{n}R_n$ which also converges to 0. Therefore, the claim is proven. □

Note that the requirement that $\frac{d_s}{n} > \lambda$ for some $\lambda > 0$ may not always be met since d_s is a random variable, however as n becomes large we can see using Feller (1968) that for $p < 0.9$, the probability of d_s less than $0.1n$ for example converges to 0:

$$P(d_s \leq 0.1n) \leq \frac{(n - 0.1n)(1-p)}{(n(1-p) - 0.1n)^2} = \frac{0.9(1-p)}{n(0.9-p)^2} \to 0.$$

This means that asymptotically, given that we have chosen $p < 0.9$, we can assume $\frac{d_s}{n} > 0.1$ and more generally that this term is $O(n)$.

Finally, combining Theorems 2 and 3 proves the desired result stated in Theorem 1.

4.1.4 Usefulness of Asymptotic Result

Having now proven Theorem 3, we have shown that the 'root' converges to a standard normal distribution:

$$\left(\frac{nr_s}{d_s}\right)^{1/2} \frac{\hat{\theta}_s - \hat{\theta}}{\hat{\sigma}} \xrightarrow{p} N(0,1).$$

For this result to be useful, we want to show that the distribution of the standardized thinned estimate $\hat{\theta}_s$, is close to that of the standardized estimate over all the observed data, $\hat{\theta}$. To do this, we require a condition on $\hat{\theta}$ stronger than that imposed earlier. Firstly, analogously to Equation 4.4, let

$$H_0(t) = P\left\{\frac{\sqrt{n}\left(\hat{\theta} - \theta\right)}{\sigma} \leq t\right\},$$

and assume now that

$$\sup_t |H_0(t) - \Phi(t)| \xrightarrow{p} 0. \tag{4.8}$$

The following theorem then provides us with the desired result.

Theorem 6. *If $d_s/n \geq \lambda$ for some $\lambda > 0$ and $r_s \to \infty$ for as $s = 1, 2, ..., B$, then*

$$\sup_t |H(t) - H_0(t)| \xrightarrow{p} 0$$

Proof.

$$\begin{aligned}
\sup_t |H(t) - H_0(t)| &= \sup_t |H(t) - \Phi(t) + \Phi(t) - H_0(t)| \\
&\leq \sup_t (|H(t) - \Phi(t)| + |H_0(t) - \Phi(t)|) \\
&\leq \sup_t |H(t) - \Phi(t)| + \sup_t |H_0(t) - \Phi(t)|.
\end{aligned}$$

By using Theorem 3, and assuming 4.8 holds, we know that

$$\begin{aligned}
\sup_t |H(t) - \Phi(t)| &\xrightarrow{p} 0, \\
\sup_t |H_0(t) - \Phi(t)| &\xrightarrow{p} 0.
\end{aligned}$$

Since all terms are positive, we have that also

$$\sup_t |H(t) - \Phi(t)| + \sup_t |H_0(t) - \Phi(t)| \xrightarrow{p} 0,$$

and so we can use the Squeeze theorem (Sohrab, 2014) to prove the result:

$$\begin{aligned}
0 \leq \sup_t |H(t) - H_0(t)| \leq \sup_t |H(t) - \Phi(t)| + \sup_t |H_0(t) - \Phi(t)| \\
\implies \sup_t |H(t) - H_0(t)| \xrightarrow{p} 0
\end{aligned}$$

□

We have now shown that the standardised distributions of the thinned estimate and the observed estimate asymptotically approach each other. Intuitively, since we know that for large n, the distribution of $\hat{\theta}$ is close to the distribution of θ, and the distribution of $\hat{\theta}_s$ is close to the distribution of $\hat{\theta}$, then $\hat{\theta}_s$ can be used to approximate the distribution of the desired θ.

4.2 Link Between Thinning and Bootstrap Sampling

In this section, we will show that when the thinning parameter takes a value of $p = 0.5$, there is an equivalence between the distribution that arises from the thinned estimates, and the distribution arising from the classical bootstrap.

Firstly, we carry out a numerical exploration. Figure 4.1 shows histograms of the estimators calculated from thinned samples and from bootstrapped samples for varying thinning parameters, p. The plots for a Poisson distribution with $\lambda = 10$, and for an exponential distribution with $\lambda = 10$ are shown here. We see that a thinning parameter of $p = 0.2$ results in a distribution with a higher variance than the bootstrapped distribution because

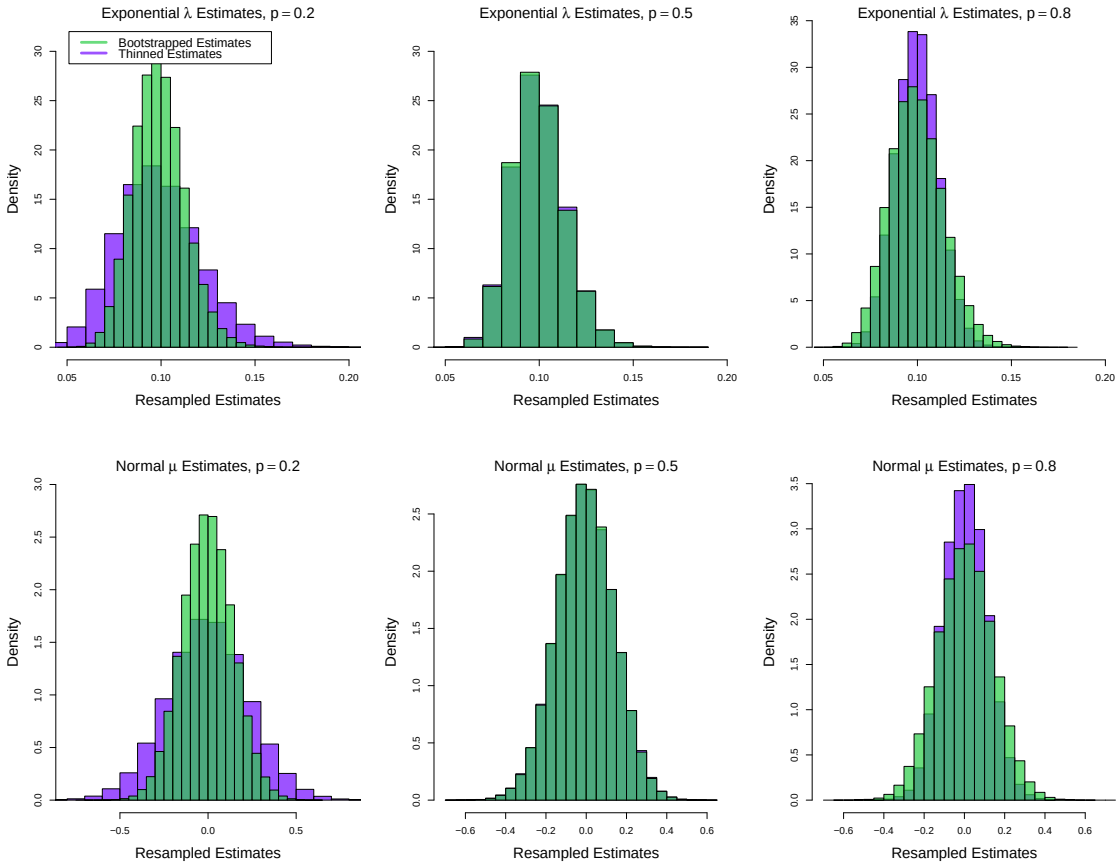

Figure 4.1: Histograms of bootstrapped estimates, $\hat{\theta}^*$, and thinned estimates, $\hat{\theta}_s$, shown in green and purple respectively. The top row uses i.i.d data generated from $X \sim Exp(10)$ and $\lambda = 10$ is estimated using its MLE, which we know satisfies the requirements of Theorems 3 and 6. The bottom row uses data distributed according to $X \sim N(0, 1)$ and again, $\mu = 0$ is estimated using its MLE. These results are produced by carrying out 500 simulations of 100 data points each. For each other these simulations, 1000 thinning samples are obtained and an estimate of the desired statistic is calculated. These estimates are then used to form the histograms.

fewer observations are sampled, and therefore, we do not capture enough information. On the other hand, for $p = 0.8$, the variance of the thinned distribution is smaller than that of the bootstrapped distribution. In this case, we over-fit to the observed data and do not capture enough of the variance in the distribution. For a value of $p = 0.5$, the histograms are very similar and seem to represent the same distribution up to randomness.

We can make this observation more precise by considering the asymptotic distributions which arise from thinning and bootstrapping. Note that we have assumed that the estimator $\hat{\theta}$ obeys the limit in Equation 4.3. Let $\hat{\theta}^*$ be the estimate of the desired statistic based on the bootstrapped sample. Bickel and Freedman (1981) show that, given 4.3, we have that

$$\frac{\sqrt{m}\left(\hat{\theta}^* - \hat{\theta}\right)}{\hat{\sigma}} \to N(0, 1)$$

where m is the number of bootstrap samples selected. Here, we will consider the case $m = n$, such that for large n

$$\frac{\sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right)}{\hat{\sigma}} \sim N(0,1). \tag{4.9}$$

The result from Theorem 3 can be written as

$$\left(\frac{nr_s}{d_s}\right)^{1/2} \frac{\hat{\theta}_s - \hat{\theta}}{\hat{\sigma}} \to N(0,1).$$

Clearly, when $p = 0.5$, $r_s \approx 0.5n$ and $d_s \approx 0.5n$, and therefore for large n

$$\frac{\sqrt{n}\left(\hat{\theta}_s - \hat{\theta}\right)}{\hat{\sigma}} \sim N(0,1),$$

which is approximately the same distribution as the bootstrapped distribution, as in 4.9. Therefore, we conclude that thinning with a parameter of $p = 0.5$ performs very similarly to bootstrapping for statistics which satisfy the conditions of Theorem 3.

4.3 Asymptotics for Spatial Point Processes

4.3.1 Conditions Required

We now turn our attention to the spatial point process setting, and to whether Theorem 3 applies in this case. Various conditions are required on the statistic of interest, θ, and the underlying process. These will be explored in this section. These conditions are summarised here in terms of probability distributions:

(1) The observed data is independent and identically distributed according to some unknown process with finite variance.

(2) θ needs to be a statistical functional which maps some convex space of distributions, $\mathcal{F}$, to $\mathbb{R}$.

(3) θ can be approximated as a linear statistic with an error of order lower than $O(\sqrt{n})$.

(4) The estimate $\hat{\theta}$ obeys the central limit theorem, such that Equation 4.3 holds.

4.3.2 Conditions on Data

An important assumption we place on the sampled data is that they arise independently and are identically distributed. This assumption is much less restrictive when sampling from probability distributions. For spatial point processes, the introduction of any clustering or regularity structure also introduces dependence into the data, and these are often the features we want to understand better via thinning. This issue has been explored thoroughly for the classical bootstrap and jackknife, and as stated in Cressie (1991), 'blind application

of the classical jackknife and bootstrap techniques on dependent data can give incorrect answers'. This is because these resampling techniques do not generally preserve the underlying dependence structure according to Loh (2008), and the same may hold for thinning.

4.3.3 Conditions on Statistic

The statistics of interest that we will focus on here are the intensity in both homogeneous and inhomogeneous cases, $\hat{\lambda}(x)$, and Ripley's K, $\hat{K}(h)$. Both of these are statistical functionals applied to the empirical functions, F_n, as required. Before carrying out numerical experiments for each of these, we will discuss to what extent they satisfy the sufficient conditions above. Note that in all cases, we have a statistical functional.

- **Homogeneous Intensity**: We use the maximum likelihood estimator which we know is asymptotically normal (Sweeting, 1980) and can be approximated by a linear statistic (Serfling, 1980). If the underlying process is a homogeneous Poisson, then we also have i.i.d data, and therefore we expect Theorem 3 to hold.
- **Inhomogenous Intensity**: We will consider both a parametric and non-parametric estimator for the inhomogeneous intensity. The parametric estimator is a maximum likelihood estimator, and therefore satisfies the desired conditions (Serfling, 1980; Sweeting, 1980). It is more difficult to determine whether the same holds for the non-parametric kernel density estimate, so we will explore this numerically.
- **Ripley's K**: It is difficult to determine whether Ripley's K satisfies the differentiability conditions (seen in Section 4.1.2) which would mean it is a linear statistical functional and satisfies requirement 3. Intuitively, this is likely not the case since it is a function of the distances between points, and can also be reframed as a sum of dependent variables. Also, it is most useful when the underlying data exhibits clustering/regularity structure, and therefore is not i.i.d. Svane et al. (2023) shows however that asymptotic normality is satisfied.

4.4 Numerical Demonstration of Results

This section presents simulations which demonstrate the theorems previously proven. Firstly, we consider Theorem 3. Figure 4.2 shows the distribution of the thinned estimates, $\hat{\theta}_s$, after normalisation compared with a standard normal distribution:

$$\left(\frac{nr_s}{d_s}\right)^{1/2} \frac{\hat{\theta}_s - \hat{\theta}}{\hat{\sigma}}. \tag{4.10}$$

Here, we estimate the intensity, λ, of a homogeneous Poisson distribution which is estimated using its MLE. We choose this setting since we know it satisfies the desired conditions summarised in the previous section. As $n \to \infty$, we expect these to converge to the same

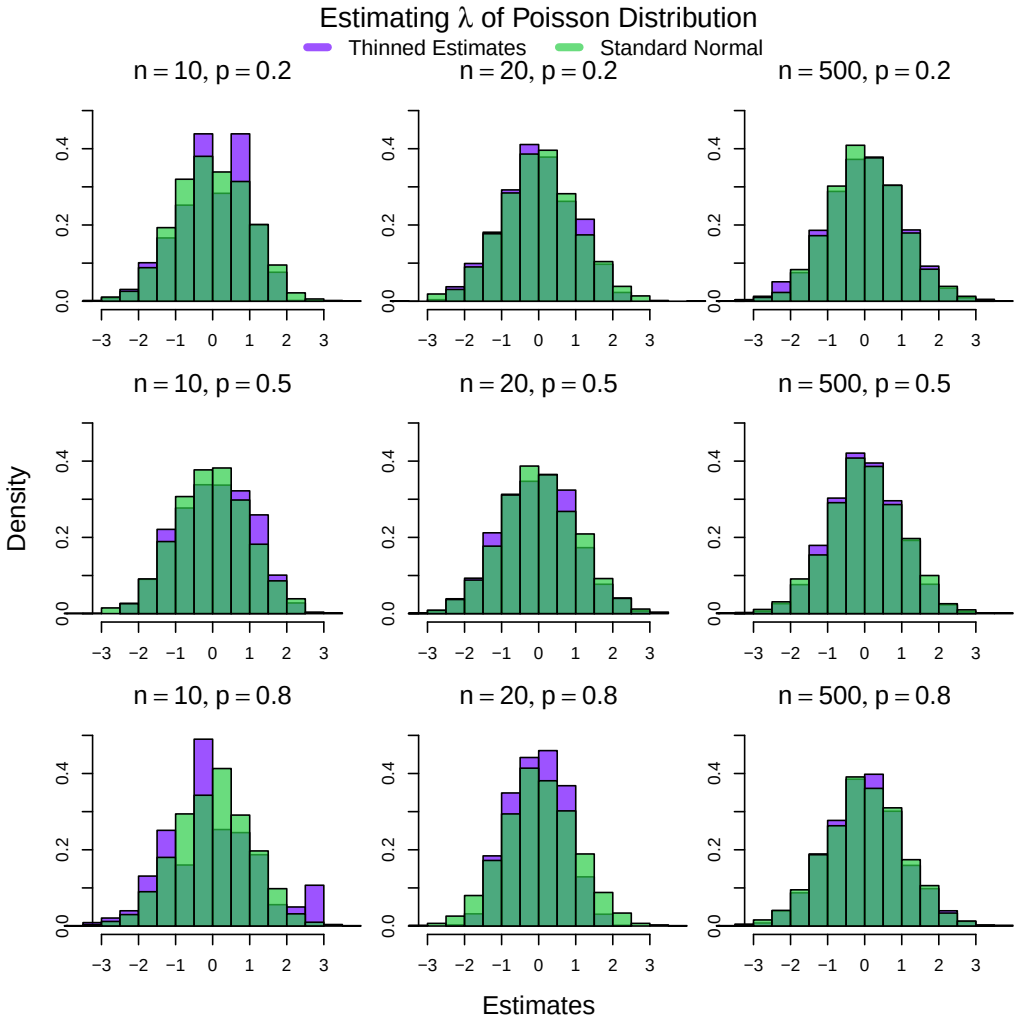

Figure 4.2: Histograms showing distributions of normalised thinned estimates, compared to a standard normal distribution. Each row represents a thinning parameter of $p = 0.2, 0.5, 0.8$ respectively, and each column represents varying observed dataset size, in particular $n = 10, 20, 500$ respectively. Each subplot is generated using n i.i.d datapoints distributed according to a $Poi(10)$ distribution. For each dataset, 2000 thinned samples are obtained and used to calculate the estimate. These are then plotted alongside 2000 samples from a standard normal distribution.

distribution. This aligns with the results in Figure 4.2, since for all thinning parameters considered, the histograms become more similar as n increases. We see also that the convergence occurs quite quickly, such that for $n = 20$, the method produces a reasonable distribution, and the improvements thereafter are less significant. Furthermore, there are no large differences in the distribution when varying the thinning parameter between $p = 0.2, 0.5, 0.8$. This is because this effect has been eliminated in the normalisation through the r_s/d_s term.

Next, we demonstrate the result of Theorem 6 in Figure 4.3. We use the same statistic and estimate as before, except in this case, we compare against the normalised estimates

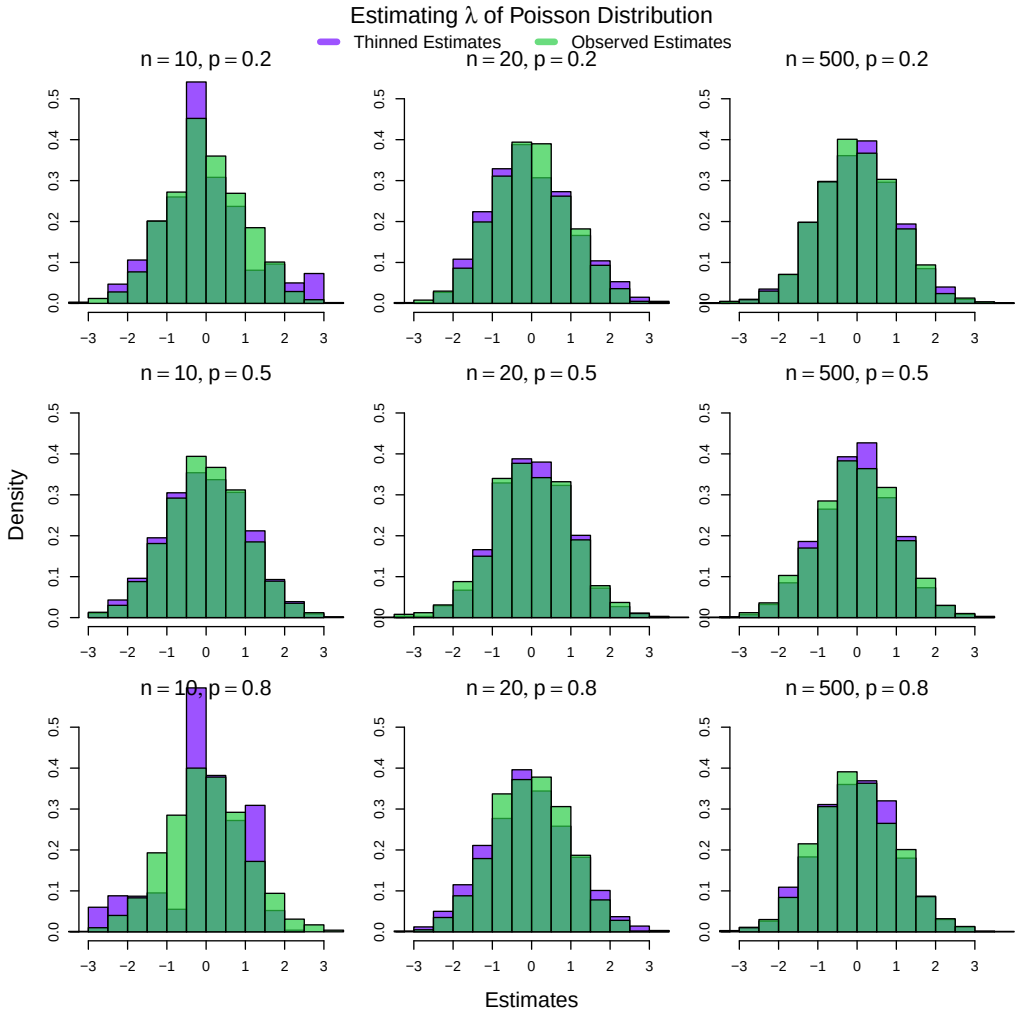

Figure 4.3: Histograms showing distributions of normalised thinned estimates, compared to the normalised estimates calculated over the entire observed dataset. Each row represents a thinning parameter of $p = 0.2, 0.5, 0.8$ respectively, and each column represents varying observed dataset size, in particular $n = 10, 20, 500$ respectively. Each subplot is generated using n i.i.d datapoints distributed according to a $Poi(10)$ distribution. For each dataset, 2000 thinned samples are obtained and used to calculate the estimate. For each case, we also generate 2000 sets of observed data from the distribution, and calculate the estimator over all of these. These estimates are then normalised and plotted against the thinned normalised estimates.

over the entire observed data rather than a standard normal:

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma}. \tag{4.11}$$

Again, we see that as n increases, the the normalised $\hat{\theta}_s$ and $\hat{\theta}$ become closer in distribution, as expected.

4.5 Example: Intensity of Homogeneous Poisson Point Process

In this section, we will explicitly show the result proven in Theorem 3 for the intensity of a homogeneous Poisson spatial point process. Firstly, we define $\hat{\lambda}$. We know that for a homogeneous Poisson process, X, the number of points within a bounded Borel set, A, is

$$N_X(A) \sim Poi(\lambda|A|), \tag{4.12}$$

where λ is the true intensity of the point process, and $|A|$ is the area of A. Let r be a random variable corresponding to the number of points which are retained after thinning. We know this is conditionally distributed as

$$r|N_X(A) = n \sim Bin(n, p). \tag{4.13}$$

By definition, the intensity of a homogeneous Poisson process over a set A with volume $|A|$ is

$$\lambda = \frac{E[N_X(A)]}{|A|}$$

and thus, given a realisation $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, the intensity can be estimated in the following way

$$\hat{\lambda} = \frac{N_X(A)}{|A|},$$

(van Lieshout, 2019). Now, to define $\hat{\lambda}_s$ using the thinned subsample, we again take the ratio of the number of points and the area of the set. This time, since the mean of the number of points retained is np, we need to also normalise by the thinning probability, p:

$$\hat{\lambda}_s = \frac{r}{|A|p},$$

where r is the number of points in the thinned sample $\mathbf{x}_s$. The aim here is to show that the distribution of $\sqrt{n}\left(\hat{\lambda} - \lambda\right)$ is asymptotically the same as that of

$$\sqrt{\frac{nr_s}{d_s}}\left(\hat{\lambda}_s - \hat{\lambda}\right).$$

Since we know $\hat{\lambda}$ has a Poisson distribution with finite variance, we can apply the central limit theorem:

$$\sqrt{n}\left(\hat{\lambda} - \lambda\right) \to N(0, \sigma^2)$$

for large n. Now, before showing the desired result, we need to find the variances of $\hat{\lambda}$ and $\hat{\lambda}_s$.

Theorem 7. *The intensity estimate has a variance of*

$$var(\hat{\lambda}) = \frac{\lambda}{|A|},$$

and the thinned intensity estimate of a homogenous Poisson point process has a variance of

$$var(\hat{\lambda}_s) = \frac{\lambda}{|A|p}.$$

Proof. See Appendix A. □

Note that during this proof, we also obtain the distribution of the number of retained points:

$$r \sim Poi(\lambda|A|p). \tag{4.14}$$

Theorem 8. *The variance of* $\hat{\lambda}_s - \hat{\lambda}$ *is given by*

$$var(\hat{\lambda}_s - \hat{\lambda}) = \frac{(1-p)\lambda}{|A|p}.$$

Proof. See Appendix A □

In summary, we have

$$\left(\frac{n|A|}{\lambda}\right)^{1/2}(\hat{\lambda} - \lambda) \to N(0,1), \quad \text{and} \quad \left(\frac{np|A|}{\lambda(1-p)}\right)^{1/2}(\hat{\lambda}_s - \hat{\lambda}) \to N(0,1),$$

but for large n, r_s/d_s is approximately $p/(1-p)$. So since $\hat{\sigma}^2 = \lambda/|A|$, we have shown that

$$\left(\frac{nr_s}{d_s}\right)^{1/2}\frac{\hat{\lambda}_s - \hat{\lambda}}{\hat{\sigma}} \to N(0,1)$$

as in Theorem 3.

4.6 Numerical Demonstration on Non-Linear Statistic

We have shown that the results in Theorems 3 and 6 hold for a linear statistic, specifically the intensity. We will now explore whether the results still hold in the case of a non-linear statistical functional. We will carry out experiments using the median, ν, which for a Poisson distribution is approximately

$$\nu \approx \lambda + \frac{1}{3} - \frac{1}{50\lambda}$$

according to Choi (1994). We will obtain observed estimates and thinned estimates using plug-in estimates of the intensity:

$$\hat{\nu} = \hat{\lambda} + \frac{1}{3} - \frac{1}{50\hat{\lambda}}, \qquad \hat{\nu}_s = \hat{\lambda}_s + \frac{1}{3} - \frac{1}{50\hat{\lambda}_s}.$$

Analogously to Section 4.4, we plot histograms of 4.10 against a standard normal in Figure 4.4, and 4.10 against 4.11 in Figure 4.5. Interestingly, we see that the distributions converge for large n, despite the statistic no longer being linear. This means that condition 3 in Section 4.3.1 is not a necessary condition, and therefore, the thinning method may be useful in more complex non-linear scenarios such as predicting Ripley's K.

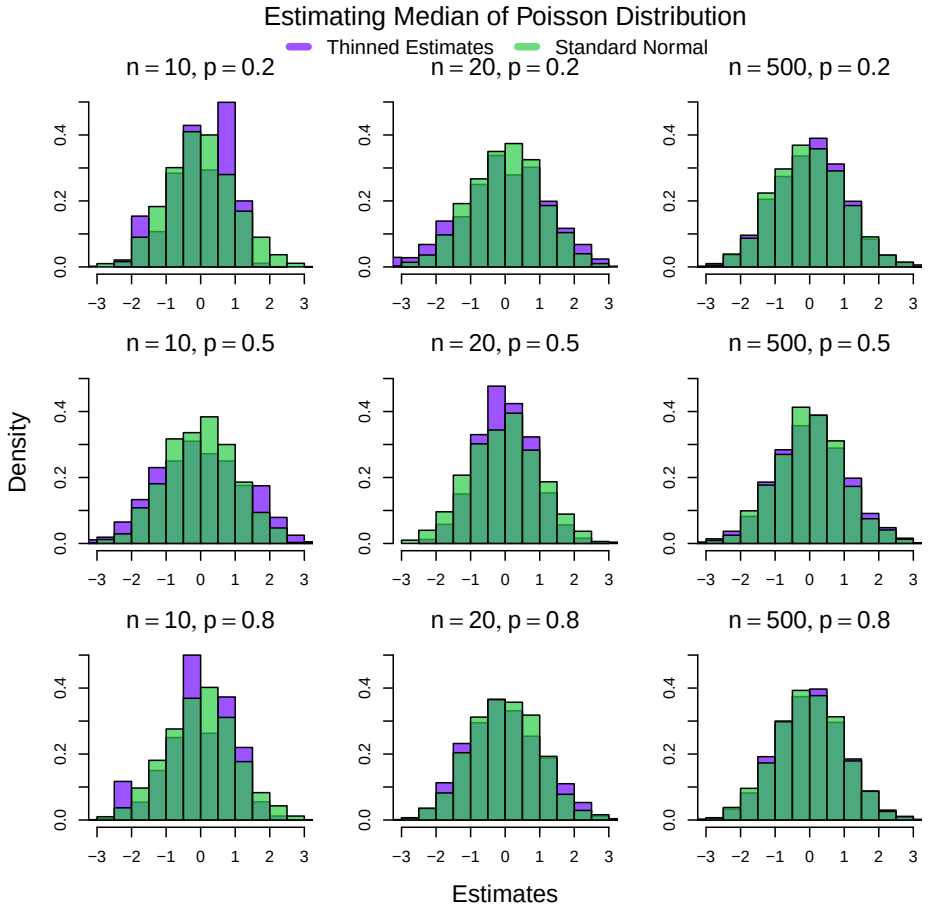

Figure 4.4: Histograms showing distribution of normalised thinned estimates compared to a standard normal distribution for the estimate of the median of a Poisson distribution. Simulation follows the same specification as Figure 4.2.

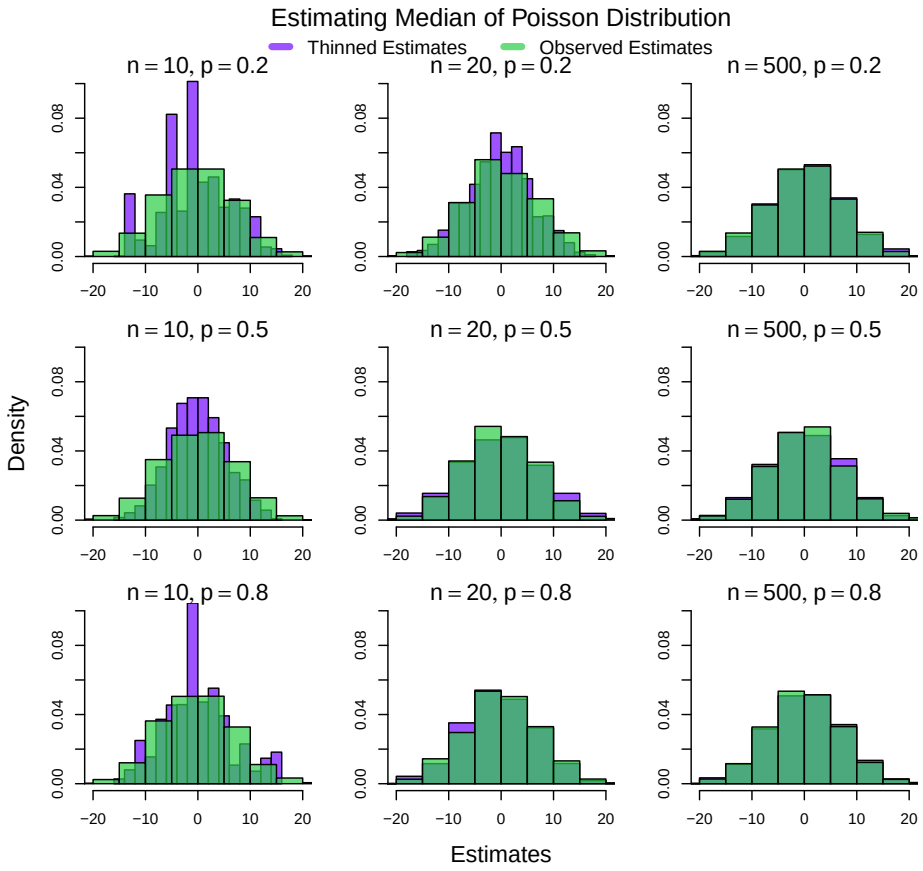

Figure 4.5: Histograms showing distribution of normalised thinned estimates compared to normalised estimates calculated over the entire observed dataset for the estimate of the median of a Poisson distribution. Simulation follows the same specification as Figure 4.3

Chapter 5

Confidence Interval Formation

In this chapter, we will investigate how to best form a confidence interval for a desired statistic, given the results from the previous chapter. We will begin by using the basic boostrap confidence interval, which is most prominently used by Loh and Stein (2004), and adapt this appropriately.

5.1 Naive Confidence Interval

We first attempt to form confidence intervals using the popular method of the basic bootstrap, developed by Davison and Hinkley (1997). It is designed for independently and identically distributed samples, which is something we have also assumed earlier for Theorem 3, so is not restrictive.

Let R be the number of thinned samples we obtain and calculate the estimator from, and α be the confidence level. Let $\hat{\theta}_s^{(R+1)(1-\alpha/2)}$ and $\hat{\theta}_s^{(R+1)\alpha/2}$ be the $(R+1)(1-\alpha/2)$th and $(R+1)\alpha/2$th ordered values of $\hat{\theta}_s$ respectively. Then, using the distribution of $\hat{\theta}_s - \hat{\theta}$ to estimate that of $\hat{\theta} - \theta$, we can then derive a confidence interval in the following way:

$$\begin{aligned}
1-\alpha &= P\left(\hat{\theta}_s^{(R+1)\alpha/2} \leq \hat{\theta}_s \leq \hat{\theta}_s^{(R+1)(1-\alpha/2)}\right) \\
&= P\left(\hat{\theta}_s^{(R+1)\alpha/2} - \hat{\theta} \leq \hat{\theta}_s - \hat{\theta} \leq \hat{\theta}_s^{(R+1)(1-\alpha/2)} - \hat{\theta}\right) \\
&\approx P\left(\hat{\theta}_s^{(R+1)\alpha/2} - \hat{\theta} \leq \hat{\theta} - \theta \leq \hat{\theta}_s^{(R+1)(1-\alpha/2)} - \hat{\theta}\right) \\
&= P\left(\hat{\theta}_s^{(R+1)\alpha/2} - 2\hat{\theta} \leq -\theta \leq \hat{\theta}_s^{(R+1)(1-\alpha/2)} - 2\hat{\theta}\right) \\
&= P\left(2\hat{\theta} - \hat{\theta}_s^{(R+1)(1-\alpha/2)} \leq \theta \leq 2\hat{\theta} - \hat{\theta}_s^{(R+1)\alpha/2}\right).
\end{aligned}$$

The approximate confidence interval is then

$$\left(2\hat{\theta} - \hat{\theta}_s^{(R+1)(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}_s^{(R+1)\alpha/2}\right).$$

This method is implemented in Figure 5.1 for $\alpha = 0.05$ on order to obtain 95% confidence intervals. This is done for the simple case of estimating the intensity, λ of a homogeneous Poisson distribution, as the number of observed points, n, increases. The red dotted line represents a cover of 0.95 which is the desired behaviour, given the choice of α here. We see that only a thinning probability of $p = 0.5$ results in the desired behaviour, while $p = 0.2$

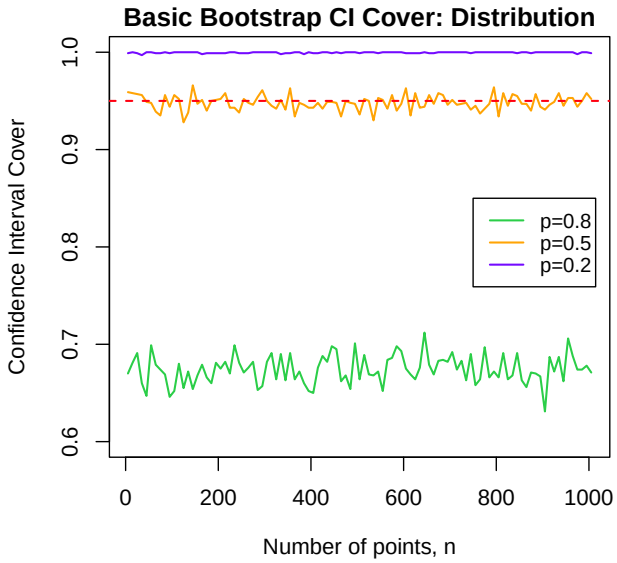

Figure 5.1: Line plot showing the empirical coverage using the basic bootstrap confidence interval as the number of observed datapoints, n, increases from $n = 5$ to $n = 1005$ in steps of 10. The statistic being calculated here is the intensity of a homogenous Poisson distribution. 1000 simulations are carried out, each of which use 500 thinned subsamples. The purple, orange and green line represent thinning parameters of $p = 0.8, 0.5, 0.2$ respectively, and the red dashed line is the desired cover of 0.95.

results in thinned samples with too much variance, and $p = 0.8$ does not capture enough of the variance. The benefit of this confidence interval formation is that the distribution of the true statistic, or its estimator do not need to be known. The assumption made which does not align with thinning is that the distribution of $\hat{\theta}_s - \hat{\theta}$ approximates $\hat{\theta} - \theta$. We have instead that the asymptotic distribution of

$$\left(\frac{nr_s}{d_s}\right)^{1/2} \frac{\hat{\theta}_s - \hat{\theta}}{\hat{\sigma}}$$

approximates $\sqrt{n}\left(\hat{\theta} - \theta\right)$. It is now clear why a thinning probability of $p = 0.5$ produces a good confidence interval cover using the basic bootstrap method. Since $E(r_s) = np$ and $E(d_s) = 1 - np$, when $p = 0.5$, the extra term $\sqrt{r_s/d_s}$ is approximately 1, and the two distributions coincide.

5.2 Studentized Confidence Interval

We now formulate an alternative confidence interval which should adjust for the thinning probability used. To do this, we use a studentized confidence interval. This typically provides a better confidence interval when the distributions of $\sqrt{n}(\hat{\theta} - \theta)$ and $\sqrt{n}(\hat{\theta}_s - \hat{\theta})$ are not close but the distributions of $\sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma}$ and $\sqrt{n}(\hat{\theta}_s - \hat{\theta})/\hat{\sigma}_s$ are. Here, we have the case that the asymptotic variance of $\hat{\theta}_s$ is dependent on the thinning probability used, which is not taken into account with the basic bootstrap confidence intervals.

The interval we are trying to approximate is an interval around $\hat{\theta}$, given the true θ. We

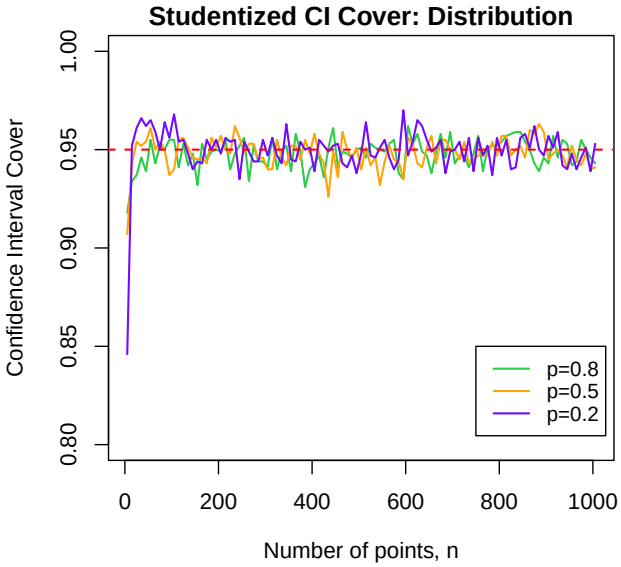

Figure 5.2: Line plot showing the empirical coverage using the Studentized confidence interval as the number of observed datapoints, n, increases from $n = 5$ to $n = 1005$ in steps of 10. The statistic being calculated here is the intensity of a homogeneous Poisson distribution. 1000 simulations are carried out, each of which use 500 thinned subsamples. The purple, orange and green line represent thinning parameters of $p = 0.8, 0.5, 0.2$ respectively, and the red dashed line is the desired cover of 0.95.

are working with statistics with the assumed property that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \sim N(0, \hat{\sigma}^2)$$

which leads to a confidence interval of

$$\left(\hat{\theta} - t_{R-1,1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\theta} + t_{R-1,1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right),$$

where $t_{R-1,1-\alpha/2}$ is the $(1-\alpha/2)$th percentile of a Student's t statistic with $R-1$ degrees of freedom. Note that R is the number of thinned subsamples we obtain in each simulation. Since we only have $\hat{\theta}_s$ and $\hat{\theta}$, we can instead form the interval in the following way

$$\hat{\theta}_s \sim N\left(\hat{\theta}, \hat{\sigma}_s^2\right) \implies \left(\hat{\theta} - t_{R-1,1-\alpha/2}\hat{\sigma}_s, \hat{\theta} + t_{R-1,1-\alpha/2}\hat{\sigma}_s\right).$$

where $\hat{\sigma}_s$ is the variance of the thinned samples. Using the distribution derived in Theorem 3, this is the same as

$$\left(\hat{\theta} - t_{R-1,1-\alpha/2}\hat{\sigma}\sqrt{\frac{d_s}{nr_s}}, \hat{\theta} + t_{R-1,1-\alpha/2}\hat{\sigma}\sqrt{\frac{d_s}{nr_s}}\right)$$

which has an extra $\sqrt{\frac{d_s}{r_s}}$ multiplying the desired variance, $\hat{\sigma}$. We want to use $\hat{\sigma}_s$ to obtain an interval, so we can just cancel out this extra term to obtain

$$\left(\hat{\theta} - t_{R-1,1-\alpha/2}\hat{\sigma}_s\sqrt{\frac{r_s}{d_s}}, \hat{\theta} + t_{R-1,1-\alpha/2}\hat{\sigma}_s\sqrt{\frac{r_s}{d_s}}\right). \tag{5.1}$$

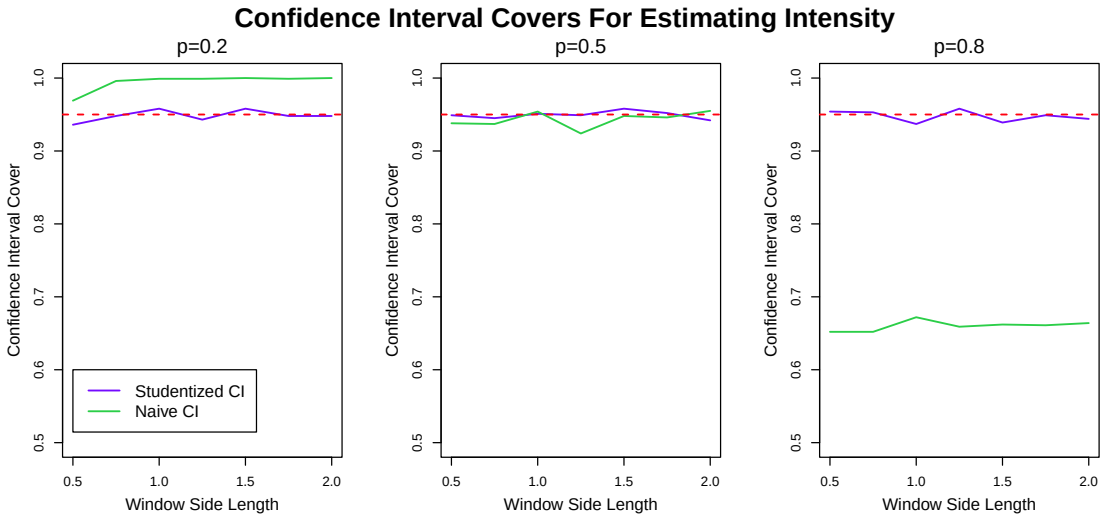

Figure 5.3: Line plots showing the cover of confidence intervals calculated for the estimation of the intensity of a homogeneous Poisson point process with a true intensity of $\lambda = 250$. The green lines show the cover using the naive basic bootstrap method, and the purple lines show the results of the studentized confidence intervals. Each plot shows the covers using a thinning parameter of $p = 0.2, 0.5, 0.8$ respectively from left to right. These covers are shown for various sizes of observation window, A, with side lengths varying from 0.5 to 2 in steps of 0.25. For each calculated cover, 1000 simulations are carried out, and for each of these a realisation of a homogeneous Poisson process is generated, and 500 thinned subsamples are selected to form the confidence interval.

This formation of the confidence interval clearly eliminates the effect of the thinning parameter, p, which is encoded into r_s and d_s, the number of points retained and discarded respectively. Similarly to Figure 5.1, Figure 5.2 shows the covers calculated for the estimating of the intensity of a Poisson distribution with intensity, $\lambda = 10$. However in this case, we now used the studentized confidence interval, derived in Equation 5.1. We can see that this method performs much better, with the covers for $p = 0.2, 0.5, 0.8$ all sitting around the desired value of 0.95 show by the red dashed line. The coverage is no longer dependent on the thinning probability, and the convergence to the desired value occurs quite quickly.

5.3 Example: Intensity of Homogeneous Poisson Point Process

The same theory also applies for various statistics of spatial point processes. Here we will present the simple example of estimating the intensity of a homogeneous Poisson point process. We have seen already in Section 4.5 that this intensity estimator satisfies Theorem 3, and so we can expect to form accurate confidence intervals as in the case of the Poisson distribution.

Figure 5.3 shows how both the basic bootstrap and studentized methods of forming confidence intervals perform. Very similarly to the case of a Poisson distribution, the basic bootstrap covers, shown in green, are dependent on the thinning probability which is not desirable. The studentized confidence intervals on the other hand perform well for all values shown here, $p = 0.2, 0.5, 0.8$. This shows that the use of thinning to obtain confidence intervals is applicable to spatial point processes. In the following chapters we will explore where it is particularly useful, and where it fails.

Chapter 6

Ripley's K

In this chapter, we will explore how well thinning performs when applied to Ripley's K estimates. We focus in particular on Ripley's K here because it is a widely used second moment statistic that provides information on the clustering of the underlying process. We have so far shown the validity of the thinning method for first-order intensities and the second-order median of a probability distribution, so it will be useful to extend this to second-order statistics in the spatial setting.

6.1 Results

6.1.1 Basic Bootstrap Confidence Intervals

In the previous chapter, we showed that, given Theorem 6 holds, we can form correct covers using studentized confidence intervals with scaled variance. We also saw that the naive (basic bootstrap) confidence interval doesn't produce good results since we typically cannot assume that $\hat{\theta}_s - \hat{\theta}$ approximates $\hat{\theta} - \theta$, especially in the case that $p \neq 0.5$. We will now explore how these methods perform for estimating Ripley's K with CSR, clustered and regular processes.

Figure 6.1 shows the covers of confidence intervals formed using the naive basic bootstrap. Thinning parameters of $p = 0.2, 0.5, 0.8$ are shown in purple, orange and green respectively. We use $\alpha = 0.05$, and the red dotted line at 0.95 represents the desired cover. Since Ripley's K requires a radius parameter, we consider values between 0.01 and 0.14 with steps of 0.01. We use the same simulation setting as Loh and Stein (2004). It is clear that these covers are dependent on the thinning parameter, which is not ideal since the true underlying process would need to be known in order to find the optimal p. A clustered Neyman-Scott process requires a larger p than the other cases, which suggests that the optimal value of p is dependent on the amount of clustering, but this makes the estimation of K redundant. Note that the four methods explored in Loh and Stein (2004) have a similar issue with respect to the subregion size.

6.1.2 Studentized Confidence Intervals

Figure 6.2 shows the confidence interval covers for the same setting as before, but now using a studentized confidence interval with scaled variance. We see that the dependence on the thinning parameter, p, still persists meaning that Theorem 6 does not hold. Since a homogeneous Poisson point process is made up of i.i.d observations, the reason the theorem is not satisfied may be that Ripley's K is not a linear functional.

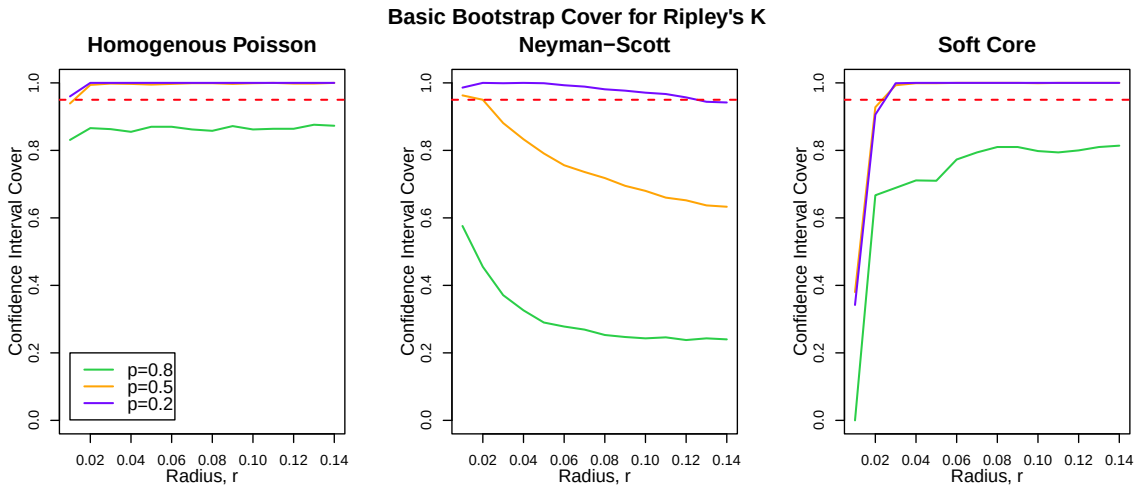

Figure 6.1: Line plots showing the cover of confidence intervals for estimates of Ripley's K using the basic bootstrap. Covers for thinning parameters of $p = 0.2, 0.5, 0.8$ are plotted with purple, orange and green lines respectively. The first subplot shows these simulations carried out using a homogeneous Poisson Point Process with $\lambda = 250$. The second subplot uses a Neyman-Scott Matérn cluster process with $\lambda_p = 25$ and $\lambda_d = 10$. The third subplot uses a Soft Core process with $\lambda_b = 500$, and $F(\rho) = 800\rho$ for $0 < \rho < 1$, and a random mark m uniform on $[0, 1]$. For each of these processes, Ripley's K is calculated with radius parameters varying from 0.01 to 0.14 in steps of 0.01 (same as Loh and Stein (2004)). For each cover, 1000 simulations are carried out, and for each of these 500 thinned subsamples are generated to form the confidence interval.

In the case of a homogeneous Poisson process, we see that thinning provides a linear cover with respect to the radius, which is a benefit over the marked point, subsets and tiling methods described in Loh and Stein (2004). This property is desirable since it means the optimal thinning parameter does not need to be re-calibrated for different radii.

For a clustered process, the cover seems to decrease and then plateau, similar to the tiling method. This behavior may be linked to the amount of clustering in the underlying process. The marked point method and subsets method seem to generally give a more accurate cover for all subregion sizes tested, meaning that the thinning method is not useful for obtaining confidence intervals for Ripley's K.

6.2 Distribution of Thinned Estimates

The studentized interval shown above does not produce the desired cover, meaning that the result shown in Theorem 3 does not hold for Ripley's K. More precisely, the variance of the thinned estimates is not $\sqrt{r_s/d_s}$ times that of the estimates of the entire observed data. This may be because Ripley's K cannot be written as a linear statistical functional, as required. This is difficult to show analytically, but we will show that the estimates do not have the desired distributions through simulation.

6.2.1 Normal Distribution of Thinned Estimates

Firstly, we can check whether $\sqrt{n}\left(\hat{K}_s(h) - \hat{K}(h)\right)$ has an asymptotically normal distribution. This is done by simulating a homogeneous Poisson point process, estimating $\hat{K}(h)$ from

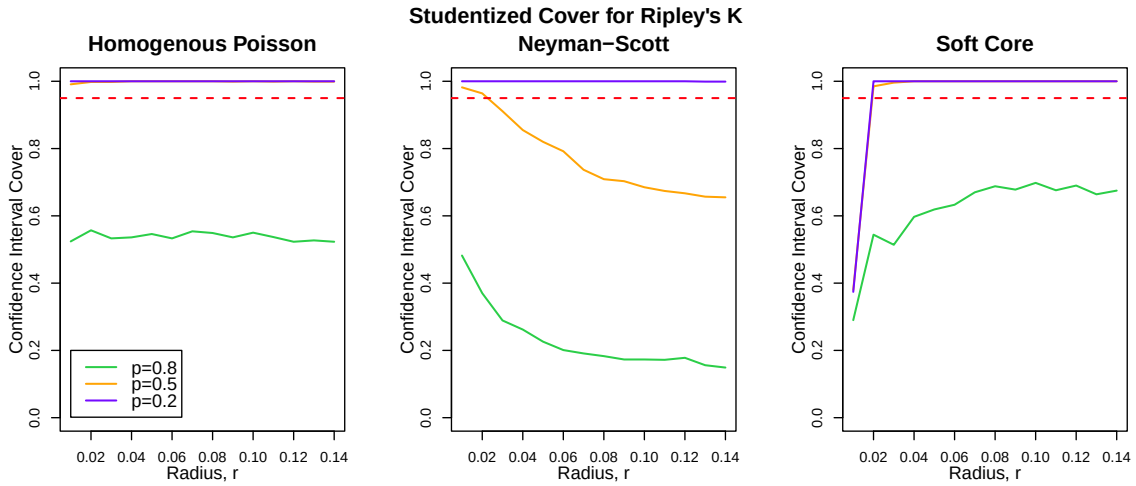

Figure 6.2: Line plots showing the cover of confidence intervals for estimates of Ripley's K using the studentized confidence intervals with a scaled variance. Covers for thinning parameters of $p = 0.2, 0.5, 0.8$ are shown by purple orange and green lines respectively. The underlying point processes of each subplot are the same as those in Figure 6.1.

this realisation, then generating 500 thinned samples and estimating $\hat{K}_s(h)$ from each of these. The value of $\sqrt{n}\left(\hat{K}_s(h) - \hat{K}(h)\right)$ is then stored. The Poisson process used is one with an intensity of $\lambda = 10,000$, so n is large enough to observe the asymptotic effects. The QQ-plots with thinning parameters of $p = 0.2, 0.5, 0.8$ are shown in Figure 6.3.

We see that for a large n, the distributions are roughly normal, however it is clear that there is a slight curvature to the plots. More specifically, the tails tend to curve above the line, and this is more visible for $p = 0.2$ than for the larger thinning parameters. This shape indicates that the samples are skewed to the right, meaning that more than half of the points are greater than 0. This effect is more prominent for lower values of the thinning probability, p, possibly because the thinned samples contain fewer points. Overall, we see that the distribution of $\sqrt{n}\left(\hat{K}_s(h) - \hat{K}(h)\right)$ is close to having an asymptotically Gaussian distribution, however is not consistent at the tails. This could be the reason that thinning does not produce the desired cover, along with the lack of i.i.d data.

6.2.2 Variance of Thinned Estimates

We now consider the variance of these quantities, and more specifically their relationship. Let

$$var\left(\sqrt{n}(\hat{K}(h) - K(h))\right) = \hat{\sigma}^2,$$

$$var\left(\sqrt{n}(\hat{K}_s(h) - \hat{K}(h))\right) = \hat{\sigma}_s^2.$$

Then if Theorem 3 holds, it suggests that for large n,

$$\frac{r_s \sigma_s^2}{d_s} \approx \frac{p\hat{\sigma}_s^2}{1-p} \approx \hat{\sigma}^2.$$

We can determine whether this relationship holds through simulation, as in the second subplot of Figure 6.4. This is done by simulating a homogeneous Poisson point process of intensity $\lambda = 250$ on a unit square 10000 times, and calculating $\hat{K}(h)$ on each of these with $h = 0.1$. The variance of all of these samples is then taken to be $\hat{\sigma}^2$. To approximate $\hat{\sigma}^2$ using $\hat{\sigma}_s^2$, we

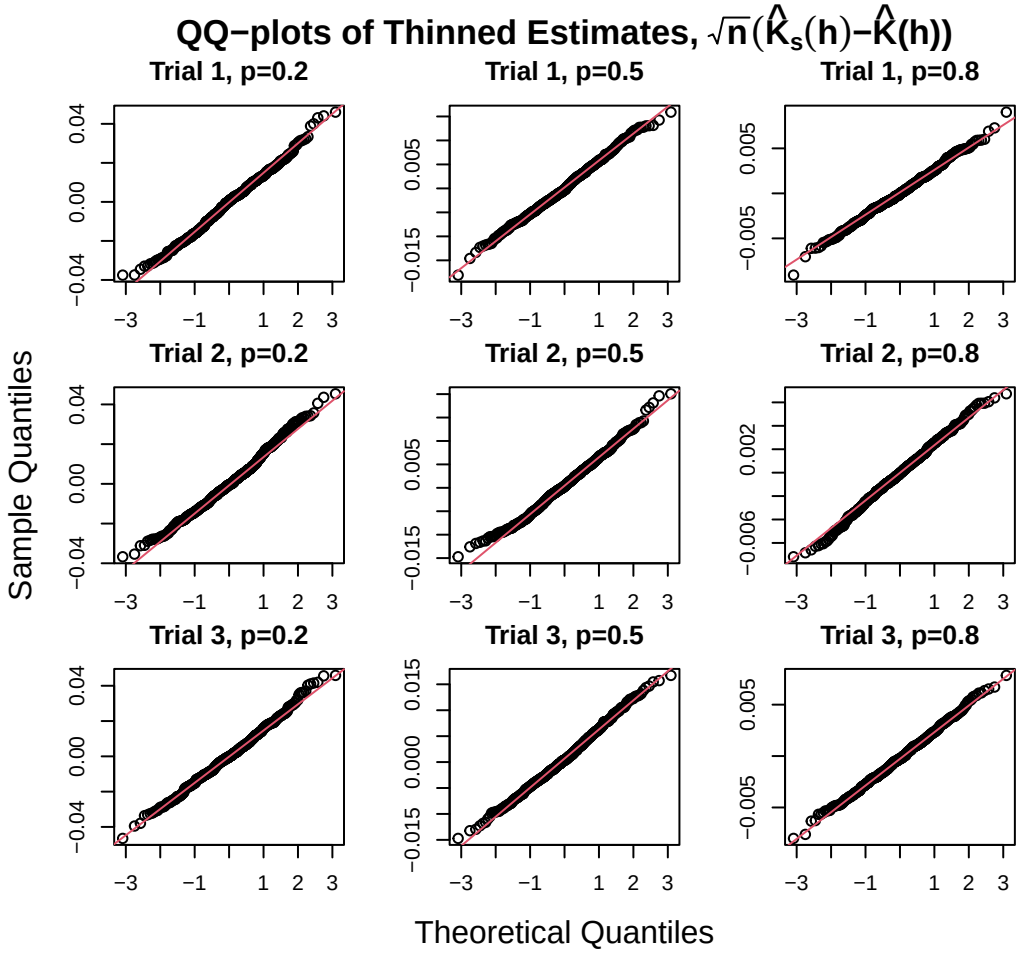

Figure 6.3: QQ-plots of Ripley's K estimates on thinned point processes after normalisation, $\sqrt{n}(\hat{K}_s(h) - \hat{K}(h))$. Values are calculated by simulating a realisation of a homogeneous Poisson point process with intensity $\lambda = 10000$ on a unit square, and generating 500 thinned processes. Thinned Ripley's K is then estimated on each of these, and normalised to obtain the desired value. This is carried out for thinning parameters $p = 0.2, 0.5, 0.8$ which are shown in the subfigures from left to right columns respectively. This process is carried out 3 times, each trial corresponding to a row.

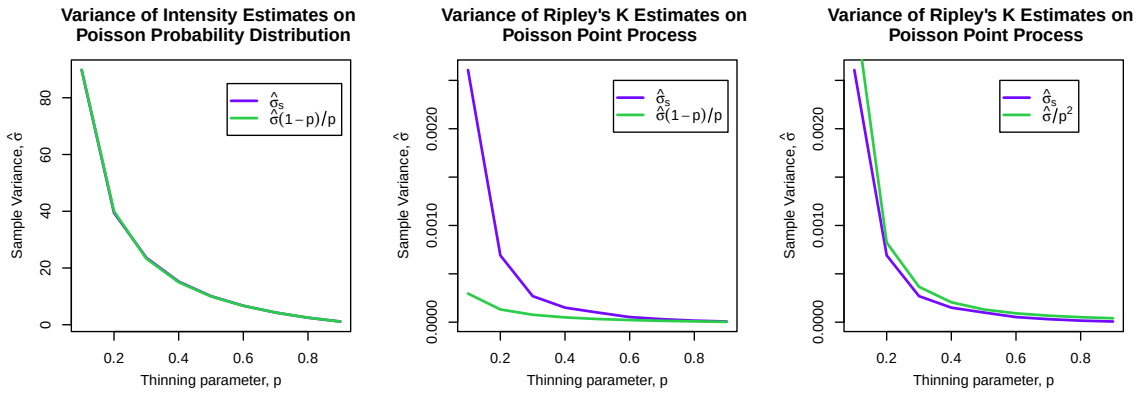

Figure 6.4: Line plots comparing the variance of the estimator of a statistic versus the variance of the thinned replication. The first subplot shows the case where the statistic is the intensity of a Poisson probability distribution with $\lambda = 10$. The scaling used is the one derived in Theorem 3. The second and third plots consider estimates of Ripley's K with radius $h = 0.1$ on homogeneous Poisson point processes with $\lambda = 250$. The second subplot again uses the scaling derived in Theorem 3, and the third subplot uses the scaling formed from the analytic variance.

simulate the same process 100 times, and for each of these, generate 500 thinned subsamples and calculate the thinned replication $\hat{K}_s(h)$, then we take the variance, $\hat{\sigma}_s^2$. Figure 6.4 shows the variance of the thinned estimates, $\hat{\sigma}^2$, scaled by $(1-p)/p$ for a range of p from 0.1 to 0.9, which should be close to $\hat{\sigma}_s^2$ if the result holds. We see that this relationship does not hold for Ripley's K. For comparison, the first subplot of Figure 6.4 shows the same process carried out but with a Poisson probability distribution where the statistic is the intensity. We have seen in this case that the Theorem 3 holds, and as a result, the two quantities are almost equal at all values of p.

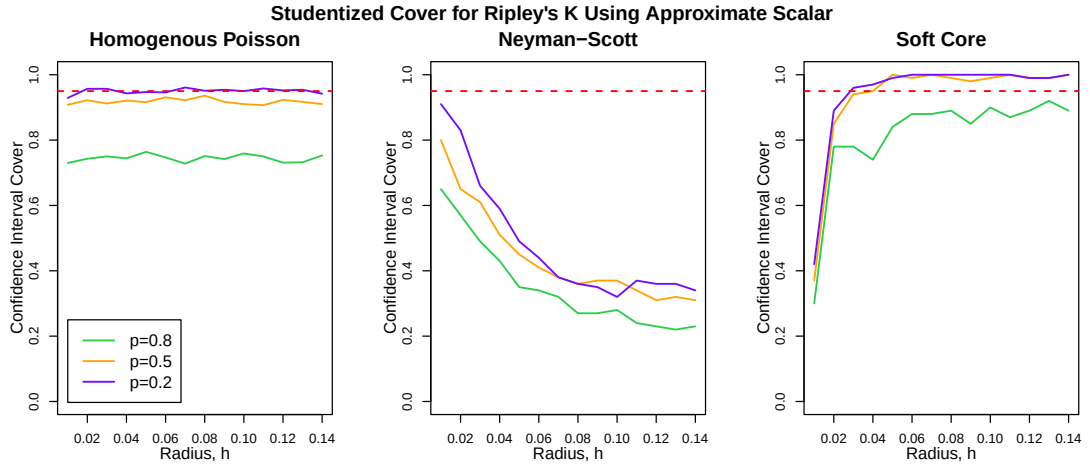

Figure 6.5: Line plots showing the cover of confidence intervals for estimates of Ripley's K using the studentized confidence intervals with scaling of $\hat{\sigma}^2 \approx p^2\hat{\sigma}_s^2$. Simulation setting is the same as that of Figure 6.1.

6.3 Finding Appropriate Scaling

It is clear that the distribution derived in Chapter 4 does not apply to Ripley's K. In particular, the variance of the thinned replications is not what we expect. We will attempt to find a different scalar for the variance of the thinned applications using the analytic form of the variance of Ripley's K. We know that under a homogeneous Poisson point process, the variance of $\hat{K}(h)$ takes the form seen in Equation 2.4. The only difference in the variance of a thinned replication, $\hat{K}_s(h)$, is that the number of points changes to r_s:

$$\hat{\sigma}_s^2 = var(\hat{K}_s(h)) = \frac{2|A|^2\beta(h)}{r_s^2}(1 + 0.305\gamma(h) + \beta(h)(-1 + 0.0132r_s\gamma(h)))$$

where $\beta(h) = \frac{\pi h^2}{|A|}$ and $\gamma(h) = \frac{|\partial A|h}{|A|}$. If we assume that the final two terms are small, we can approximate the ratio of the variances:

$$\frac{\hat{\sigma}_s^2}{\hat{\sigma}^2} = \frac{var(\hat{K}_s(h))}{var(\hat{K}(h))} \approx \frac{n^2}{r_s^2} \approx \frac{1}{p^2} \implies \hat{\sigma}_s^2 \approx \frac{\hat{\sigma}^2}{p^2}.$$

This relationship is shown in the final subfigure of Figure 6.4. We see that the two quantities match up much better than with the previous scaling constant, however, they do not align as well as the first subplot, since we have made a strong assumption to carry out this approximation.

We can use this new scaling to compute confidence interval covers which are plotted in Figure 6.5. The simulations setting is the same as the one described for Figure 6.2, but of course with the the variance of the thinned replications being scaled differently. We see that this significantly improves the covers. Firstly, covers vary less with respect to the thinning parameter, p, which is useful as ideally, we would like to eliminate dependence so that we do not require any additional information about the distribution. We see that the dependence is still present, but the effect is reduced. Alongside this, the covers are generally closer to the desired level of $1 - \alpha$, especially for the Poisson and Soft Core experiments. The covers reach the extreme value of 1 less often, meaning that the width of the confidence intervals has decreased.

Overall, use of this scalar results in an improvement, however we still face some significant issues with thinning. It is clear from Figure 6.5 that the same thinning parameter performs very differently for different amounts of clustering. This renders the method useless in this case, since the use of Ripley's K is the measure the amount of clustering, which is typically unknown.

6.4 Issues

We have seen in this chapter that thinning is not an effective resampling method when the statistics of interest is Ripley's K, and even more so when the underlying data is not i.i.d. Intuitively, we can think of Ripley's K as a sum of random variables, where each random variable is the number of points within a radius h of a point, and we sum over each point forming the center of the circle. It is intuitively clear from this that Ripley's K is not a linear statistical functional. However, we showed in Section 4.5 that linearity is not a necessary condition for Theorems 3 and 6 to hold. Additional issues may arise from the edge effects of the bounded observation window, or the fact that Ripley's K is more dependent on the interactions between points than the median is.

Chapter 7

Inhomogenous Intensity Estimates

So far, we have demonstrated the usefulness of the thinning method on distributions, and shown that there are issues when we try to apply the method to second order statistics of point processes. We will now explore whether thinning can be used to find confidence intervals for first order statistics on point processes, specifically inhomogeneous intensities, $\lambda(x, y)$. We will experiment with two methods which are popular in the literature for inhomogeneous intensity estimates, the parametric maximum likelihood estimate, and the non-parametric kernel density estimator, both of which are outlined in Chapter 2. Throughout this chapter we will consider an inhomogeneous Poisson point process with log-linear intensity, as described in Equation 2.5.

7.1 Experiments

7.1.1 Parametric Estimator

In order to calculate confidence intervals for parametric estimates of the intensity, we will use a first-order potential, meaning the intensity is of the form $\lambda(x, y) = \exp(a + bx)$. The values we are attempting to estimate are the coefficients a and b. We will attempt to use both the basic bootstrap and studentized confidence intervals like in the previous chapter, to deduce whether the distribution derived in Chapter 4 holds in this case.

Figure 7.1 shows the covers resulting from the basic bootstrap, as the side length of the observed window, A, increases, and therefore the number of observed points increases. Out of the thinning probabilities shown here, $p = 0.2, 0.5, 0.8$, only a value of $p = 0.5$ produces the desired behavior for estimating b. It is clear that the cover is dependent on the thinning probability in the other cases.

On the other hand, Figure 7.2 shows the analogous covers, but this time using studentized confidence intervals, which assume that the results in Chapter 4 hold. As expected, the performance has significantly improved using this method, and the covers are no longer dependent on p, meaning that Theorem 6 holds in the higher dimensional spatial setting.

7.1.2 Non-parametric Estimator

In order to obtain a non-parametric estimate of of the intensity function $\lambda(x, y) = \exp(a + bx)$, we use the kernel density estimate as described in Section 2.3.2. This method produces a prediction for the value of the intensity at each point on a chosen grid. Therefore, in this case we will formulate the experiment slightly differently to the the way we have previously.

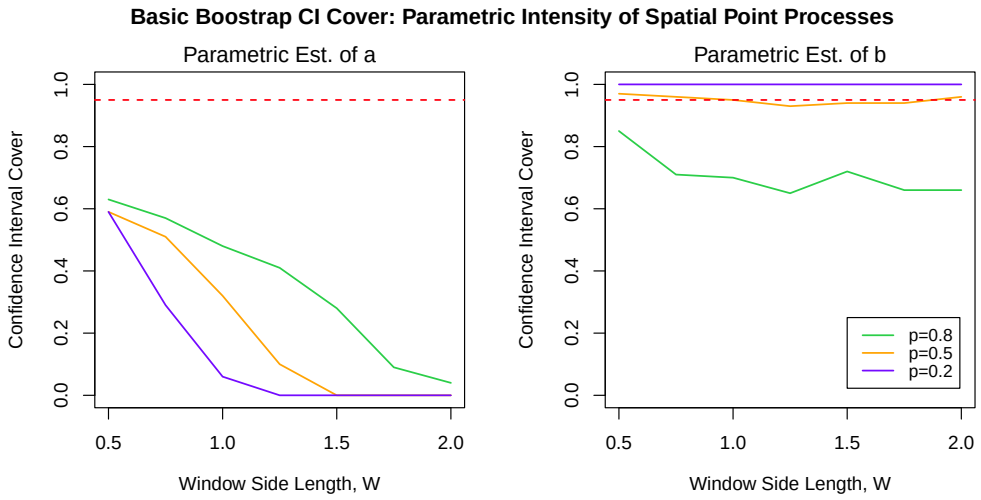

Figure 7.1: Line plots showing the empirical coverage using the basic bootstrap confidence interval as the size of the observed window, A, increases. This side length of the window varies from $W = 0.5$ to $W = 2.0$ in steps if 0.25. The process used is an inhomogenous Poisson point process with an intensity function of $\lambda(x, y) = \exp(a + bx)$, and the values being estimated are a and b. The left plot shows the covers for a, and the right plot shows the covers for b. 100 simulations are carried out, each of which use 500 thinned subsamples. The purple, orange and green line represent thinning parameters of $p = 0.8, 0.5, 0.2$ respectively, and the red dashed line is the desired cover of 0.95.

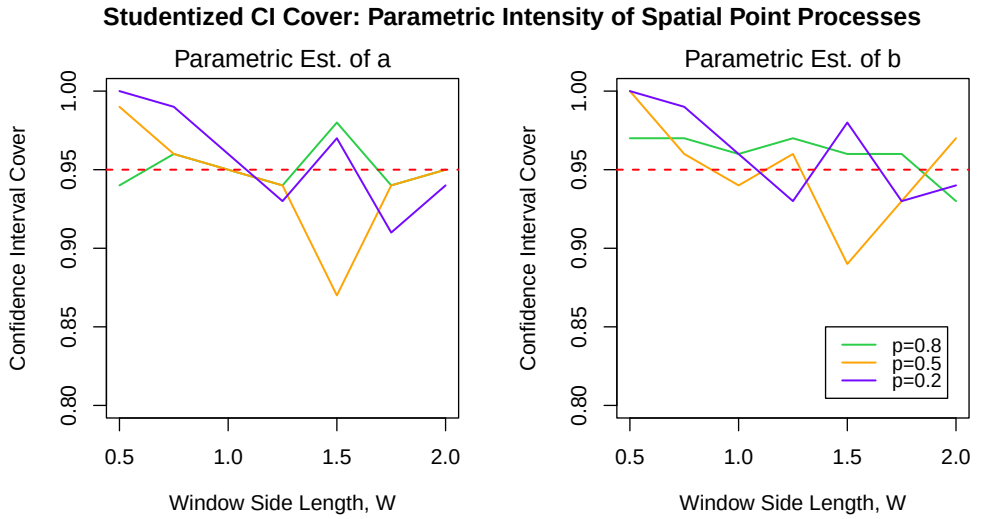

Figure 7.2: Line plots showing the empirical coverage using the studentized confidence interval as the size of the observed window, A, increases. This side length of the window varies from $W = 0.5$ to $W = 2.0$ in steps if 0.25. The process used is an inhomogenous Poisson point process with an intensity function of $\lambda(x, y) = \exp(a + bx)$, and the values being estimated are a and b. The left plot shows the covers for a, and the right plot shows the covers for b. 100 simulations are carried out, each of which use 500 thinned subsamples. The purple, orange and green line represent thinning parameters of $p = 0.8, 0.5, 0.2$ respectively, and the red dashed line is the desired cover of 0.95.

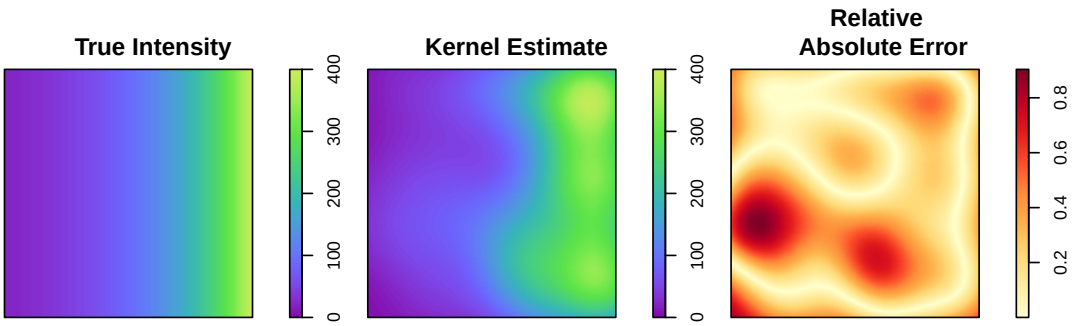

Figure 7.3: Heatmaps showing the true intensity, estimated intensity, and relative absolute error of the estimates. These use an example inhomogeneous intensity function of $\lambda(x, y) = \exp(3+3x)$ on the unit square. The kernel density estimate is calculated using one realisation of a Poisson point process with this intensity, and the relative absolute error is calculated as $|\lambda(x, y) - \hat{\lambda}(x, y)|/\lambda(x, y)$ at each point (x, y).

We will continue to carry out thinning in the same way across the window to obtain B thinned replications, which are each a grid of intensity predictions. For each of the points on the grid, we use the corresponding predictions across all of the thinned replications to form a confidence interval, and use this to calculate a cover for each point.

Firstly, Figure 7.3 shows heat maps of an example intensity function $\lambda(x, y) = \exp(3 + 3x)$ on a unit square, the kernel density estimate using a realisation of an inhomogeneous Poisson point process with this intensity, and the relative absolute error between the true and estimated intensity. Note that this is a kernel estimate arising from one random realisation of the process. Different realisations will of course produce varying estimates, but the errors typically follow a similar pattern. The edge effects result in higher errors along the borders, and there are typically some patches with high errors within the region. Understanding the accuracy of the estimates will help us to interpret the structure of the covers.

Figure 7.4 shows the cover at each point on a 128×128 grid on the unit square, using the studentized confidence interval. The first column shows the covers with the colour bar ranging from 0 to 1, and the second column shows the same values but with the colour bar restricted between 0.93 and 0.97. Since we have $\alpha = 0.05$ as before, the coloured regions in the second column are the regions where the covers are close to the desired level. This is shown for thinning parameters taking values $p = 0.2, 0.5, 0.8$. There are no significant differences for different values of p, and the variation we do see is most likely down to randomness. So similarly to the parametric estimator, this tells us that Theorem 6 holds in the spatial setting, when the statistic of interest is an inhomogeneous intensity.

We see that the covers at the borders are consistently outside of the range $[0.93, 0.97]$, as well as some patches within the window. The areas where the covers are poor are similar to the regions where the kernel density estimate itself is poor, as shown in Figure 7.3. Since the studentized confidence interval is centred around the kernel density estimate using all of the observed data, it is plausible that an inaccurate estimate produces a skewed and therefore also inaccurate confidence interval. In the more central regions where the kernel density estimate is more reliable, we have covers which are very close to the desired $1 - \alpha$. Overall, there are large regions over which we can produce reliable confidence intervals, however, if we do not have an idea of whether the initial prediction is accurate, then it is difficult to know whether the resulting confidence interval is useful.

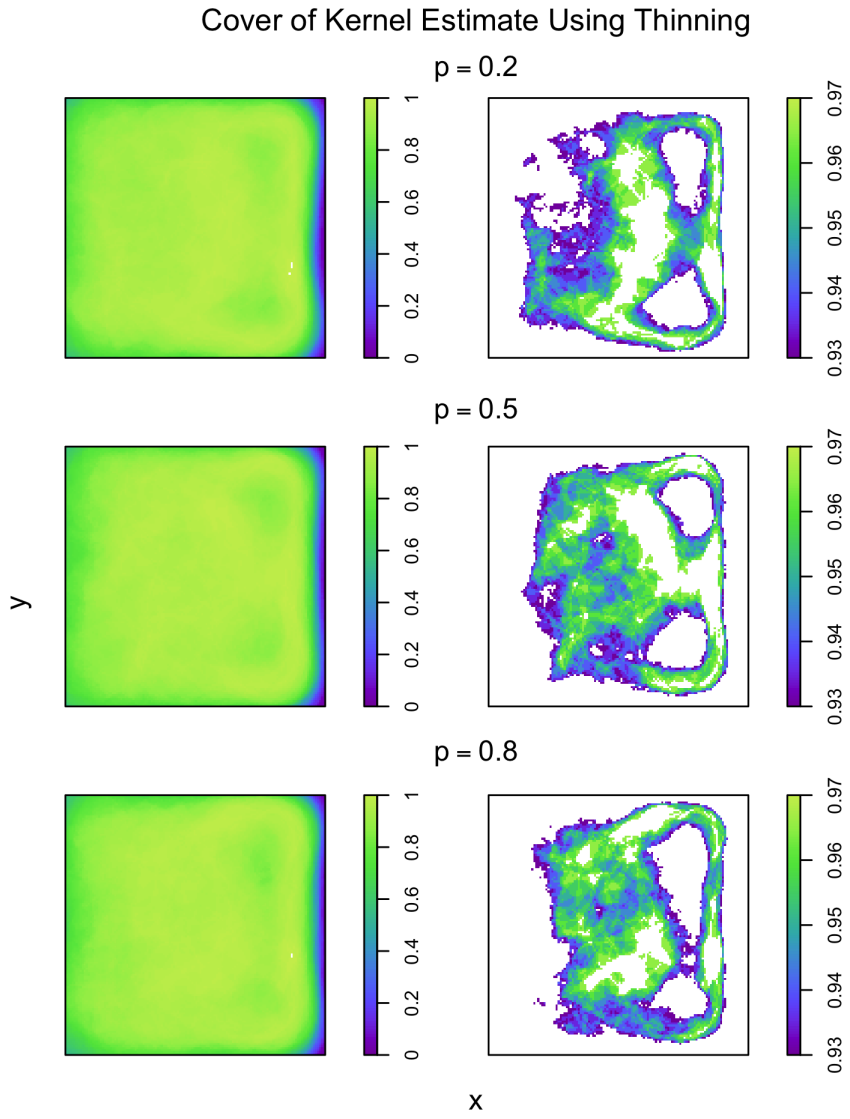

Figure 7.4: Heatmaps showing covers for the inhomogenous intensity of a Poisson point process, $\lambda(x, y) = \exp(3 + 3x)$, estimated using the kernel density method. The rows show results for the thinning parameter talking values $p = 0.2, 0.5, 0.8$ respectively. The covers are formed using 200 realisations of the point process. For each of these, we generate 500 thinned samples and an intensity estimate over the entire grid for each sample. The 500 samples are combined to form a studentized confidence interval for each point on the grid, and the process is repeated for each simulation. At each point, we compare the confidence intervals to the true intensity, and calculate a cover. The first column shows a colour bar ranging from 0 to 1, and the second from 0.93 to 0.97. The desired cover is $1 - \alpha = 0.95$.

Chapter 8

Conclusion

In summary, we have explored the application of thinning in various settings, where it performed well in some cases and poorly in others.

Firstly, we proved that thinning is a valid method to use for approximating the distribution of a linear statistical functional of a distribution. In particular, Theorem 6 proved that the distribution of the normalised thinning replication $\hat{\theta}_s$ approaches the normalised distribution of $\hat{\theta}$ which is desirable. Although this is a positive result, Efron's bootstrap is a very popular method for these cases, which is the current standard in the literature with few issues with i.i.d one-dimensional data. One advantage that thinning may have over bootstrapping is the computational cost, especially if the size of the bootstrap samples is designed to be greater than the number of observed points, n. For thinning, the selection of one point consists of generating a uniform random variable, u, and accepting the point if $u < p$, where p is the thinning parameter. On the other hand, the selection of one point in a bootstrap sample consists of sampling from a discrete uniform distribution of the observed points, possibly using inversion sampling (Tang and Ng, 2011). The differences between these two procedures are minor, so if there is any difference in the computational complexity of the methods, it will arise as a result of the size of the bootstrap samples taken. We have also shown that if $p = 0.5$, the two methods are equivalent, and so if one is computationally more efficient in some circumstances, then it can be used knowing the distribution is the same in both cases.

When applying the thinning method to higher dimensional spatial data, we observed that it worked well for estimating both the homogeneous or inhomogeneous intensity function of a Poisson point process. We also demonstrated explicitly that the asymptotic distribution of thinned intensity replications have the desired distribution, meaning that Theorem 6 holds in this spatial setting. However, we saw through numerical experiments that the same does not hold for Ripley's K.

Sufficient conditions for the desired asymptotic distribution to hold are that the statistic of interest is a linear functional, and the underlying data is independent and identically distributed. Of course, a clustered or regular process does not obey the i.i.d assumption, however, neither does an inhomogeneous Poisson point process, but we saw in Chapter 7 that the result still holds in this case. This suggests that i.i.d data is not a necessary condition. The most prominent roadblock with thinning is the requirement that the statistic is linear. In particular, Ripley's K is a function of the interactions between points and therefore does not obey this. Howver, recall that we saw in Section 4.5 that linearity of the statistical functional was also not a necessary condition, so there may be other useful non-linear statistics for which thinning is a useful resampling method.

Further Work

We observed in Figure 6.2 that in some cases, the confidence interval covers are a function of the radius parameter of Ripley's K. In particular, we see this behavior for clustered and regular processes, but not for the CSR Poisson. It is possible that when the underlying data is not i.i.d, the scaling constant is also dependent on the amount of clustering/regularity. This could be investigated further by computing the covers similarly to Figure 6.2, with the radius h kept constant, and the amount of clustering in the underlying data being varied.

More generally, it is possible that the observation that the method applies to intensity but not Ripley's K generalises in the spatial setting to first and second-order statistics, or linear and non-linear statistical functionals. So far we have only proven that the desired asymptotic distribution holds for linear statistics, but we do not understand the asymptotic distributions of higher-order statistics. This work can also be extended of course by assessing the covers arising from various other more complex statistics and underlying distributions, in particular, it would be interesting to find out if there are any non-linear statistics for which thinning is useful in the spatial setting.

Another possible extension of this work is to adapt the thinning method such that it performs better for statistics such as Ripley's K. This could be by adopting a similar idea to the block bootstrap, where groups of points are considered as units to preserve interactions between pairs. Moreover, it may be possible to adapt the existing block bootstraps such that they are independent of other parameters such as the size of the blocks. This could be done using the same ideas introduced here for thinning, where we considered the relationship between the variance of the estimate and thinned replication of the statistic.

References

J. M. Loh and M. L. Stein. Bootstrapping a spatial point process. *Statistica Sinica*, 14(1): 69–101, 2004.

M. H. Quenouille. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):68–84, 1949.

M. H. Quenouille. Notes on bias in estimation. *Biometrika*, 43(3/4):353–360, 1956.

J. Tukey. Abstracts of Papers. *The Annals of Mathematical Statistics*, 29(2):614 – 623, 1958.

B. Efron and R. Tibshirani. An Introduction to the Bootstrap, 1993.

C. F. J. Wu. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14(4):1261 – 1295, 1986.

Y. G. Berger and C. J. Skinner. A Jackknife Variance Estimator for Unequal Probability Sampling. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67 (1):79–89, December 2004.

D. V. Hinkley. Jackknifing in unbalanced situations. *Technometrics*, 19(3):285–292, 1977.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, January 1979.

B. Efron. Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, 6(4):1971 – 1997, 2012.

H. R. Künsch. The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*, 17(3):1217 – 1241, 1989.

S. Wang. Optimizing the smoothed bootstrap. *Annals of the Institute of Statistical Mathematics*, 47(1):65–80, January 1995.

N. Cressie. *Statistics for Spatial Data*. A Wiley-interscience publication. J. Wiley, 1991.

K. Singh. On the asymptotic accuracy of efron's bootstrap. *The Annals of Statistics*, 9(6): 1187–1195, 1981.

P. Hall. Resampling a coverage pattern. *Stochastic Processes and their Applications*, 20(2): 231–246, 1985a.

P. Hall. On symmetric bootstrap confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(1):35–45, 1988.

T. Mattfeldt, H. Häbel, and F. Fleischer. Block bootstrap methods for the estimation of the intensity of a spatial point process with confidence bounds. *Journal of microscopy*, 251, 05 2013.

M. I. Borrajo, W. González-Manteiga, and M.D. Martínez-Miranda. Bootstrapping kernel intensity estimation for inhomogeneous point processes with spatial covariates. *Computational Statistics Data Analysis*, 144:106875, 2020.

M. A. Kiskowski, J. F. Hancock, and A. K. Kenworthy. On the use of ripley's k-function and its derivatives to analyze domain size. *Biophysical Journal*, 97(4):1095–1103, 2009.

M. Jafari Mamaghani, M. Andersson, and P. Krieger. Spatial point pattern analysis of neurons using ripley's k-function in 3d. *Frontiers in Neuroinformatics*, 4, 2010.

W. Hamer, D. Knitter, S. Grimm, B. Serbe, B. Eriksen, O. Nakoinz, and R. Duttmann. Location modeling of final palaeolithic sites in northern germany. *Geosciences (Switzerland)*, 9, 10 2019.

J. Negre, F. Muñoz, and J. An Barceló. A Cost-Based Ripley's K Function to Assess Social Strategies in Settlement Patterning. *Journal of Archaeological Method and Theory*, 25 (3):777–794, September 2018.

Y. Ge, M. Sun, and Y. Pu. Geographic information system-based edge effect correction for ripley's k-function under irregular boundaries. *Geographical Research*, 57(4):436–447, 2019.

J. M. Loh. Bootstrapping an inhomogeneous point process. *Journal of Statistical Planning and Inference*, 140(3):734–749, 2010.

P. J. Diggle. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2013.

B. D. Ripley. *Spatial Statistics*. Wiley Series in Probability and Statistics. Wiley, 1981.

J. Neyman and E. L. Scott. Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(1):1–43, 1958.

D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic Geometry and Its Applications*. Inorganic Chemistry. Wiley, 1995.

D. Stoyan. Statistical analysis of spatial point processes: A soft-core model and cross-correlations of marks. *Biometrical Journal*, 29(8):971–980, 1987.

M. N. M. van Lieshout. *Theory of Spatial Statistics: A Concise Introduction*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2019.

O. C. Ibe. 15 - Markov Point Processes. In Oliver C. Ibe, editor, *Markov Processes for Stochastic Modeling (Second Edition)*, pages 453–480. Elsevier, Oxford, second edition edition, 2013.

S. Shekhar, P. Zhang, and Y. Huang. Spatial Data Mining. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 848–8549. Springer US, Boston, MA, 2005.

P. A. W. Lewis. A branching poisson process model for the analysis of computer failure patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(3):398–456, 1964.

M. S. Bartlett. The spectral analysis of two-dimensional point processes. *Biometrika*, 51 (3/4):299–311, 1964.

N. R. Aryal and O. D. Jones. Spatial-temporal rainfall models based on poisson cluster processes. *Stochastic Environmental Research and Risk Assessment*, 35(12):2629–2643, December 2021.

G. Faÿ, B. González-Arévalo, T. Mikosch, and G. Samorodnitsky. Modeling teletraffic arrivals by a Poisson cluster process. *Queueing Systems*, 54(2):121–140, October 2006.

M. Westcott. On existence and mixing results for cluster point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(2):290–300, 1971.

H. ElSawy and E. Hossain. A modified hard core point process for analysis of random csma wireless networks in general fading environments. *IEEE Transactions on Communications*, 61(4):1520–1534, 2013.

D. Stoyan and A. Penttinen. Recent applications of point process methods in forestry statistics. *Statistical Science*, 15(1):61–78, 2000.

B. Matern. *Spatial Variation*. Lecture Notes in Statistics. Springer New York, 1986.

B. D. Ripley. *Statistical Inference for Spatial Processes*. Cambridge University Press, 1988.

A. J. Baddeley, J. Møller, and R. Waagepetersen. Non-and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54(3):329–350, 2000.

R. Waagepetersen. Estimating functions for inhomogeneous spatial point processes with incomplete covariate data. *Biometrika*, 95(2):351–363, 2008.

S. L. Rathbun, S. Shiffman, and C. J. Gwaltney. Modelling the effects of partially observed covariates on poisson process intensity. *Biometrika*, 94(1):153–165, 2007.

S. L. Rathbun and N. Cressie. A space-time survival point process for a longleaf pine forest in southern georgia. *Journal of the American Statistical Association*, 89(428):1164–1174, 1994.

P. Diggle. A kernel method for smoothing point process data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(2):138–147, 1985.

P. Hall. Resampling a coverage pattern. *Stochastic Processes and their Applications*, 20(2): 231–246, 1985b.

S. N. Lahiri. On the moving block bootstrap under long range dependence. *Statistics Probability Letters*, 18(5):405–413, 1993.

D. N. Politis and J. P. Romano. *A Circular Block-resampling Procedure for Stationary Data*. Purdue University. Department of Statistics, 1991.

R. Y. Liu. Moving blocks jackknife and bootstrap capture weak dependence. 1992.

P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1986.

H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17, 06 1993.

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

J. M. Loh and M. L. Stein. Spatial bootstrap with increasing observations in a fixed domain. *Statistica Sinica*, 18:667–688, 04 2008.

D. N. Politis and J. P. Romano. Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions. *The Annals of Statistics*, 22(4):2031 – 2050, 1994.

W. J. Branu and R. J. Kulperger. A bootstrap for point processes. *Journal of Statistical Computation and Simulation*, 60(2):129–155, 1998.

R. von Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.

B. Fristedt and L. Gray. *A Modern Approach to Probability Theory*. Birkhauser, 1997.

R. S. Strichartz. *The way of analysis*. World Publishing Corporation, 2010.

L. T. Fernholz. *Von Mises calculus for statistical functionals*. Number 19 in Lecture notes in statistics. Springer, 1983.

R. J. Serfling. *Approximation theorems of mathematical statistics*. Wiley series in probability and mathematical statistics : Probability and mathematical statistics. Wiley, New York, NY [u.a.], [nachdr.] edition, 1980.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

C. F. J. Wu. On the asymptotic properties of the jackknife histogram. *The Annals of Statistics*, 18(3):1438–1452, 1990.

J. Hájek. Limiting distributions in simple random sampling from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:361–374, 1960.

H. Wang and J. Zou. A comparative study on sampling with replacement vs poisson sampling in optimal subsampling. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 289–297. PMLR, 13–15 Apr 2021.

W. Feller. *An introduction to probability theory and its applications*, page 150–152. John Wiley, 1968.

G. Pólya. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, 8:171–181, 1920.

H. H. Sohrab. *Basic Real Analysis*. Springer New York, 2014.

P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

J. M. Loh. A valid and fast spatial bootstrap for correlation functions. *The astrophysical journal*, 681(1):726, 2008.

T. J. Sweeting. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, 8(6):1375–1381, 1980.

A. M. Svane, C. Biscio, and R. Waagepetersen. A functional central limit theorem for the k-function with an estimated intensity function, 2023.

K. P. Choi. On the medians of gamma distributions and an equation of ramanujan. *Proceedings of the American Mathematical Society*, 121(1):245–251, 1994.

M. Tang and H. K. T Ng. Inverse sampling. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 688–690. Springer, 2011.

A. DasGupta. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics*. Springer Texts in Statistics. Springer New York, 2011.

Appendix A

Variance of Intensity Estimates

Here we will provide the proof to Theorem 7 in Chapter 4.

Theorem. *The intensity estimate has a variance of*

$$var(\hat{\lambda}) = \frac{\lambda}{|A|},$$

and the thinned intensity estimate of a homogenous Poisson point process has a variance of

$$var(\hat{\lambda}_s) = \frac{\lambda}{|A|p}.$$

Proof. The first of these is straight forward, using the known distribution of $N(A)$:

$$var\left(\hat{\lambda}\right) = var\left(\frac{N_X(A)}{|A|}\right) = \frac{1}{|A|^2} var\left(N_X(A)\right) = \frac{\lambda|A|}{|A|^2} = \frac{\lambda}{|A|}.$$

Finding the variance of the thinned estimate is trickier, since we only know the conditional probability of $r|N = n$, whereas the distribution of r is required. To find this, we use the law of total probability:

$$P(r = x) = \sum_{m=0}^{\infty} P(r = x|N = m)P(N = m).$$

Since the terms where $m < x$ are zero, the index of the sum can begin at x. Using the known distributions 4.12 and 4.13:

$$
\begin{aligned}
P(r=x) &= \sum_{m=x}^{\infty} P(r=x|N=m)P(N=m) \\
&= \sum_{m=x}^{\infty} \binom{m}{x} p^r (1-p)^{m-r} \frac{(\lambda|A|)^m \exp(-\lambda|A|)}{m!} \\
&= \sum_{m=x}^{\infty} \frac{m!}{r!(m-r)!} p^r (1-p)^{m-r} \frac{(\lambda|A|)^m \exp(-\lambda|A|)}{m!} \\
&= \sum_{m=x}^{\infty} \frac{1}{r!(m-r)!} p^r (1-p)^{m-r} (\lambda|A|)^m \exp(-\lambda|A|) \\
&= \frac{p^r \exp(-\lambda|A|)}{r!} \sum_{m=x}^{\infty} \frac{1}{(m-r)!} (1-p)^{m-r} (\lambda|A|)^m \\
&= \frac{(\lambda|A|p)^r \exp(-\lambda|A|)}{r!} \sum_{m=x}^{\infty} \frac{1}{(m-r)!} (1-p)^{m-r} (\lambda|A|)^{m-r}
\end{aligned}
$$

Applying a change of variables $k = m - r$, and recognising the Taylor expansion of an exponential:

$$
\begin{aligned}
P(r=x) &= \frac{(\lambda|A|p)^r \exp(-\lambda|A|)}{r!} \sum_{k=0}^{\infty} \frac{(\lambda|A|(1-p))^k}{k!} \\
&= \frac{(\lambda|A|p)^r \exp(-\lambda|A|)}{r!} \exp\left(\lambda|A|(1-p)\right) \\
&= \frac{(\lambda|A|p)^r \exp(-\lambda|A|p)}{r!}
\end{aligned}
$$

which we recognise as a Poisson probability distribution, such that

$$
r \sim Poi(\lambda|A|p). \tag{A.1}
$$

Using this, we can now find the variance of the thinned estimator:

$$
var\left(\hat{\lambda}_s\right) = var\left(\frac{r}{|A|p}\right) = \frac{1}{(|A|p)^2} var(r) = \frac{\lambda|A|p}{(|A|p)^2} = \frac{\lambda}{|A|p}.
$$

The resulting relationship between the variances of the two estimators $\hat{\lambda}$ and $\hat{\lambda}_s$ is

$$
var\left(\hat{\lambda}_s\right) = \frac{var\left(\hat{\lambda}\right)}{p}. \tag{A.2}
$$

□

Here we provide the proof of Theorem 8 in Chapter 4.

Theorem. *The variance of $\hat{\lambda}_s - \hat{\lambda}$ is given by*

$$
var(\hat{\lambda}_s - \hat{\lambda}) = \frac{(1-p)\lambda}{|A|p}.
$$

Proof.

$$\begin{aligned} var\left(\hat{\lambda}_s - \hat{\lambda}\right) &= var\left(\frac{r}{|A|p} - \frac{N_X(A)}{|A|}\right) \\ &= \frac{1}{|A|^2} var\left(\frac{r}{p} - N_X(A)\right) \\ &= \frac{1}{|A|^2}\left[var\left(\frac{r}{p}\right) + var(N_X(A)) - 2cov\left(\frac{r}{p}, N_X(A)\right)\right] \end{aligned}$$

The first two variances are known using Equation 4.12 and 4.14, so we need to find the covariance term:

$$cov\left(\frac{r}{p}, N_X(A)\right) = E\left(\frac{rN_X(A)}{p}\right) - E\left(\frac{r}{p}\right)E(N_X(A)).$$

DasGupta (2011) shows that

$$E\left(rN_X(A)\right) = p\left(\lambda|A| + \lambda^2|A|^2\right)$$

$$\implies cov\left(\frac{r}{p}, N_X(A)\right) = \lambda|A| + \lambda^2|A|^2 - \lambda^2|A|^2 = \lambda|A|$$

Now, combining these results

$$\begin{aligned} var\left(\hat{\lambda}_s - \hat{\lambda}\right) &= \frac{1}{|A|^2}\left[\frac{\lambda|A|}{p} + \lambda|A| - 2\lambda|A|\right] \\ &= \frac{\lambda(1-p)}{|A|p} \end{aligned}$$

which is the desired result. □