4-22-2020 3:30 PM

Point Process Modelling of Objects in the Star Formation Complexes of the M33 Galaxy

Dayi Li, *The University of Western Ontario*

Supervisor: McLeod, Ian A., *The University of Western Ontario*
Joint Supervisor: Barmby, Pauline, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Statistics and Actuarial Sciences

ABSTRACT

In this thesis, Gibbs point process (GPP) models are constructed to study the spatial distribution of objects in the star formation complexes of the M33 galaxy. The GPP models circumvent the limitations of the two-point correlation function employed in the current astronomy literature by naturally accounting for the inhomogeneous distribution of these objects. The spatial distribution of these objects serves as a sensitive probe in understanding the star formation process, which is crucial in understanding the formation of galaxies and the Universe. The objects under study include the CO filament structure, giant molecular clouds (GMCs) and young stellar cluster candidates (YSCCs). A hierarchical model is adopted to account for the natural formation hierarchy among these objects. The effect of the properties of GMCs on their spatial correlation with YSCCs is also investigated. A Bayesian paradigm is employed for model inference. Potential physical implications are obtained and addressed through model criticism.

KEY WORDS: Spatial statistics, Gibbs point processes, Statistical modelling, Bayesian inference, Markov chain Monte Carlo, Star formation, Galaxies: individual: M33

SUMMARY FOR LAY AUDIENCE

The star formation process is crucial in understanding the formation of galaxies and the Universe. Stars are understood to form in an aggregated manner from their stellar nurseries — giant molecular clouds, leading to the formation of compact groups of stars called star clusters. Since the time scale of star formation surpasses human lifetime by orders of magnitude, studying the spatial distribution of giant molecular clouds, stars, and star clusters then serves as an indirect but sensitive probe for understanding the formation of these objects. While the spatial distribution of stars is relatively well-understood, this is not the case for giant molecular clouds or star clusters. In the current astronomy literature, the two-point correlation function is used for studying the spatial distribution of star clusters. However, it poses severe limitations and drawbacks when applied to studying the highly complex distribution of giant molecular clouds and star clusters. To address this issue, I adopt the framework of Gibbs point process models from spatial statistics and study its performance when applied to the point patterns of giant molecular clouds and young star clusters in the nearby M33 galaxy. Potential physical implications for the star formation process obtained from the models are also addressed.

CO-AUTHORSHIP STATEMENT

I hereby declare that I am the sole author of this thesis and that I have not used any sources other than those listed in the bibliography and identified as references. Dr. Ian McLeod and Dr. Pauline Barmby made their contribution only in terms of typo correction and commentary on wording.

Partial work from Chapters 4 and 5 is currently being formulated in a manuscript, titled "Untapped Power of Spatial Modelling in Astronomy: Gibbs Point Process Model for Objects in the Star Formation Complexes of M33" (Dayi Li, Pauline Barmby, A. Ian McLeod) and is being prepared for submission to the journal *Monthly Notices of the Royal Astronomical Society*.

All models are wrong, but some are useful.

George E.P. Box

To the Universe

ACKNOWLEDGEMENTS

I thank my parents and my sister for the love and support throughout the years, without whom I would not be able to pursue what I love. I thank my dearest friends, Wiseley, Fahim, Jovial and Zoya for their support and tolerance. I am indebted to them to help me grow and learn to become a better person. I thank Dr A. Ian McLeod, Dr Pauline Barmby for supervision and guidance. For their academic insights and intuition are the reasons I was able to finish this thesis. Thank Mr. Eric Koch, Dr Erik Rosolowsky for permission to use their data. Thank Mr. Baolai Ge for help on SHARCNET and SHARCNET for supporting the computation carried out in this work.

CONTENTS

LIST OF ABBREVIATIONS

BDMH	Birth-Death Metropolis-Hastings
CMB	Cosmic Microwave Background
CO	Carbon Monoxide
CSR	Complete Spatial Randomness
DMH	Double Metropolis-Hastings
GC	Globular Cluster
GMC	Giant Molecular Cloud
GNZ	Georgii-Nguyen-Zessin (formula)
GPP	Gibbs Point Process
GRF	Gaussian Random Field
HST	Hubble Space Telescope
LGCP	Log-Gaussian Cox Process
MCMC	Markov Chain Monte Carlo
MCMC-MLE	Markov Chain Monte Carlo Maximum Likelihood Estimation
MH	Metropolis-Hastings

MPLE	Maximum Pseudo-likelihood Estimation
MW	Milky Way
NND	Nearest-Neighbor Distance
OG	Open Cluster
PCF	Pair Correlation Function
SAVM	Single Auxiliary Variable Method
SC	Star Cluster
SPDE	Stochastic Partial Differential Equation
YSC	Young Star Cluster
YSCC	Young Stellar Cluster Candidate
2PCF	2-Point Correlation Function

LIST OF TABLES

LIST OF FIGURES

Chapter 1

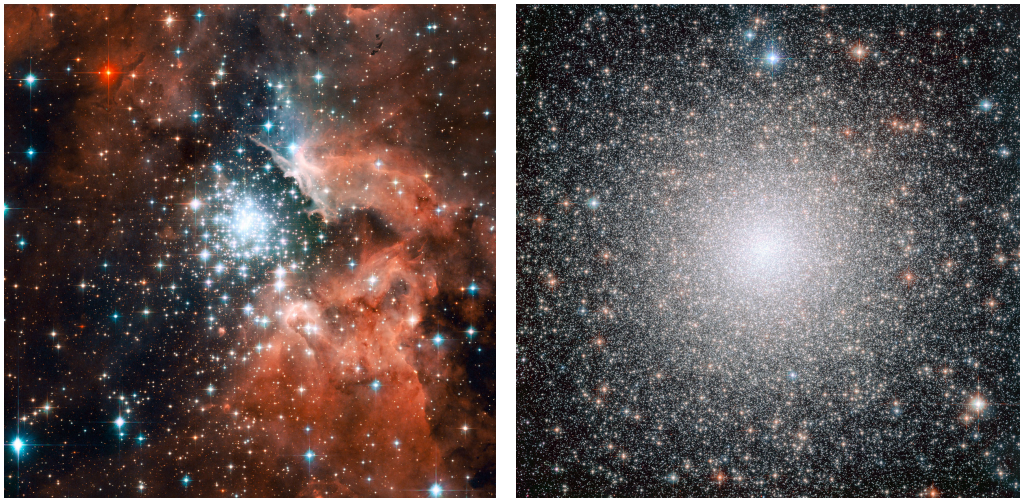
Introduction

1.1 Background: Star Clusters and Giant Molecular Clouds

The night sky has bewildered human beings since the dawn of our civilizations with galaxies being one of the most mysterious and majestic. A galaxy is a gravitationally bound system consisting of stars, gas and dust as well as dark matter. We reside in one of the many trillions of galaxies in the Universe–the Milky Way (MW)–an average barred spiral galaxy spanning approximately 150-200 thousand light years. However, the notion of galaxies was only conceived in the last century by Edwin Hubble, and there are still important questions about them that are directly linked to the formation of the Universe and our origin. Understanding the formation of galaxies and their constituents plays an important role in our understanding of the Universe.

As we sit inside of a galaxy, much information about our own galaxy is hidden from us. Studying nearby galaxies is then crucial to gaining insights on the formation of our own galaxy and the Universe. As the most luminous constituents of any galaxy are

stars, it is then natural to understand their formation processes. Current observations on star formation show that stars are formed in an aggregated manner (Lada and Lada, 2003; Portegies Zwart, McMillan, and Gieles, 2010), i.e., multiple stars are formed simultaneously in a relatively small and compact region that consequently forms a star cluster (SC). In general, SCs are mainly divided into two types, globular clusters (GCs) and open clusters (OCs). GCs are usually made up of around ten thousands to millions of old stars grouped into a roughly spherical region. OCs generally consist of only several hundred newly formed young stars that are not confined in a regular shape. Figure 1.1 below shows a super OC HD 97950 in the nebula NGC 3603 situated in the Carina spiral arm and NGC 6388, a GC in the constellation Scorpius.



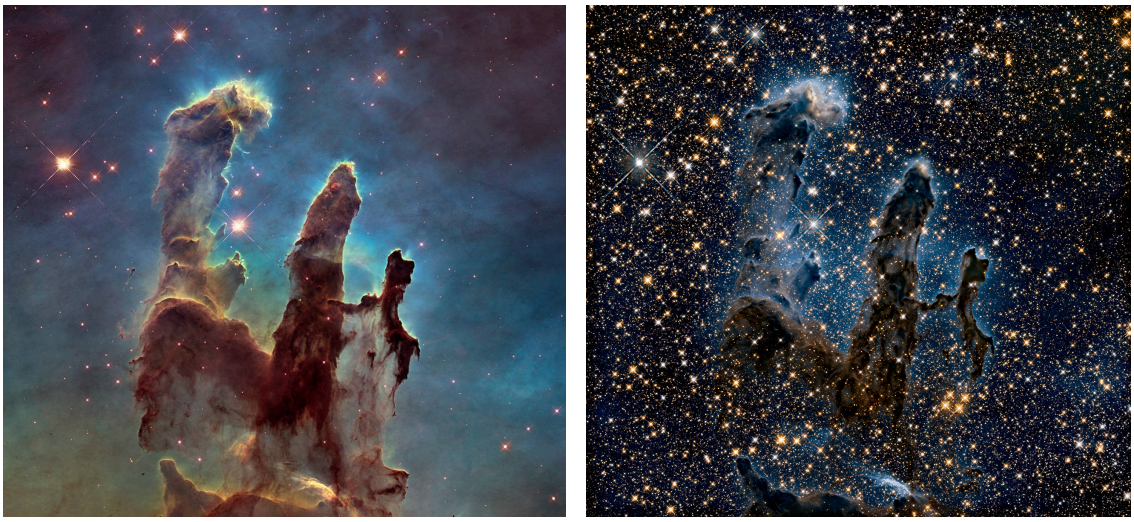
(a) HD 97950 in NGC 3603 (b) NGC 6388

Figure 1.1: Two types of star clusters observed in the Milky Way: (a) Open Cluster; (b) Globular Cluster.
Credits: (a) NASA, ESA and the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration; (b) NASA, ESA, F. Ferraro (University of Bologna)

SCs are found in every galaxy where we are able to observe and understanding their structure, distribution and evolution is a fundamental step to understand star formation as well as the formation and evolution of galaxies.

One key piece of current understanding of star formation is that stars form in

an aggregated manner due to their birth in giant molecular clouds (GMCs) (Lada and Lada, 2003; Portegies Zwart, McMillan, and Gieles, 2010). GMCs are massive collections of dense molecular gas with mass from 10^3 to 10^7 solar masses [1] in a tight region consisting of large amounts of raw material for star formation, mostly molecular hydrogen H_2. Figure 1.2 shows famous images taken by the Hubble Space Telescope (HST) of molecular clouds, the Pillars of Creation in the Eagle Nebula in the Milky Way Galaxy. Though emitting in the visible light spectrum, the stars are obscured by the gas and dust; observations in infrared reveal a significant amount of stars formed inside the gas towers.



(a) Visible Spectrum (b) Infrared Spectrum

Figure 1.2: The "Pillars of Creation" (a dense clump of gas and dust with star formation activity) in the Eagle Nebula: (a) in the visible spectrum, the gas and dust of the cloud is seen; (b) in the infrared spectrum, stars obscured by gas and dust can be observed.
Credits: NASA, ESA and the Hubble Heritage Team (STScI/AURA)

The distribution of star formation is understood to result from GMC fragmentation (Carlberg and Pudritz, 1990; McLaughlin and Pudritz, 1996), under the influence of gas collapsing under gravitational effects (Vega, Sánchez, and Combes, 1996; Kuznetsova, Hartmann, and Ballesteros-Paredes, 2018), turbulence in the local

[1] solar mass: $M_\odot = 1.989 \times 10^{30}$ kg.

environment (Elmegreen and Scalo, 2004; Federrath, Klessen, and Schmidt, 2009; Girichidis et al., 2012; Hopkins, Narayanan, and Murray, 2013; Guszejnov, Hopkins, and Krumholz, 2017) or feedback processes that suppress the star formation (Krumholz, 2014). Investigating the spatial distribution of SCs provides a sensitive and direct observational signature of the star formation process. However, it is not well understood to what extent the galactic environment, locally and globally, influences the evolution of SCs (Grasha et al., 2019). Understanding the spatial distribution and making quantitative measurements of it is then a crucial task. One current method used in understanding this distribution is called the two-point correlation function (2PCF) in astronomical literature (Peebles, 1980) or the pair correlation function (PCF) in spatial statistics literature. This tool was originally developed to measure how the distribution of galaxies behaves in order to investigate the large scale structure of the Universe. It measures the excess probability of finding two objects at a certain distance away to that of a completely random Poisson distribution of objects. Recently, it was used to investigate the distribution of SCs in multiple galaxies (Grasha et al., 2015; Grasha et al., 2017; Grasha et al., 2019). There are also studies (Grasha et al., 2019; Corbelli et al., 2017) done investigating the spatial relationship between SCs and giant molecular clouds (GMCs).

In this research, I propose a novel method through point process modelling to quantitatively measure multiple aspects of the spatial distribution properties of GMCs and SCs in the galaxy M33, such as inhomogeneity of GMCs and SCs as well as the correlation structure between GMCs and SCs. Since high resolution observations of GMCs in nearby galaxies are still relatively few, M33 became the sole target being investigated in this research due to the fact that it is the second closest (approximately 2.7 million light years away) spiral galaxy to us which is well-studied and for which high quality observational data are available. Figure 1.3 is the newest image of the M33 galaxy captured by the Hubble Space Telescope in the visible spectrum.

Figure 1.3: The largest mosaic image ever of the M33 (Triangulum) galaxy
Credits: NASA, ESA, and M. Durbin, J. Dalcanton, and B.F. Williams (University of Washington)

In the next section, I will address some of the current methodological issues when analyzing the spatial distribution of stellar objects in nearby galaxies. Subsequently, I will provide the motivation for point process modelling approach. I will also conduct a general review of point process modelling methodologies and issues regarding their existence, construction and inference procedures.

1.2 Literature Review

1.2.1 2-Point Correlation Function in Stellar Population Studies

The two-point correlation function (2PCF) was first derived by Peebles (1980) for trying to understand the large scale structure of the Universe. However, in spatial statistics literature, it took up the name pair correlation function (PCF) due to its origin in statistical mechanics for studying the distributional structure of molecules in complex systems. Nevertheless, they are exactly the same thing except that they

are sometimes normalized in different ways in astronomy and spatial statistics. In this research, I will use the terms interchangeably depending on the context.

2PCF is a simple yet powerful quantitative measure that tells us how certain point patterns behave compared to a Poisson (completely random) process at different scales. In spatial statistics, PCF is defined to be a non-negative function of the pairwise distance of two typical points in a point pattern. At any distance r, it measures the ratio of the probability that we observe a point at a distance r away from another point to that of a completely random distribution of points. Note that this ratio is in the sense of expectations as considering a single pair of points does not make any practical sense. Given necessary conditions, a PCF with value 1 at a certain distance r indicates that the point pattern analyzed has the same behavior as what is expected from a Poisson process at r. In astronomy, 2PCF is normalized by subtracting 1 from the PCF so that complete spatial randomness is denoted by 0.

Following Peebles (1980) and Peebles (2001), the 2PCF is defined as follows: let n denote the number density, i.e., average number of points per unit region, of a stationary point process (see Definition 2.2.3), then the probability of a point occurring in a typical volume element dV is given by

$$dP = ndV,$$

and the 2PCF is then related to the probability that there is a point occurring in each of the typical volume elements dV_1 and dV_2 with separation r_{12} through the following equation:

$$dP = n^2 dV_1 dV_2 [1 + \xi(r_{12})],$$

where $\xi(r_{12})$ is then the 2PCF. We can see that if a point process is Poisson, then $\xi(r_{12}) \equiv 0$ since the probability of observing a point occurring in each volume element is exactly $dP = n^2 dV_1 dV_2$ for a Poisson process. In fact, 2PCF can be considered

as the spatial counterpart of the autocorrelation function for time series, i.e., the time lag in the autocorrelation function for time series is now substituted by distance separation. 2PCF is important as it directly gives us the power spectrum of a point process (Peebles, 1980; Peebles, 2001) through the following Fourier transform:

$$P(k) = \int d^3 r \xi(r) e^{i\vec{k}\cdot\vec{r}}$$

where k is the frequency. Note that the power spectrum here characterizes the density contrast of matter as a function of scale. The direct relations between 2PCF and power spectrum is highly useful as the power spectrum is highly sensitive in detecting small fluctuations in the distribution of points (Blackman and Tukey, 1958).

However, as noted in both the spatial statistics literature (Baddeley, Rubak, and Turner, 2015; Møller and Waagepetersen, 2003) and astronomy literature (Peebles, 1980; Peebles, 2001), a crucial assumption on the validity of 2PCF is that the point pattern has to be stationary. This includes homogeneous and second-order stationary. Homogeneous in this case means that the number density n of a point process is constant everywhere. This can also be regarded as first-order stationary. This is apparent from the previous derivations that n is not a function of location or other environmental covariates. Second-order stationary means that the relationship between any two points does not depend on the absolute positions of the points but their relative positions or distance. If we make the further assumption that the point pattern is second-order stationary, the 2PCF then only depends on the distance between two points.

In a usual data analysis or modelling context, inhomogeneity has to be accounted for while second-order stationarity is generally assumed; methodology for analyzing point patterns with non-stationary second-order property is scarce as it has always been difficult to account for. In fact, this is an ongoing research topic in spatial

statistics (Risser, 2016).

For analyzing the large scale structure of the Universe, there has been accumulating evidence supporting the claim of stationarity (Peebles, 1993; Davis, Miller, and White, 1997; Peebles, 2001) of galaxy distributions on the scales of $10 \sim 200$ Mpc[2]. Therefore, the application of 2PCF in this context is justified and generally gives us accurate information about the spatial structure of galaxies.

Recently, 2PCF has been applied to analyze the spatial distribution of SCs by (Grasha et al., 2015; Grasha et al., 2017; Grasha et al., 2019; Corbelli et al., 2017) where the conclusion obtained from 2PCF suggests a power law clustering behavior between SCs. However, the use of 2PCF in these studies seem to have not met the crucial assumption of stationarity due to the apparent inhomogeneity in the number density of SCs across galaxy disk. In the case where it was considered, the inhomogeneity was not accounted for sufficiently. Indeed, it is quite obvious that the distribution of SCs in any galaxy would not be homogeneous due to the highly varied mass distribution in the galactic disk. Furthermore, local environmental effects such as the presence of GMCs will also produce inhomogeneity at local scales.

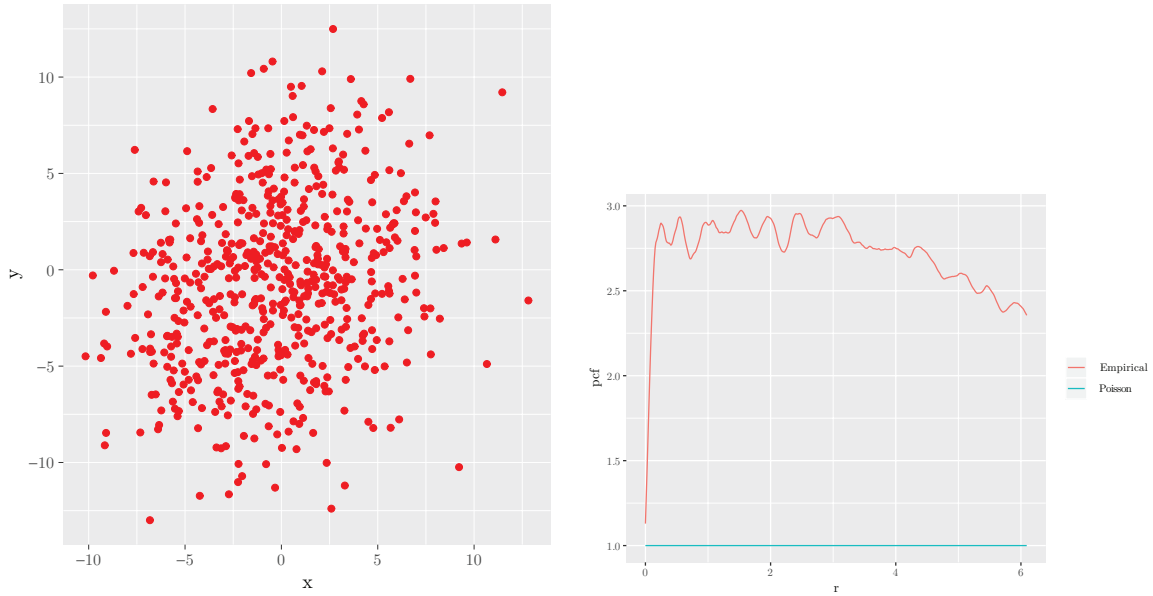
I will here provide reasons for the importance of accounting for inhomogeneity in the point pattern before directly using 2PCF/PCF. It is noteworthy that in the context of stellar population studies, the aim of 2PCF is to measure the interpoint interaction effect, i.e., whether the occurrence of a point is likely to be accompanied by another point at certain distance compared to a random distribution. This means that the violation of homogeneity can lead to drastically different conclusions from the fitted 2PCF. The reason is as follows: imagine that we have a point pattern where we know there are environmental effects exerting influence on the number density of the points in different regions. Then it is likely that a region with high number

[2]1 pc (parsec) $\approx$ 3.26 light-year(3.086×10^{16} m). Solar system is on the scale of $\ll$ 1 pc; star formation complex is on the scale of $\sim$ 100 pc; galaxies are on the scale of 1$\sim$100 kpc in diameter and generally on the scale of 1$\sim$10 Mpc apart from each other.

density will appear to be more clustered than a Poisson process. Consequently, fitting a 2PCF directly to the point pattern will always lead to the conclusion that the point pattern is clustered compared to a Poisson process. As noted, the clustering conclusion obtained from 2PCF here is a form of second order clustering resulting from the interpoint interaction. However, it is completely possible that there exists no interpoint interaction between points due to inhomogeneity of the number density. Below is a simple demonstration of how inhomogeneity can completely derail the conclusion of a directly fitted PCF.

We simulate an inhomogeneous Poisson process by mimicking the distribution of SCs in a galaxy where we assume the number density n is the strongest at the origin, and falls off according to some exponential power law as we move away from the origin. Figure 1.4(a) shows one simulation of said inhomogeneous Poisson process and Figure 1.4(b) shows its corresponding empirical PCF, where r is the pairwise distance between two points. The PCF indicates that the point pattern is clustered at all scales since the PCF is greater than 1 at all scales. However, the process is in fact a Poisson process and the actual PCF should be approximately 1 at all scales. The reason for the drastic difference is precisely the inhomogeneity and this example perfectly demonstrates how far away from the truth our conclusions will be when fitting a PCF without properly accounting for inhomogeneity.

From the previous arguments, therefore, it is important to differentiate the subtle difference between the effects from inhomogeneity and the interpoint interaction. Inhomogeneity exerts its influence on the occurrence of a point (the number density) independently of another point. We can think of this as a "fertility" effect, i.e., how much resource there is in a certain region to produce one point. The interpoint interaction, however, is the influence exerted from the occurrence of a point to another point, i.e., there exists a notion of dependence structure. We can think of this as competition for resources in the case of repulsion and triggering of occurrences of multiple



(a) Simulated inhomogeneous Poisson pattern

(b) Fitted Empirical PCF

Figure 1.4: Simulated Inhomogeneous Poisson process with its corresponding directly fitted PCF

points in the case of clustering. Subsequently, when analyzing the distribution of SCs, we need to distinguish between these two effects, and the conclusions from these two effects consequently answer different questions. Studying the inhomogeneity of the SCs answers questions such as "how do SCs distribute around GMCs?" while the 2PCF answers questions such as "how do SCs distribute among themselves?" while all other factors are assumed to be accounted for. It is straightforward to see that if the conclusion is SCs are more likely to occur closer to GMCs compared to a random Poisson distribution, this does not imply that the occurrence of a SC is likely to trigger the occurrence of another SC compared to a random distribution. However, there is a pitfall in separating the inhomogeneity effect and interpoint interaction in that if there is only one realization of some random point process, it is impossible to tell whether a clustering/repulsive feature is due to inhomogeneity of the point process or interpoint interaction. Therefore, separating these two effects should be dictated by theories and necessary assumptions. Nevertheless, excluding the effect of

inhomogeneity can lead to drastic differences in conclusions from a fitted 2PCF.

Grasha et al. (2015), Grasha et al. (2019), and Grasha et al. (2017) concluded that the 2PCF follows a power law as a function of distance between SCs and shows a clustering feature at short distances compared to a random distribution of points. It is likely that the 2PCF indeed follows a power law and exhibits clustering features at certain distances. However, the fitted parameters are not likely reflecting the true degree of the relationship for the reasons mentioned above.

It is certainly tempting to find a way to account for the inhomogeneity and then fit a 2PCF. But there is a big problem if 2PCF is the only tool we have. For almost all point patterns, there is no numerical measurement of inhomogeneous effects so that we can eliminate them and refit the 2PCF. For example, there is no way for us to know the accurate numerical measurement on the effects of galactic structure on the distribution of SCs without other additional tools. This is one of the most fundamental limitation of 2PCF since for a vast amount of real world point pattern data, there is always some degree of inhomogeneity.

Even though there are attempts (Grasha et al., 2019; Corbelli et al., 2017) made to address the issue coming from galactic structure, the method used is rather ad-hoc and prone to information loss. Corbelli et al. (2017) accounted for the large scale galactic structure effect in their study of the relationship between GMCs and young SCs (YSCs) in M33. However, due to the limitations of 2PCF, they choose to separate the galactic plane into three radial regions encompassing the galactic center so that the large scale galactic structure effect could be regarded as homogeneous in each region. Similarly, Grasha et al. (2019) divided the galactic plane of M51 into two regions encompassing the galactic center and conducted 2PCF analysis for SCs. Though this method could potentially eliminate the galactic structure effect, it still cannot account for local effects. These can be due to local inhomogeneity of interstellar medium such as gas and dust that fuel the star formation process.

Since uneven distribution of the raw materials for star formation will lead to the inhomogeneous distribution of stars, especially for the younger generation of stars. Furthermore, the determination of region boundaries is arbitrarily defined by users and how changes of the boundaries would affect the resulting 2PCF is not exactly clear. Grouping the data also introduces information loss since information on a continuous space is cut into several non-communicating subspaces.

Another limitation of 2PCF is its restriction on investigating how the properties of SCs and GMCs affect their spatial relationships. Grasha et al. (2015), Grasha et al. (2019), and Grasha et al. (2017) investigated the effect of age on the clustering strength of SCs. The data has to be grouped by age to provide an analysis from the 2PCF. This grouping of data loses a significant amount of information since a continuous variable is reduced to a categorical variable.

It is imperative to ask for a new method to address the issues of current methodologies of using 2PCF. An immediate candidate is parametric modelling where a point process model with physically meaningful parameters is constructed. In the next subsection, I will present a review on the state of the art of point process modelling methodologies and their related issues.

1.2.2 Spatial Point Process Modelling

Spatial point process modelling takes up a significant part of spatial statistical research. It concerns the study of the locations of the occurrence of random objects or events. The most important question of interest in understanding spatial point processes is the behavior of point patterns compared to a Poisson process. There are two types of characteristics of interest that can subsequently provide information on related scientific questions — repulsive (regular/inhibitive) and clustering (attraction/aggregated) behaviors. Two types of point process modelling paradigms are widely considered — Cox point processes (Cox, 1955) and Gibbs/Markov point

processes (GPP) (Ripley and Kelly, 1977; Lieshout, 2000).

Cox point processes are generally employed for studying clustering point patterns. Cox point processes are also termed doubly stochastic Poisson processes. The idea is that there exists a latent high-level Poisson process where its realization gives rise to the observed low-level point process. It is then assumed that the observed process is a realization of Poisson process centered around the high-level points. The construction of Cox point processes naturally lends them the ability to model clustering point patterns where the point pattern is essentially treated as a realization of an inhomogeneous Poisson process. For example, Cox processes were used to simulate spike train data (Krumin and Shoham, 2009) as well as in financial mathematics where credit risk is a significant factors when pricing financial instruments (Lando, 1998).

Much more recently, an extension based on Cox processes was proposed by Møller (Møller, Syversveen, and Waagepetersen, 1998), called log-Gaussian Cox processes (LGCP). LGCP have provided a wide range of applicability in modelling real world clustering spatial/spatio-temporal data, and have been used in vastly different fields including ecology (Brix and Moller, 2001; Serra et al., 2014; Waagepetersen et al., 2016), pattern recognition (Nguyen, Fablet, and Boucher, 2011), epidemiology (Li et al., 2012), criminology (Rodrigues and Diggle, 2012; Shirota and Gelfand, 2017), neuroscience (Samartsidis et al., 2019), etc. The difference between LGCP and standard Cox processes is that instead of assuming the inhomogeneity is a result of realization of a latent Poisson process, LGCP assumes that the latent process is a Gaussian random field (GRF) which can be a function of location and observed or unobserved covariates. The realization of the exponentiated GRF then produces a continuous intensity (number density) surface which gives rise to the observed point pattern. The observed point pattern is assumed to be a realization of an inhomogeneous Poisson process with inhomogeneity specified by the previous intensity surface. The main focuses of modelling of LGCP are the covariate effects and the covariance structure

of the latent GRF as these two aspects govern the behavior of the low level inhomogeneous Poisson process.

Currently, the main research focus on LGCP is inference computation. Due to its inherent hierarchical structure, it is naturally modelled through a Bayesian hierarchical framework. However, the computational bottleneck manifests from the inference for the continuous latent GRF. For any continuous GRF, the inference computation requires the following: (i) Construct a fine lattice grid over the observation window (ii) Count the number of points in each region to approximate the number density in each region (iii) Fit the GRF based on the approximation. Needless to say, the finer the lattice grid, the better the approximation and more accurate the result. However, increasing the number of grid points causes a computational bottleneck since fitting the GRF requires inverting a dense covariance matrix with its dimension equal to the number of grid points. The most widely employed method for dealing with this is to model the latent GRF as a conditional autoregressive model which renders fast computation possible (Rue and Held, 2005). However, it is suggested that computation on a fine lattice is highly wasteful since the lattice method cannot be locally refined (Simpson et al., 2016). Recent attempts by (Lindgren, Rue, and Lindström, 2011) suggested a link between stochastic partial differential equations (SPDE) and a continuous random field. This leads to the result that a continuous random field can be effectively approximated by a Gaussian Markov random field where existing methods such as the integrated nested Laplace approximation (INLA) can be used for fast Bayesian computation. However, the modelling techniques and computational structure of the SPDE approach are still in the early stages of development. Therefore, for problems of highly complex structure such as ones in this research, the SPDE approach is not yet suitable.

The main reason preventing me from using LGCP in this research is problem specific: in terms of modelling the distribution of stellar populations, it is better to use

models more suitable for modelling physical objects. Furthermore, it is physically more reasonable to impose the notion of interpoint interaction in the model as there exist direct physical counterparts of the notion of interaction between stellar objects, e.g., gravitational interaction, competition for the fuel of star formation, feedback process suppression and so forth. LGCP does not seem suitable as it assumes the observed pattern to be a realization of an inhomogeneous Poisson process which necessarily strips away the concept of influence between points. Instead, the dependence structure is fully captured by the covariance structure of the latent GRF. However, how to interpret the physical implication of the fitted covariance structure is unclear since there does not yet exist an immediate physical counterpart of the latent GRF which supposedly gives rise to the physical processes governing star formation. Interestingly, there does exist physical manifestation of GRF in cosmology — the cosmic microwave background (CMB) (Wandelt, 2013). However, CMB is a continuous field of remnant electromagnetic radiation shortly after the Big Bang rather than a point pattern. Nevertheless, treating the generation process of stellar populations as a GRF can potentially open interesting ideas for future research in astrophysics. However, as the focus of this research is not geared towards astrophysical theory, I do not pursue the LGCP approach here.

On the other hand, the GPP model is a class of point processes emphasizing interaction between the points. It can be used to model interaction of clustering, repulsive, or both types of interaction. Compared to the Cox process, the GPP model is a much better route to go in the context of this research. The statistical structure of interpoint interaction is already established in astronomy through the development and application of 2PCF. GPP also has a very close tie to the 2PCF. Furthermore, it is very easy to incorporate physically interpretable parameters in a GPP model and it is generally suited for modelling point patterns of physical objects.

GPP are ubiquitous for modelling repulsive point patterns due to the model's

emphasis on the notion of interaction between points. This is in general not possible for the Cox processes. Models of clustering patterns using GPP are also available. Originating from statistical physics, these models were first employed to study the behavior of physical systems with complex dependent structure. One of the first models of this type is the famous Boltzmann distribution (Gibbs distribution) (Gibbs, 1902) where the aim of modelling is the probability of observing a system being in a certain state as a function of the energy and temperature of the system. Subsequently, there was the Ising model (Ising, 1925) for studying the magnetic dipole moments of atomic spins.

GPP were only much later introduced to the spatial statistics community by Ripley and Kelly (1977), sparking application for point process modelling under the name of Markov point processes. Because of the ease of construction and the ability to incorporate physically meaningful parameters, there is a countless number of possible model constructions that can be tailored for different problems. For modelling repulsive patterns, there exists hard-core process, Strauss process, soft-core process, etc. For modelling clustering process, one can employ Geyer's saturation process and triplet process as well as area-interaction process. There also exist models that can account for spatially varying behavior, e.g., repulsion-attraction processes where point pattern exhibits a repulsive pattern at short range and a clustering pattern at mid to long range. Due to its high flexibility, GPP has seen wide applications in forestry (Goulard, Särkkä, and Grabarnik, 1996; Picard et al., 2009), ecology (Isham, 1984; Högmander and Särkkä, 1999; Rajala, Murrell, and Olhede, 2018), and neuroscience (Johnson, 1996), as well as cosmology (Tempel et al., 2016).

With the ability to parametrically model the spatial distribution of SCs, we can model the inhomogeneous effects with a flexible structure based on empirical observations and existing physical theory. Furthermore, it gives us the ability to simultaneously model the adjusted 2PCF as well as the effect of the properties of SCs and

GMCs on their distributions.

In this research, I propose two novel GPP models tailored for modelling stellar populations specifically for GMCs and SCs. The models are constructed in order to capture the empirical distributional structure exhibited by GMCs and SCs. I will also attempt to derive new physical insights from the inferred model parameters.

Although GPP have highly appealing properties in terms of interpretability and model flexibility, similar to LGCP, they too pose challenges when it comes to model inference. This is due to the fact that the likelihood function of GPP models are partially intractable, in that there exists an intractable normalizing constant which is a function of the model parameters. Due to this complication, a significant amount of literature has focused on developing inference algorithms in the maximum likelihood paradigm for the sake of computational speed. Two main inference methods are the maximum pseudo-likelihood estimation (Baddeley and Turner, 2000, MPLE) and Monte Carlo maximum likelihood estimation (Geyer and Møller, 1994, MCMCMLE). However, these methods can be limited and restrictive.

MPLE uses a local Markov-type approximation of the true likelihood to carry out MLE inference (Baddeley and Turner, 2000), hence the name pseudo-likelihood. Due to the fact that it only employs local information, MPLE usually underestimates the strength of interaction. This means that if a point pattern exhibits strong interpoint interaction, results obtained from MPLE will tend to be highly biased. Moreover, it requires the model to be in log-linear form Baddeley and Turner (2000). This is restrictive since model parameters with physical meaning, such as the typical scale of the interaction range, cannot be inferred through this approach as they are not of log-linear form with sufficient statistics. Parameters that are in log-linear form with the sufficient statistics are usually difficult to derive physical interpretation from since for various existing models, e.g., the Strauss process, the function characterizing the interpoint interaction is not a continuous function of the distance. In a physical

context, discontinuity in functions is generally not desired.

MCMCMLE has the restriction that the derivative of the log-likelihood with respect to model parameters must have analytical gradients which might not be possible for certain models, e.g., Goldstein et al. (2014). Furthermore, for models with complex likelihood functions, the computation of gradients of the log-likelihood can be costly and sometimes the complexity of the analytical gradient can be such that it causes computational overflow/underflow. Furthermore, in the context of this research, it is more suitable that the model parameters follow certain probability distributions rather than being a fixed value. This then naturally leads to the Bayesian inference paradigm.

For Bayesian inference of GPP, unlike the MLE methods mentioned above, there is no restriction on the form of the likelihood, hence it is much more suitable for modelling physical systems. Furthermore, a Bayesian paradigm is a natural approach for problems in astronomy since new observational data will become available with the employment of more powerful telescopes. Bayesian inference is also much easier to implement compared to both MPLE and MCMCMLE. However, just like the problem faced by MLE approaches, it is also hindered by the intractable normalizing function as mentioned before. This renders the posterior distribution to be doubly-intractable (since there is an intractable normalizing term in the likelihood and an intractable term for the posterior distribution) and standard Markov chain Monte Carlo (MCMC) algorithms cannot be used for Bayesian inference.

Methods to facilitate MCMC algorithms for GPP only appeared recently due to the explosive increase in computational power in recent years. The first attempt at dealing with this issue is the ingenious auxiliary variable/exchange algorithm by Møller et al. (2006) and Murray, Ghahramani, and MacKay (2006) where they proposed to work around the intractable normalizing constant by simulating an auxiliary variable at each Metropolis-Hasting iteration. This auxiliary variable will then causes

the unknown ratio of the normalizing constants to be canceled and standard MCMC can proceed. However, the method is highly restrictive since it requires one to perfectly simulate the auxiliary variable which is usually not possible except for some toy examples. A much more practical algorithm was later proposed by Liang (2010), called the double Metropolis-Hasting algorithm (DMH). DMH relaxes the requirement of perfect simulation of the auxiliary variable by replacing it with simulation from a standard MCMC run. This made Bayesian inference much more practical for real world complex GPP. The DMH algorithm will be employed for inference purposes in this research.

Next, I formalize the scientific problems of interest and necessary assumptions which will dictate the construction of the models.

1.3 Problems and Assumptions

- **How to model the highly inhomogeneous distribution of stellar populations?**

  To model the inhomogeneity of stellar distributions, I assume the following: The inhomogeneity of GMCs are attributed to the CO filament. Numerous observations indicate that CO molecules generally forms in filamentary structure, with GMCs born and gradually separated from the filament and eventually dispersed. These observations suggest that GMCs positions are strongly affected by CO filament. Since GMCs generally disperse in a very short time frame ($\sim$ 60 million yrs), there is hardly enough time for them to diffuse away from CO filament and appear uncorrelated with the filament structure. Hence the assumption is reasonable.

  For the inhomogeneity of SCs, I assume two forms of inhomogeneity, a global

trend and a local effect. I assume that the global trend is attributed to the general mass distribution within a galaxy, i.e., the mass density is higher in the inner region and lower in the outer region. This mass density profile is usually modelled as some form of power law as a function of the distance to the galactic center. For the local effect, I assume that it results from the presence of GMCs, as it is mentioned that GMCs are the widely accepted "stellar nurseries". I will also assume that the effect of each GMCs on the "fertility" of SCs has a finite effective range, i.e., after a certain typical range, the effect of GMCs on SCs will become negligible.

- **What is the correlation between GMCs and SCs?** The correlation structure between GMCs and SCs is modeled as an asymmetrical hierarchical relationship since there exists different levels of hierarchy between the underlying processes that generate GMCs and SCs. I will assume that the process generating GMCs takes the higher level of hierarchy than the process for SCs. This is because GMCs are considered as the birthplace of SCs as mentioned. Therefore, a natural formation hierarchy exists between GMCs and SCs and the model needs to take this formation structure into account. The correlation among GMCs and SCs will then arise from this hierarchical structure.

- **How do the properties of GMCs affect the distribution of SCs around them?** To infer the effect of properties of GMCs on the distribution of SCs, I will assume that the effect of GMCs on SCs is of some generic functional form. A simple example is to model the effect as a linear combination of the properties of GMCs. I will discuss this in detail in Chapter 5.

- **What is the second order behavior of SCs' spatial distribution after accounting for inhomogeneous intensity?**

  The second order behavior is rather difficult to infer from summary statistics.

I will assume that there exists short range repulsion between SCs as it complies with physical reality since SCs have physical sizes. Several other physical observations/evidence suggesting short range repulsion behavior between SCs. For mid to long range interaction, I will assume a Poisson structure. If any deviance exists, it could be detected through model criticism and it can provide us with important physical implications.

This thesis is organized in the following way. Chapter 2 introduces the necessary definitions and theories on point processes. Chapter 3 introduces the formalism regarding the meaningful construction of GPP models as well as the details of simulation and Bayesian inference algorithms for GPP models. Chapter 4 provides the details on how I construct the models. Chapter 5 consists of data analysis on stellar objects in M33 using the proposed models as well as physical implications derived from models. Chapter 6 provides conclusions.

Chapter 2

Spatial Point Process

Spatial point process models are important tools employed by various scientific disciplines, such as ecology, epidemiology, criminology, seismology, etc., to study how the locations of a collection of random events or random objects distribute in space. For example, ecologists may be interested how certain species of trees distribute on a forest floor so they may get insights on forest management. Criminologists would like to know where burglary might occur in Toronto and provide knowledge to the police for more efficient theft prevention.

In astronomy, spatial point processes are essentially everywhere. Looking up at the night sky and there is an extremely complex and beautiful point pattern that consists of something that eluded humans for centuries — the stars. However, spatial point process modelling hasn't found its way in astronomy most likely due to the in depth knowledge and training required in theoretical statistics which most astronomers are not well-acquainted. Furthermore, literature on spatial statistics can be terse and dull towards non-practitioners.

This chapter is then dedicated as a basic introduction of point process and provides

some technical details on the modelling tools that will be used for our problems. Since this thesis is not pivoted towards the theory and methodology of point process, which requires extensive measure theory probability, I gear the approaches toward more applied lenses.

2.1 Point Process

Before defining a point process, a space S is required for the points to live in. Usually, $S \subset \mathbb{R}^d$ and typically a d-dimensional box or sphere in R^d. For our problems, since M33 is extremely far away, only a two dimensional projection of the spatial positions of the objects is available. Therefore, I will focus on $S \subset \mathbb{R}^2$ and model the distribution of objects in M33 as a planar process. In most cases, however, the points observed will be bounded by an observation window W, e.g., one can only observe the patch of sky that can be captured by the field of the camera of a telescope.

Now a point process $\mathbf{X}$, with realization or configuration $\mathbf{x}$, is defined as a finite and countable process. Note that we use $\mathbf{x}$, $\mathbf{y}$, ... to denote a set of points which is a realization/configuration of $\mathbf{X}$ and use x, y, ξ, η, ... to denote a singleton.

Below, I present several important definitions and theorems regarding point process that will provide us with basic groundwork for constructing our model. Note that for all definitions and theorems in this chapter, I follow the discussion from Møller and Waagepetersen (2003) and Baddeley, Rubak, and Turner (2015).

Definition 2.1.1. (Møller and Waagepetersen, 2003, pp. 7) *Let $n(\mathbf{x})$ be the cardinality of a subset $\mathbf{x} \subset S$. Let $\mathbf{x}_A = \mathbf{x} \cap A$ for $A \subset S$. $\mathbf{X}$ is called locally finite if $n(\mathbf{x}_A) < \infty$ for all bounded A.*

This means that for any realization $\mathbf{x}$ of $\mathbf{X}$, there are only a finite number of points

in any bounded region. This provides us the support of $\mathbf{X}$. We denote this as

$$N_{lf} = \{\mathbf{x} \subset S : n(\mathbf{x}_A) < \infty,\ \forall A \subseteq S \text{ and } |A| < \infty\}$$

Here N_{lf} denotes the set of all *locally-finite point realization.* Note that $\mathbf{X}$ is not a random variable since it does not take any numeric values. Rather, the "value" that $\mathbf{X}$ can take is any point pattern realization $\mathbf{x}$ that satisfies definition 2.1.1 or simply $\mathbf{x} \in N_{lf}$ and a point process model specifies a probability density function that gives us the likelihood of observing the realization $\mathbf{x}$ of $\mathbf{X}$ given a certain specification of the model structure.

The precise definition of a point process requires an extensive amount of measure theory. Since this research should also be easily accessible for astronomers, I will avoid giving a mathematically precise definition of point process. Assume S has a defined metric (usually the Euclidean distance), it is sufficient to know that once a point process is defined on S, it is equipped with a probability distribution

$$P_{\mathbf{X}}(F) = P(\mathbf{X} \in F).$$

Here $F \subset N_{lf}$ is a collection of different point pattern realizations. This distribution gives us the probability of $\mathbf{X}$ producing a realization $\mathbf{x} \in F$. An analogy from random variable is that the probability distribution function P_X of a random variable X gives us the probability $P(X \in B)$ where B is well-defined and $B \subset R$.

Definition 2.1.2. *A point process on S is simple if no two points from its realizations are at the same location.*

Point process modelling does not address point patterns with coincidental points and most real world point processes, such as the ones we are addressing in this research, are simple.

2.2 Poisson Point Process

2.2.1 Definition and Basic Properties

The simplest point process model is the Poisson point process. It represents the idea of complete spatial randomness (CSR). Another way to look at it is that the location of points are completely independent from each other, i.e., there is no interaction between any point in a Poisson point process. This does not sound very interesting since CSR almost never exists in real life point pattern data. However, Poisson point process serves as an anchor point and a reference model for point pattern analysis and it is the most fundamental building block of more sophisticated point process models.

To define any point process model, we need the concept of an intensity function and intensity measure. An intensity function satisfies $\lambda(\geq 0) : S \to [0, \infty)$ and $\int_A \lambda(\xi) d\xi < \infty$ for all bounded $A \subset S$. As the name suggests, this function specifies the rate of point occurrence at location $\xi \in A$, i.e., it gives us the expected number of points in an infinitesimal neighborhood of ξ. An intensity measure on A of a point process with intensity function λ is defined as $\mu(A) = \int_A \lambda(\xi) d\xi$.

Before going into Poisson process, we have to introduce another point process that is closely related to Poisson process.

Definition 2.2.1. (Møller and Waagepetersen, 2003, pp. 14) *Let f be a probability density function on a set $A \subset S$, and let $n \in \mathbf{N}^+$. A point process $\mathbf{X}$ consisting of n i.i.d. points with density f is called a binomial point process of n points in A with density f. We denote $\mathbf{X} \sim$ binomial(A, n, f).*

This definition gives us the following important property of a binomial point

process. For any $B \subset A$, let $p_B = \int_B f(\xi)d\xi \in [0,1]$.

$$P\left(n(\mathbf{X}_B) = k\right) = \binom{n}{k} p_B^k (1-p_B)^{n-k}. \tag{2.1}$$

This means the number of points $n(\mathbf{X}_B)$ in any sub-region B of A is a binomial random variable with parameter n and p_B hence the name binomial point process. Furthermore, the occurrence of one point has no effect on the occurrence of another point which is crucially linked to the Poisson point process. One important special case of binomial point process would be the uniform point process on B, Uniform(B), conditioned on there being n i.i.d. points. This is the case where $f(\xi) = 1/|A|$.

Definition 2.2.2. (Møller and Waagepetersen, 2003, pp. 14)*A point process* $\mathbf{X}$ *on* S *is a Poisson point process with intensity function* λ *if :*

(a) *for any* $A \subset S$ *and* $0 < \mu(A) < \infty$, $n(\mathbf{X}_A) \sim$ *Poisson*$(\mu(A))$ *;*

(b) *for any* $n \in \mathbf{N}^+$ *and* $A \subset S$ *with* $0 < \mu(A) < \infty$, *conditional on* $n(\mathbf{X}_A) = n$, $\mathbf{X}_A \sim$ *binomial* (A, n, f) *with* $f(\xi) = \lambda(\xi)/\mu(A)$.

We write $\mathbf{X} \sim$ *Poisson*(S, λ).

Following Definition 2.2.2, if λ is constant, then Poisson(S, λ) is called a homogeneous Poisson process on S with rate or intensity λ; otherwise, it is called an inhomogeneous Poisson process on S. If $\lambda(\xi) \equiv 1$, then the process is called a unit rate Poisson process.

Definition 2.2.3. (Møller and Waagepetersen, 2003, pp. 14)*A point process* $\mathbf{X}$ *on* R^2 *is stationary if its distribution is translation-invariant, i.e.,* $P_{\mathbf{X}} = P_{\mathbf{X}+s}$ *for any* $s \in R^2$. *It is isotropic if its distribution is rotation-invariant about the origin, i.e.,* $P_{\mathbf{X}} = P_{\mathcal{R}\mathbf{X}}$ *for any rotation* $\mathcal{R}$ *around the origin.*

Stationarity is an important consideration as it is crucial to the formulation of vast number of point process models.

2.2.2 Distribution Function for Poisson Process

Proposition 2.2.1. *(Møller and Waagepetersen, 2003, pp. 15)* $\mathbf{X} \sim Poisson(S, \lambda)$ *iff for all* $A \subset S$ *with* $\mu(A) = \int_A \lambda(\xi)d\xi$ *and all* $F \subset N_{lf}$,

$$P(\mathbf{X}_A \in F) = \sum_{n=0}^{\infty} \frac{\exp(-\mu(A))}{n!} \int_A \cdots \int_A \mathbf{1}[\{x_1, \ldots, x_n\} \in F] \prod_{i=1}^{n} \lambda(x_i) dx_1 \ldots dx_n \tag{2.2}$$

where the integral for $n = 0$ *is read as* $\mathbf{1}[\emptyset \in F]$.

To see that $X \sim \text{Poisson}(S, \lambda)$ implies equation 2.2, we have

$$\begin{aligned} P(\mathbf{X}_A \in F) &= \sum_{n=0}^{\infty} P(n(\mathbf{X}_A) = n) P(\mathbf{X}_A \in F \mid n(\mathbf{X}_A) = n) \\ &= \sum_{n=0}^{\infty} P(n(\mathbf{X}_A) = n) \int_A \cdots \int_A P(\{x_1, \ldots, x_n\} \in F \mid \mathbf{X}_A = \{x_1, \ldots, x_n\}) \times \\ &\quad f(\mathbf{X}_A = \{x_1, \ldots, x_n\} \mid n(\mathbf{X}_A) = n) dx_1 \ldots dx_n \\ &= \sum_{n=0}^{\infty} \frac{\exp(-\mu(A))}{n!} \int_A \cdots \int_A \mathbf{1}[\{x_1, \ldots, x_n\} \in F] \prod_{i=1}^{n} \lambda(x_i) dx_1 \ldots dx_n. \end{aligned} \tag{2.3}$$

This gives us the distribution function of $X \sim \text{Poisson}(S, \lambda)$.

Even though a Poisson process has a distribution function, it does not have a well-defined density on its own. However, a density with respect to another Poisson process exists. Usually, this is the unit rate Poisson point process.

2.2.3 Densities for Poisson Process

To introduce a density for a Poisson process, we need the idea of *absolutely continuous* (Møller and Waagepetersen, 2003, pp. 25) between point processes. If $\mathbf{X}$ and $\mathbf{Y}$ are two point processes defined on S, then $\mathbf{X}$ is absolutely continuous with respect to $\mathbf{Y}$ iff $P(\mathbf{Y} \in F) = 0$ implies $P(\mathbf{X} \in F) = 0$ for any $F \subset N_{lf}$. Furthermore, there exists

a function $f : N_{lf} \to [0, \infty]$ such that

$$P(\mathbf{X} \in F) = \mathbb{E}[\mathbf{1}[\mathbf{Y} \in F] f(\mathbf{Y})], \ \forall F \subset N_{lf}. \tag{2.4}$$

Here f is defined as the *density* for $\mathbf{X}$ w.r.t. $\mathbf{Y}$.

Proposition 2.2.2. *(Møller and Waagepetersen, 2003, pp. 25) Suppose two Poisson point processes, $\mathbf{X}$ and $\mathbf{Y}$, have intensity $\lambda_{\mathbf{X}}, \lambda_{\mathbf{Y}} : S \to [0, \infty)$ so that $\mu_{\mathbf{X}}(S), \mu_{\mathbf{Y}}(S) < \infty$, and that $\lambda_{\mathbf{Y}}(\xi) > 0$ implies $\lambda_{\mathbf{X}}(\xi) > 0$. Then $\mathbf{X}$ is absolutely continuous w.r.t. $\mathbf{Y}$, with density*

$$f(\mathbf{x}) = \exp(\mu_{\mathbf{Y}}(S) - \mu_{\mathbf{X}}(S)) \prod_{\xi \in \mathbf{x}} \lambda_{\mathbf{X}}(\xi)/\lambda_{\mathbf{Y}}(\xi). \tag{2.5}$$

for any finite point realizations $\mathbf{x} \subset S$.

Note that if $\mathbf{X} \sim \text{Poisson}(S, \lambda)$ and $\mathbf{Y}$ a unit rate homogeneous Poisson process, we then have the density of $\mathbf{X}$ as

$$f(\mathbf{x}) = \exp\left(|S| - \int_S \lambda(\xi) d\xi\right) \prod_{i=1}^{n(\mathbf{x})} \lambda(x_i). \tag{2.6}$$

Furthermore, if $\mathbf{X}$ is homogeneous, then

$$f(\mathbf{x}) = \exp((1 - \lambda)|S|)\lambda^{n(\mathbf{x})}. \tag{2.7}$$

We will see later, in terms of Gibbs point process models, a Poisson process is the only model with a probability density that can be expressed fully with a tractable normalizing constant given that the intensity λ is specified.

2.3 Intensity Measures

Up to this point, we have a formal understanding of point processes. Just like ordinary random variables, there are crucial intensity measures that resembles the moments of a random variable. These are fundamental in helping us understand the structure in point pattern data. In this section, I provide some theoretical basis for the intensity measures of spatial point patterns.

2.3.1 First Order Intensity Measure

The first order intensity measure for a point process is analogous to the expectation for a random variable. It is given by

$$\mu(A) = \mathbb{E}[n(\mathbf{X}_A)],\ A \subset \mathbb{R}^2, \tag{2.8}$$

i.e., it measures the expected number of points in a certain region, A, for a point process $\mathbf{X}$. As in the previous section, the intensity function λ is defined as a function from S to $[0, \infty)$ such that

$$\mu(A) = \int_A \lambda(\xi) d\xi,\ A \subset \mathbb{R}^2. \tag{2.9}$$

Heuristically, $\lambda(\xi)d\xi$ is the probability for the point process $\mathbf{X}$ to have a point occurring in an infinitesimally small ball centered at ξ with volume $d\xi$. If $\mathbf{X}$ is homogeneous, then λ is constant, and it represents the expected number of points per unit volume. In a modelling context, we will usually model the intensity function λ to account for potential inhomogeneity.

2.3.2 Second Order Intensity Measure

For a second order intensity measure, it measures the behavior of pair of points. We first consider the second order factorial moment measure $\alpha^{(2)}$ on $\mathbb{R}^2 \times \mathbb{R}^2$:

$$\alpha^{(2)}(B) = \mathbb{E}\left[\sum_{\xi,\eta\in\mathbf{X},\xi\neq\eta} \mathbf{1}[(\xi,\eta) \in B]\right],\ B \subset \mathbb{R}^2 \times \mathbb{R}^2. \tag{2.10}$$

This gives us the expected number of distinctive pairs of points in any given bivariate product space B. This is analogous to the second moment of a random variable. In fact, it gives us the second moment of the random variable $n(\mathbf{X}_A)$:

$$\mathbb{E}[n(\mathbf{X}_A)^2] = \alpha^{(2)}(A) + \mu(A),\ A \subset \mathbb{R}^2. \tag{2.11}$$

since

$$\alpha^{(2)}(A) = \mathbb{E}[n(\mathbf{X}_A)^2 - n(\mathbf{X}_A)]$$

which is immediate from equation 2.10. In terms of variance of $n(\mathbf{X}_A)$, we have

$$\mathrm{Var}[n(\mathbf{X}_A)] = \alpha^{(2)}(A) + \mu(A) - \mu(A)^2. \tag{2.12}$$

. More generally, for any $B, C \subset S$, we have

$$\mathrm{COV}[n(\mathbf{X}_B), n(\mathbf{X}_C)] = \alpha^{(2)}(B \times C) + \mu(B \cap C) - \mu(B)\mu(C). \tag{2.13}$$

Now if $\alpha^{(2)}$ can be written as

$$\alpha^{(2)}(B) = \int\int \mathbf{1}[(\xi,\eta) \in B]\lambda^{(2)}(\xi,\eta)d\xi d\eta,\ B \subset \mathbb{R}^2 \times \mathbb{R}^2, \tag{2.14}$$

with $\lambda^{(2)} : \mathbb{R}^2 \times \mathbb{R}^2 \to [0,\infty)$, then $\lambda^{(2)}$ is called the *second order product density* or *second order intensity measure.*

Heuristically, $\lambda^{(2)}(\xi,\eta)d\xi d\eta$ is the probability of observing a pair of points occurring in each of two infinitesimally small ball centered at ξ,η with volume $d\xi, d\eta$ respectively. The second order intensity measure is a measure of the correlation between two points in a point pattern. However, it has to be with respect to a Poisson process, i.e., how strong is the correlation compared to CSR. Hence, we usually analyze the normalized second order intensity measure $\lambda^{(2)}(\xi,\eta)$ by dividing $\lambda(\xi)\lambda(\eta)$.

Definition 2.3.1. (Møller and Waagepetersen, 2003, pp. 31)*If both* $\lambda, \lambda^{(2)}$ *are well-defined, the pair-correlation function (PCF) is defined by*

$$g(\xi,\eta) = \frac{\lambda^{(2)}(\xi,\eta)}{\lambda(\xi)\lambda(\eta)}. \tag{2.15}$$

This is widely used in astronomy and astrophysics and it is termed the two-point correlation function (2PCF) in the astronomical literature. PCF measures the ratio of the probability that the point process of interest having two points at ξ and η to that of a Poisson process. If $g(\xi,\eta) = 1$, then it means the point process has the same probability of observing two points each at ξ,η compared to a Poisson point process. If $g(\xi,\eta) > 1$, then it has a greater probability than a Poisson process and vice versa.

To truly facilitate the usage of PCF, we need to have the assumption of stationarity. If we also assume isotropy, the PCF reduces to the following form:

$$g(\xi,\eta) = \rho(||\xi - \eta||), \tag{2.16}$$

where $||\cdot||$ is the usual Euclidean distance and $\rho(\cdot)$ is function taking non-negative values. This means the PCF is only a function of the distance between any pair of points. For the rest of the chapter, I assume we are dealing with stationary and isotropic point processes.

However, for almost all real-life data, stationarity is not satisfied, mainly due to the inhomogeneity present in the first-order intensity and it has to be accounted for.

After addressing the inhomogeneity in the first-order intensity, it is generally assumed that the point process is second-order stationary. This is also referred to as the *second order intensity reweighted stationary* (Møller and Waagepetersen, 2007), i.e., the point process is second order stationary after accounting for potential inhomogeneity in the first order intensity.

It is important to note that when estimating an empirical PCF, the first order intensity $\lambda(\cdot)$ is usually not available and assumed to be constant. This means that obtaining an empirical PCF for an inhomogeneous point pattern is a direct violation of the assumption required for empirical PCF to provide correct second-order property of a point pattern.

There is, however, attempt to construct empirical second order intensity reweighted PCF (Baddeley, Rubak, and Turner, 2015), but a major issue is that the numerical measurement of first order intensity is rarely known. Certain model assumption is usually required. However, even if these model assumption are made, the estimates are usually highly biased. Therefore, we do not pursue this approach here.

2.4 Empirical Summary Statistics

I will now introduce some of the summary statistics and methods for empirically analyzing the behavior of the second order property.

For estimating the empirical PCF, a naive estimator is constructed by constructing a series of concentric annuli encompassing each point in the point pattern and count the number of points in each annuli. Subsequently, an empirical PCF is obtained by averaging the count for each point and plotted against the pairwise distance, after adjusting for the first order intensity of the point pattern. However, this method is highly prone to bias, and the width of the annuli is rather difficult to choose.

A much better estimator is given by the following (Baddeley, Rubak, and Turner,

2015, pp. 228):

$$\hat{\rho}(r) = \frac{|W|}{2\pi r n(n-1)} \sum_{i \neq j}^{n} \kappa(r - d_{ij}) e_{ij}(r) \tag{2.17}$$

where $\kappa(\cdot)$ is a smoothing kernel centered at d_{ij} with a specified kernel bandwidth h which is usually selected by cross-validation. This estimator ensures the estimated empirical PCF is smooth and less prone to bias. Note that d_{ij} here is the observed pairwise distance and e_{ij} is a correction term for edge effects. Edge effects here mean that due to the bounded of the observation window, there may be unobserved points outside of the observation window such that their existence may introduce bias. However, this is not necessary to consider in our data since the intensity of GMCs and SCs decreases drastically towards the boundary of the observation window and it is highly unlikely that there are any unobserved objects outside the observation window.

One thing to note for the estimator given in equation 2.17 is that it will always explode to infinity as $r \to 0$. This is not necessarily the case for many point process. Therefore, a modified version of the estimator is given by the following:

$$\hat{\rho}(r) = \frac{|W|}{2\pi} \sum_{i \neq j}^{n} \kappa(r - d_{ij}) e_{ij}(r)/d_{ij}. \tag{2.18}$$

As mentioned before, if the empirical PCF is compared to 1 at various distances. Clustering pattern exists if the value is greater than 1, repulsive if less 1, and Poisson if equal to 1.

The above summary statistics are defined through correlation structure among points. Another type of summary statistics to characterize the behavior of point pattern is through the empty space function or distance function.

One most used summary statistics in this regard is the nearest neighbor distance

function (Baddeley, Rubak, and Turner, 2015, pp. 262):

$$G(r) = \mathbb{P}[d(\xi, \mathbf{X} \setminus \{\xi\}) \leq r \mid \text{there is a point at } \xi \text{ in } \mathbf{X}], \tag{2.19}$$

where $d(\xi, \mathbf{X} \setminus \{\xi\})$ is the minimum distance from a point in $\mathbf{X}$ to ξ, excluding ξ itself. Equation 2.19 also assumes the first order intensity is homogeneous. Similar to the PCF, if the nearest neighbor distance function is greater than that of the Poisson process, it can be used to suggest clustering behavior, so on and so forth.

Note that equation 2.19 is in fact the cumulative distribution function of the nearest neighbor distance (NND). Therefore, estimating the G-function is quite straightforward. Only a cumulative distribution function is estimated based on the nearest neighbor distance of all points.

For a homogeneous Poisson process with intensity λ, the theoretical G-function is

$$G(r) = 1 - \exp(-\lambda \pi r^2). \tag{2.20}$$

The interpretation of empirical G-function is then the following: if the estimated $\hat{G}(r)$ is less than the theoretical $G(r)$ given in equation 2.20, then this indicates that on average the nearest-neighbor distances in the data are greater than that in a Poisson process with the same average intensity (Baddeley, Rubak, and Turner, 2015, pp. 267). This is indicative of a repulsive pattern and vice versa.

Note that exploratory tools such as PCF and G-function are only sensitive towards certain aspects of the point pattern structure due to their focus on different features of a point pattern. Using only one type of statistics can potentially have blind spots, and as suggested in Baddeley, Rubak, and Turner (2015), certain non-Poisson point pattern can have exactly the same theoretical PCF as a Poisson process. Analyzing point pattern should then in general consider the different features of its structure.

It has been stressed that the comparison using summary statistics, whether through

correlation or spacing, requires the assumption of homogeneity in first order intensity. However, this is predicated on the assumption that the underlying reference point pattern is the homogeneous Poisson process. Directly using empirical PCF or G-function on inhomogeneous point patterns is applicable if the purpose is to compare the fit of a model and the real data, since the reference process is the one specified by the fitted model. However, care needs to be taken to ensure that the model in general captures the intensity variations of the data.

Chapter 3

Gibbs Point Process

In this chapter, I introduce the fundamental theory governing the meaningful construction of Gibbs point process (GPP) models and basic methods for conducting model criticism. Simulation of GPP models and Bayesian inference algorithms are also illustrated.

GPP is a highly flexible way of modelling the physical structure of point processes as the model itself is constructed by empirically modeling the distributional pattern of points. It originated from statistical mechanics under the name Gibbs distribution. It attempts to study an equilibrium system consisting of a vast amount of interacting particles exhibiting complex dependence structure such that directly describing the behavior of each individual particle is impossible. Instead, a probability distribution is constructed by connecting the system's equilibrium physical structure with the energy of the system. One of the first of such model is the infamous Boltzmann distribution. The idea is that for any physical system in equilibrium, its potential energy is most likely to be the lowest. This relates the likelihood of a certain physical structure of the system to the system's energy which in turn is used to determine

the probability measure. This probability measure is termed the Gibbs probability measure and it has close ties to the exponential family distribution.

GPP models, introduced to the statistical literature by Ripley and Kelly (1977) under the name of Markov point process models, borrow the idea of the Gibbs distribution in that they are constructed by physically describing the point pattern structure. If a GPP model is reasonably capturing the physical structure of the process from which the point pattern arises, the likelihood computed from the model is high which corresponds to a low "potential energy" of the system. However, one thing to note is that in point pattern modelling context, the model is not necessarily trying to capture the equilibrium state of the system. It is a merely an attempt to empirically capture the structure present in the point pattern.

Below I provide the basic mathematical theories and methodologies required to construct a well-defined GPP model. I will also introduce several widely used GPP models and their applications. To offer a less demanding and rigorous introduction on GPP models, I will follow Baddeley et al. (2007) and partially Møller and Waagepetersen (2003) as well as Baddeley, Rubak, and Turner (2015). For highly technical and rigorous definition and construction of GPP models, see for example Daley and Vere-Jones (2008).

3.1 Finite Point Process with a Density

To define a GPP, we need certain condition on its support. Suppose a point process $\mathbf{X}$ on S satisfies $\boldsymbol{N}(\mathbf{X}) < \infty$ where $\boldsymbol{N}(\mathbf{X})$ is the random variable counting the total number of points on S. $\mathbf{X}$ is then a finite point process (Baddeley et al., 2007, pp. 61) and it belongs to the space

$$\mathcal{N}^f = \{\mathbf{x} \subset S : \boldsymbol{N}(\mathbf{x}) < \infty\}.$$

As noted in Chapter 2, a point process density only exists with respect to another point process, usually a Poisson process. We then let π_μ be the distribution of a Poisson process with intensity measure μ.

Definition 3.1.1. (Baddeley et al., 2007, pp. 62) *Suppose* $f : \mathcal{N}^f \to \mathbb{R}^+$ *is a measurable function satisfying* $\int_{\mathcal{N}^f} f(\mathbf{x})\pi_\mu(d\mathbf{x}) = 1$. *Let*

$$\mathbf{P}(A) = \int_A f(\mathbf{x})\pi_\mu(d\mathbf{x}) \tag{3.1}$$

for any $A \in \mathcal{N}$. $\mathbf{P}$ *is then a point process distribution with respect to the Poisson process with intensity* μ *and* f *is called the probability density function of the point process.*

Now for a point process $\mathbf{X}$ with probability density f,

$$\mathbb{P}(\mathbf{X} \in A) = \sum_{n=0}^{\infty} \frac{e^{-\mu(S)}}{n!} \int_S \cdots \int_S \mathbf{1}\left[\{x_1, \ldots, x_n\} \in A\right] f\left(\{x_1, \ldots, x_n\}\right) \mu(dx_1) \ldots \mu(dx_2). \tag{3.2}$$

Now if we let $\lambda > 0$, and set

$$f(\boldsymbol{x}) = \alpha\lambda^{n(\boldsymbol{x})}$$

where α is chosen so that $f(\cdot)$ is a density function, then this immediately gives us the probability density function of a homogeneous Poisson process with intensity λ. This also corresponds to the result in Proposition 2.2.1.

Take a closer look at the form of the density function of the homogeneous Poisson process, we see that the information about the point process is completely encoded by how many points, $n(\boldsymbol{x})$, there are in the observation window. The only free parameter λ in turn controls the expected count of the process. The higher the value of λ, the more points there tend to be in the observation window. Now if the Poisson process is

inhomogeneous, then λ can now be considered as a function of location or covariate, i.e., the $\lambda = \lambda(s)$ where s denotes location.

An interesting question arises in that how does a GPP model capture interpoint interaction since Poisson process, either homogeneous or inhomogeneous, does not encode anything about dependence structure between points due to the fact it only concerns the occurrence of individual points. This leads to the Gibbs representation of the general form of GPP models.

Gibbs Representation

Definition 3.1.2. (Baddeley et al., 2007, pp. 66)*A **finite Gibbs process** is a finite point process* $\mathbf{X}$ *with probability density* $f(\cdot)$

$$f(\boldsymbol{x}) = \exp\left(V_0 + \sum_{x\in\mathbf{x}} V_1(x) + \sum_{\{x,y\}\subset\mathbf{x}} V_2(x,y) + \sum_{\{x,y,z\}\subset\mathbf{x}} V_3(x,y,z) + \dots\right) \tag{3.3}$$

where V_k *is called the* k*-th order potential or potential of order* k*.*

We can now immediately see that for a homogeneous Poisson process,

$$\alpha = \exp(V_0),$$

$$\lambda = \exp(V_1(x)),\ x \in \mathbf{x},$$

$$V_k = 0,\ \forall k \geq 2.$$

This means that for a Poisson process, any potential of order greater than or equal to 2 vanishes. It is easy to see now that the interpoint interaction is in fact captured by the potential of order greater than or equal to 2. For a Poisson process, this precisely conveys the notion of independence between points, i.e., there is no interpoint interaction. In statistical mechanics, the negative of the second order potential, $-V_2$,

is simply called the potential energy. It characterizes the amount of energy the system has to overcome to place two points at x and y respectively. In a probabilistic context, it also represents the contribution to the likelihood that two points occur at x and y. Note that the quantitative relation between the potential and the likelihood is reversed in that if it takes an infinite amount of energy to place two points at x and y, it then means there is 0 probability for the process to produce two points at x and y, which makes physical and intuitive sense.

Papangelou Conditional Intensity

Constructing GPP models through the full probability density can be complicated. With sufficient conditions, however, we can fully specify a GPP model through its Papangelou conditional intensity which is much simpler and easier to interpret.

Definition 3.1.3. (Baddeley et al., 2007, pp. 67)*Let f be the probability density function of a finite point process $\mathbf{X}$ in some bounded observation window $W \subset \mathbb{R}^2$. If*

$$f(\mathbf{x}) > 0 \implies f(\mathbf{y}) > 0, \ \forall \mathbf{y} \subset \mathbf{x},$$

then f is called hereditary with respect to $\mathbf{X}$ or simply hereditary and f can be expressed in the Gibbs form given by equation (3.3).

Definition 3.1.4. (Baddeley et al., 2007, pp. 65)*Let f be the probability density function of a finite point process $\mathbf{X}$ in some bounded observation window $W \subset \mathbb{R}^2$. If f is hereditary, then the **Papangelou conditional intensity** (or simply **conditional intensity**) of $\mathbf{X}$ exists almost everywhere and is given by*

$$\lambda(\xi, \mathbf{x}) = \frac{f(\mathbf{x} \cup \xi)}{f(\mathbf{x})} \tag{3.4}$$

Conditional intensity is highly useful in that it has a one to one relationship with

the corresponding full density function as long as f is hereditary which is a quite simple condition to meet. Moreover, it does not involve the normalizing constant in the full density function, $\exp(V_0)$, which is unknown for most models. Lastly, it characterizes the contribution to the likelihood while a point ξ is added to the existing point pattern. This is very useful in understanding the structure of the point process arises from the model and as we will see later, it is a crucial component for the meaningful construction of GPP models as well as simulation of GPP models.

Repulsive and Clustering

With the definition of conditional intensity, we can subsequently precisely define repulsion and clustering for a point process.

Definition 3.1.5. (Møller and Waagepetersen, 2003, pp. 83)*Let* $\mathbf{X}$ *be a finite point process and* f *its corresponding probability density function while* λ *is the conditional intensity, then* $\mathbf{X}$ *is repulsive if*

$$\lambda(\xi, \mathbf{x}) \geq \lambda(\xi, \mathbf{y}), \quad \textit{if } \mathbf{x} \subset \mathbf{y},$$

$\mathbf{X}$ *is clustered if*

$$\lambda(\xi, \mathbf{x}) \leq \lambda(\xi, \mathbf{y}), \quad \textit{if } \mathbf{x} \subset \mathbf{y}.$$

It is intuitive here that if the conditional intensity increases when there are more points in a given region, then the point process is clustered and vice versa.

3.2 Pairwise Interaction Process

After introducing the basic theorems and definitions of GPP models, we are ready to consider some of the widely-used GPP models. One important class of model is the **pairwise interaction process** (Møller and Waagepetersen, 2003, pp. 84). This

class of model is obtained by setting $V_k = 0, \forall k > 2$ where V_k is given in equation (3.3). This means pairwise interaction model assumes the interpoint interaction only comes from pairs of points. In this case, the probability density function of a point process $\mathbf{X}$ can be expressed as

$$f(\mathbf{x}) = \alpha \prod_{i=1}^{n(\mathbf{x})} \lambda(x_i) \prod_{i \neq j} h(x_i, x_j), \ x_i, x_j \in \mathbf{x}. \tag{3.5}$$

Note that this is in fact a reparametrization of the Gibbs representation given by equation 3.3. One thing to note is that if $\lambda(\cdot)$ is constant, i.e., the point process is first order homogeneous, then $h(\cdot, \cdot)$ can serve as a parametric characterization of the PCF of $\mathbf{X}$ (Goldstein et al., 2014). Although this does not equate to saying that the pairwise-interaction is equal to the theoretical PCF. This is because pairwise-interaction model assumes the internal interaction among points is only due to second order interaction, while the interaction exhibited by PCF can be potentially affected by higher order interaction indirectly. Nevertheless, the pairwise-interaction term does indeed have a very tight relationship with one another. In a more general setting, where λ is not homogeneous, $h(\cdot, \cdot)$ will significantly deviate from the PCF and it can no longer be used as a parametric representation of the PCF.

It is noteworthy that the expected value of $n(\mathbf{X})$ in a pairwise interaction model is usually not available analytically due to the presence of interpoint interaction and the final intensity of the point process is in fact a combined effect from the first and second order potential. We can consider the first order potential as the "fertility" rate of the point process or in chemistry, it is termed the chemical activity rate/function. In the context of this research, it represents the amount "resource" there is to produce a SC. The second order potential here will have different interpretation depends on the form of $h(\cdot, \cdot)$. If h is constructed to model a process with second order clustering, it represents the amount of triggering or chain reaction of the occurrence of points. If

h is to model a repulsive process, it then conveys the idea of competition, e.g., species competing for limited resources or the fact that there is only a limited amount of resource available to produce SCs in any given region.

It is noteworthy that the pairwise interaction models are hereditary. Therefore, one can construct model by specifying its conditional intensity, i.e., one only needs to specify the following

$$\lambda(\xi, \mathbf{x}) = \lambda(\xi) \prod_{i=j}^{n(\mathbf{x})} h(\xi, x_i).$$

Below are several pairwise interaction model that are widely used.

Definition 3.2.1. (Møller and Waagepetersen, 2003, pp. 85)*A finite point process* $\mathbf{X}$ *in an observation window* $W \subset \mathbb{R}^2$ *is called a Strauss process with interacting range* r *if its conditional intensity is given as*

$$\lambda(\xi, \mathbf{x}) = \lambda \gamma^{\sum_{i=1}^{n(\mathbf{x})} \mathbf{1}[d(\xi, x_i) \leq r]}.$$

λ is the exponentiated first order potential, and $d(\cdot, \cdot)$ is the standard Euclidean distance. What makes Strauss process interesting is that the statistics considered for the second order potential is the number of pairs of points lie within distance r from each other. We can see that for the second order potential

$$V_2(x, y) = \log(\gamma) \mathbf{1}[d(x, y) \leq r].$$

It is then natural to see that γ here controls the second order behavior. If $\gamma \in (0, 1)$, then Strauss process is a repulsive process in the range $[0, r]$. If $\gamma = 1$, it then reduces to a homogeneous Poisson process. Naturally, one would think that $\gamma > 1$ will result in a clustered process. However, for a Strauss process, the probability density function is undefined when $\gamma > 1$ since the normalizing constant α does not exist. To see this, consider a scenario where all the points in $\mathbf{x}$ are within r distance from each other,

then from equation (3.2),

$$\begin{aligned}\frac{1}{\alpha} &\geq \sum_{n=0}^{\infty} \frac{e^{-\mu(S)}}{n!} \int_S \cdots \int_S \lambda^n \gamma^{n(n-1)/2} \mu(dx_1) \ldots \mu(dx_2) \\ &= \sum_{n=0}^{\infty} \frac{e^{-\mu(S)}}{n!} \lambda^n \gamma^{n(n-1)/2} \mu(S)^n\end{aligned}$$

and this infinite sum goes to infinity which can be shown through the Stirling's approximation. Therefore,

$$\frac{1}{\alpha} \geq \infty,$$

hence, the probability density function is undefined when $\gamma > 1$. If one tries to simulate a Strauss point process with $\gamma > 1$, the resulting simulation will produce point patterns with super-clustering behavior, i.e., majority of the points will clump together in several small regions.

Now for a Strauss process, if $\gamma \to 0$, it then becomes a so-called *hard-core* process.

Definition 3.2.2. (Møller and Waagepetersen, 2003, pp. 85-86) *A finite point process* $\mathbf{X}$ *in an observation window* $W \subset \mathbb{R}^2$ *is called a hard-core process with interacting range* r *if its conditional intensity is given as*

$$\lambda(\xi, \mathbf{x}) = \lambda \mathbf{1}[d(\xi, x_i) > r, x_i \in \mathbf{x}]. \tag{3.6}$$

This means that in a hard-core process, no pair of points can be less than r distance from each other. We can think of the points as the center of physical objects such as marbles with radius $r/2$. This lends the hard-core process the ability to model various point patterns consisting of physical objects such as cells or stars.

As noted before, if the first order potential is homogeneous, the pairwise interaction term is a parametric representation of the PCF if it is a function of the pairwise distance. However, we have seen that all pairwise interaction models presented above have interaction term as a piecewise function of the pairwise distance and all have

discontinuous jumps at the interaction range r. The main reason for choosing such a form is due to the difficulty in inference methodologies. GPP models are traditionally fitted using MLE-based approach and MLE approaches work the best when the model is in log-linear form. This restriction forces many pairwise interaction term to be a piecewise function of the pairwise distance. In many applications of GPP models, such as in ecology, a piecewise form of the interaction function is usually acceptable. However, in the context of this research, model the interaction function as a continuous function of the pairwise distance is better suited to obtain physical insights. Below we present several GPP models that possess continuous interaction term.

Definition 3.2.3. (Baddeley, Rubak, and Turner, 2015, pp. 515)*A finite point process* $\mathbf{X}$ *in an observation window* $W \subset \mathbb{R}^2$ *is called a soft-core process with scale* σ *and shape* κ *if its conditional intensity is given as*

$$\lambda(\xi, \mathbf{x}) = \lambda \exp\left(-\sum_{i=1}^{n(\mathbf{x})} \left(\frac{\sigma}{d(\xi, x_i)}\right)^{2/\kappa}\right). \tag{3.7}$$

Definition 3.2.4. (Møller and Waagepetersen, 2003, pp. 88)*A finite point process* $\mathbf{X}$ *in an observation window* $W \subset \mathbb{R}^2$ *is called a very soft-core process with scale* σ *if its conditional intensity is given as*

$$\lambda(\xi, \mathbf{x}) = \lambda \prod_{i=1}^{n(\mathbf{x})} \left(1 - \exp\left(-\frac{d(\xi, x_i)^2}{\sigma^2}\right)\right). \tag{3.8}$$

Definition 3.2.5. (Møller and Waagepetersen, 2003, pp. 88)*A finite point process* $\mathbf{X}$ *in an observation window* $W \subset \mathbb{R}^2$ *is called a Lennard-Jones process with characteristic diameter* σ *and well depth* ϵ *if its conditional intensity is given as*

$$\lambda(\xi, \mathbf{x}) = \lambda \exp\left[-4\epsilon \sum_{i=1}^{n(\mathbf{x})} \left(\frac{\sigma}{d(\xi, x_i)}\right)^{12} - \left(\frac{\sigma}{d(\xi, x_i)}\right)^{6}\right]. \tag{3.9}$$

The processes presented above are all considered infinite range interaction models in that the interaction potential is not equal to 0 and it only approaches 0 as pairwise distance goes to infinity. It is important to note that the soft-core and very soft-core process are both repulsive processes while Lennard-Jones process is neither repulsive nor clustering since it exhibit strong repulsion at very short distance and clustering at medium distance. These processes have been quite successful in understanding numerous physical phenomena such as the distribution of molecules using Lennard-Jones process. However, as noted before, all the above processes are not in log-linear form and inference using MLE approach can be quite difficult. Furthermore, models such as soft-core and Lennard-Jones process are highly prone to numerical instability since the pairwise distance occur in the denominator.

As we have seen, there can be countless ways to construct a pairwise interaction model. However, meaningful construction of a new pairwise interaction model is not arbitrary and one still needs to adhere to certain ground rules. First, the interaction term has to be non-negative to ensure the probability density is non-negative. Second, the second order potential has to eventually approaches 0 as pairwise distance goes to infinity since for any two points at distance far away from each other, there should be negligible or no interaction between them. Most importantly, sensible construction of any GPP models have to adhere to a certain stability criteria so that simulation and inference is possible.

Proposition 3.2.1. *(Ruelle, 1969; Møller and Waagepetersen, 2003) Let* $\phi : S \to \mathbb{R}^+$ *be some function s.t.* $\int_S \phi(\xi)\mu(d\xi) < \infty$. *Let* $\lambda(\cdot, \cdot)$ *be the conditional intensity of a finite point process* $\mathbf{X}$ *with probability density* f. $\mathbf{X}$ *(or* f*) is called locally stable if*

$$\lambda(\xi, \mathbf{x}) \leq \phi(\xi).$$

X *(or f) is called Ruelle stable if* $\forall \mathbf{x} \in \mathcal{N}^f$, $\exists c > 0$,

$$f(\mathbf{x}) \leq c \prod_{x \in \mathbf{x}} \phi(x).$$

Proposition 3.2.1 provides the theoretical groundwork for ensuring that a GPP model is defined. Locally stable implies Ruelle stable. A finite point process is Ruelle stable means that the measure of the process is dominated by a Poisson process and therefore, the point process exists with respect to a Poisson process. Local stability, on the other hand, is a stronger criteria and it ensures the successful simulation of point patterns from given GPP models. We now can see a much general reason why a Strauss process is undefined when $\gamma > 1$ since it is not Ruelle stable.

3.3 Multivariate Point Processes

If the point patterns considered consist of multiple types of points, such GMCs and YSCCs studied in this research, and interest is on how they interact/correlate with each other, then a bivariate point process should be considered. Here I give a brief outline for the bivariate point process under the framework of GPP. For simplicity, I only discuss the definition of multivariate point process under the framework of pairwise-interaction. Furthermore, only point patterns of two types are considered. Models for point patterns consist of more than two types of points can be easily extended.

Suppose two point processes, $\mathbf{X}_A$ and $\mathbf{X}_B$, form a bivariate Gibbs point process (Isham, 1984), defined on the same observation window $W \subset \mathbb{R}^2$, then it has the following joint probability density function

$$f(\mathbf{x}_A, \mathbf{x}_B) = \alpha \prod_{i=1}^{n(\mathbf{x}_A)} \lambda_A(x_i) \prod_{j=1}^{n(\mathbf{x}_B)} \lambda_B(x_j) \phi_A(\mathbf{x}_A) \phi_B(\mathbf{x}_B) \phi_{AB}(\mathbf{x}_A, \mathbf{x}_B). \tag{3.10}$$

Similar to the univariate point process, λ_A, λ_B control the first-order potential, $\phi_A(\mathbf{x}_A)$, $\phi_B(\mathbf{x}_B)$ characterise the intra-type interaction in $\mathbf{x}_A, \mathbf{x}_B$ respectively. The extra term $\phi_{AB}(\mathbf{x}_A, \mathbf{x}_B)$ denotes the inter-type interaction/correlation between the points of $\mathbf{x}_A, \mathbf{x}_B$. It is noteworthy that similar to the relationship between the empirical 2PCF/PCF and the intra-type interaction term, $\phi_{AB}(\mathbf{x}_A, \mathbf{x}_B)$ also has an empirical counterpart in that it represents the cross-type 2PCF/PCF (Baddeley, Rubak, and Turner, 2015) between $\mathbf{x}_A$ and $\mathbf{x}_B$. Certainly, the assumption that both types of point processes are first order homogeneous is required. The cross-type PCF is a generalisation of the PCF in that it looks at the ratio of the probability of observing a point in type A at r distance away from a point in the type B to that of a case where the two are uncorrelated, assuming stationarity between the two types of points.

3.3.1 Motivation for Hierarchical Interaction

If further information is available that there exists a form of hierarchy between two types of points, i.e., one type takes precedence before another, then it is more appropriate to consider a hierarchical structure between the two processes through conditional probability density.

The model in equation 3.10 was first formally studied by Isham (1984) after Harkness and Isham (1983) studied the distribution of nests of two types of ants. However, Högmander and Särkkä, 1999 noticed the fact that one species of ants *Cataglyphis* sets up nests closer to other species of ants, while one of the other species, *Messor*, chooses to live close to food sources but does not compete with *Messor*. In this scenario where there is a natural order or asymmetry between types of points, it is no longer appropriate to formulate the model through bivariate point process defined by equation 3.10.

Högmander and Särkkä (1999) then introduced a hierarchical model through conditional probability arguments. They separate the joint probability density of two

point processes through conditioning where a notion of high and low-level processes is installed. They regarded the high-level point process (*Messor*) as a univariate process that takes precedent before the low-level point process (*Cataglyphis*) and first fit a model for the high-level process only. Conditioning on the realization of the high-level process, they then fit a model for the low-level process including an interaction term to model the influence from the high process to the low process.

This form of hierarchy is very similar to a phenomena of asymmetric interaction mentioned by Rajala, Murrell, and Olhede (2018) where they analyzed the interaction between around 300 types of species. They found that a certain type of species interacts significantly with only one species when the other species are randomized in a MC test, but it interacts significantly with 22 species when itself is the one randomized in the test. This asymmetric effect is present possibly due to factors such as ecological dominance. However, the spatial correlation emerging from any form of asymmetric interaction should still be symmetric (Rajala, Murrell, and Olhede, 2018). Hence, even if asymmetric hierarchical structure exists, the cross-type interaction should still be symmetric and any probability model constructed for the processes should take that into account.

It is noteworthy that hierarchical structure does not mean that the low level process does not affect the high level process. It only means that the high-level process takes precedence before the low-level process and the dependence of high-level process on the low-level process is not explicitly specified (Baddeley, Rubak, and Turner, 2015, pp. 623).

The reason for considering hierarchy structure in this research is that there is a natural formation hierarchy from CO filament to GMCs and GMCs to SCs. Therefore, the process generating the CO filament takes precedence before GMCs and GMCs before SCs. It is then natural to incorporate a hierarchical structure to model this formation hierarchy.

Next, we specify the hierarchical GPP model. Consider first the high-level process $\mathbf{X}_A$, which is the process for GMCs. It has a probability density

$$f(\mathbf{x}_A;\ \boldsymbol{\theta}_A) = \alpha_A \prod_{i=1}^{n(\mathbf{x}_A)} \lambda_A(x_i)\phi_A(\mathbf{x}_A) \tag{3.11}$$

where $\boldsymbol{\theta}_A$ is the vector of model parameters. In this setting, we treat the point pattern of $\mathbf{x}_A$ as random and it is object of concern.

For the lower level process, we now treat the realization of $\mathbf{x}_A$ as given, and $\mathbf{x}_B$ has probability density

$$f_{\mathbf{x}_A}(\mathbf{x}_B;\ \boldsymbol{\theta}_B) = \alpha_B(\mathbf{x}_A) \prod_{i=1}^{n(\mathbf{x}_B)} \lambda_B(x_i)\phi_B(\mathbf{x}_B)\phi_{AB}(\mathbf{x}_A, \mathbf{x}_B). \tag{3.12}$$

Note that now the normalizing constant in the density of $\mathbf{x}_B$ depends on the realization of $\mathbf{x}_A$.

Notice that the joint probability density for $\mathbf{x}_A$ and $\mathbf{x}_B$ is now

$$f(\mathbf{x}_A, \mathbf{x}_B) = \alpha_A\alpha_B(\mathbf{x}_A) \prod_{i=1}^{n(\mathbf{x}_A)} \lambda_A(x_i) \prod_{j=1}^{n(\mathbf{x}_B)} \lambda_B(x_j)\phi_A(\mathbf{x}_A)\phi_B(\mathbf{x}_B)\phi_{AB}(\mathbf{x}_A, \mathbf{x}_B) \tag{3.13}$$

which seems very similar to equation 3.10 but the model is inherently asymmetric and fundamentally different from the symmetric model.

The detailed model construction will be deferred to Chapter 4. Next, I will provide the basic methodology for carrying out simulation and Bayesian inference for GPP models.

3.4 Simulation and Inference for Gibbs Point Process Models

This section introduces the basics for simulating point patterns based on a given GPP model as well as the inference algorithms for obtaining the model parameters. For

simulation of a GPP model, it is done through the spatial birth and death Metropolis-Hasting algorithm by Geyer and Møller (1994). For inference on the model parameters, there are usually two approaches, one being the frequentist approach by using maximum likelihood estimation, the other being Bayesian inference through MCMC sampling.

However, for a GPP model and many other complex statistical and probabilistic model, conducting inference is problematic. This is because GPP models and many other complex models belong to a family of partially-intractable distributions since the likelihood is specified up to an intractable normalizing constant.

Baddeley and Turner (2000) proposed a maximum pseudo-likelihood estimator (MPLE) for Markov type point process model where a pseudo-likelihood is constructed to approximate the true likelihood. However, this model has a restriction in that the unnormalized likelihood must be of log-linear form. This is problematic for models with irregular parameters, i.e., parameters that are not of log-linear form with sufficient statistics. Although methods such as profile likelihood can be used to estimate irregular parameters, it does not provide the ability to construct confidence intervals for them. Geyer (1991) proposed MCMC-MLE to approximate the unknown normalizing constant using an importance sampling scheme with a series of MCMC simulations. However, this performs poorly when the likelihood function is complex and not of log-linear form, as MCMC-MLE requires a good approximation of the gradient of the unnormalized likelihood. For complex and non log-linear model such as the one constructed in this research, the gradient estimate can be highly prone to numerical instability and even unavailable. Furthermore, initialization of the algorithm is troublesome as it requires a grid search over the parameter space.

Bayesian inference through MCMC sampling, however, does not have the restriction on the model being of any form. It is also much easier to implement compared to the aforementioned MLE approaches. Furthermore, it naturally fits into the nature of

scientific process of astronomy where new data and information will usher in with the construction and implementation of more and more powerful telescopes. This form of updating in information and knowledge fits very well with the Bayesian paradigm. However, the issue of unnormalized likelihood is still present and the intractable likelihood results in a"doubly-intractable" (Murray, Ghahramani, and MacKay, 2006) posterior distribution when Bayesian inference is applied. Several methods based on the standard MCMC algorithm, such as Murray, Ghahramani, and MacKay (2006), Møller et al. (2006), and Liang (2010), have been proposed to deal with this issue. I will illustrate the basics of the algorithms in later sections.

Next, I will provide an introduction to the construction of a simulation algorithm for point process with a specified GPP models as well as Bayesian inference algorithms for GPP models.

3.4.1 Simulation of Gibbs Point Process

Simulation of point patterns is crucial for inference and model criticism, and to ensure the model constructed can indeed capture the spatial distribution of GMCs and YSCCs to a reasonable extent, we need to be able to simulate the point patterns based on the model. In this section, therefore, we introduce the birth-death Metropolis-Hasting algorithm (BDMH) (Geyer and Møller, 1994) and specify how to adjust it to suit our simulation purpose.

The BDMH algorithm is a variant of the famous Metropolis-Hasting (MH) algorithm developed to simulate spatial point patterns with a specified unnormalized probability density $h(\mathbf{x})$. The state of the Markov chain at each time step is a point pattern, and we denote the state of the chain at time t by $\mathbf{X}_t$. At each t, a point is either being added ("born") to the point pattern with probability p_b or removed ("dead") from the point pattern with probability $p_d = 1 - p_b$. If a point is to be born, it is selected according to some probability density $b(\mathbf{X}_t; \xi)$ over the observation win-

dow where ξ is the newly added point; if a point is to be removed, it is selected with another probability density $d(\mathbf{X}_t; \xi)$ where ξ is the point to be removed. Lastly, we calculate the acceptance probability for the proposed move and determine whether the proposal is accepted or not.

To formalize the algorithm, let

$$\mathbf{X}^+ = \mathbf{X}_t \cup \{\xi\}$$

be the point pattern formed when adding ξ into $\mathbf{X}_t$ and

$$\mathbf{X}^- = \mathbf{X}_t \setminus \{\xi\}$$

be the point pattern formed when removing ξ from $\mathbf{X}_t$. A pseudo-code of the BDMH algorithm is then given in Algorithm 1:

Algorithm 1: Metropolis-Hasting Birth and Death Algorithm

Input: Initial point pattern $\mathbf{X}_0$, number of iterations T, birth probability p_b, death probability p_d, birth density $b(\cdot;\cdot)$, death density $d(\cdot;\cdot)$;
for *t = 1,...,T* **do**
  Draw $U \sim \text{unif}(0, 1)$;
  if $U < p_b$, **then**
    Generate $\xi \sim b(\mathbf{X}_t; \xi)$;
    Calculate $r_b = \dfrac{h(\mathbf{X}^+)d(\mathbf{X}^+; \xi)p_d}{h(\mathbf{X}_t)b(\mathbf{X}_t; \xi)p_b}$;
    Accept $\mathbf{X}^+$ with probability $a_b = \min(1, r_b)$;
  else
    Select $\xi \sim d(\mathbf{X}_t; \xi)$ from $\mathbf{X}_t$;
    Calculate $r_d = \dfrac{h(\mathbf{X}^-)b(\mathbf{X}^-; \xi)p_b}{h(\mathbf{X}_t)d(\mathbf{X}_t; \xi)p_d}$;
    Accept $\mathbf{X}^-$ with probability $a_d = \min(1, r_d)$;
  end
end

For a normal planar GPP, such as the Strauss process, we usually set

$$p_b = p_d = \frac{1}{2}; \quad b(\mathbf{X}_t; \xi) = \frac{1}{|D_s|}, \; d(\mathbf{X}_t; \xi) = \frac{1}{n(\mathbf{X}_t)},$$

where $|D_s|$ is the area of the observation window S. In this case, a point is selected uniformly from the observation window when it is to be added and uniformly selected from the existing pattern when it is to be removed. However, this form of specification performs extremely poorly for simulating the point pattern resembling the distribution of GMCs and YSCCs. This is because the majority of GMCs are extremely close to the CO filament, randomly choosing a point in the observation window for the birth proposal is highly unlikely to fall close enough to the CO filament. This results in almost all birth proposal being rejected and the algorithm will take extremely long time to converge or may not converge at all. The same happens for simulating YSCCs as YSCCs lie extremely close to GMCs as well.

To avoid this problem, I choose to adopt the birth density as

$$b(\mathbf{X}_t; \xi) \propto \left(1 + \frac{d^2(\xi, y)}{h^2}\right)^{-1} \tag{3.14}$$

where $d(\xi, y)$ is the distance from ξ to the closest point y on the CO filament and h is some parameter to be chosen depends on how correlated the points are with the CO filament. To obtain a birth proposal from the above birth density, I employ a simple rejection sampling procedure. Firstly, obtain a large amount of sample points that follow the above birth density through rejection sampling, then randomly choose a point from this sample each time a birth proposal is selected. A same birth density is chosen for simulating YSCCs where y is replaced by a point in GMCs. I will provide the detailed numerical choice for these hyperparameters in Chapter 5.

3.4.2 Bayesian Inference for Gibbs Point Process Models

In this section, I introduce the basic Bayesian inference algorithm including Markov chain Monte Carlo (MCMC) algorithm and its adaptation for GPP models. Bayesian inference is a natural and systematic method to infer properties of model parameters in the context of this research since astronomical observation data can be constantly updated through the construction and employment of more and more powerful telescopes. In many cases, it also possess a numerical advantage over the traditional MLE approach where gradient computation in MLE can potentially break down due to the complexity of the model. However, the standard method for Bayesian inference such as Metropolis-Hasting (MH) algorithm is not feasible for our model since our likelihood function itself contains an unnormalized constant which is a function of the parameters. I will, therefore, illustrate the basic MCMC algorithm and methods for implementing MCMC algorithm for GPP models.

Markov Chain Monte Carlo

For a usual statistical model with probability density

$$f(\mathbf{x};\boldsymbol{\theta}),$$

with $p(\boldsymbol{\theta})$ as the prior distribution, the posterior distribution for the parameters is then

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x};\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{3.15}$$

In general, the posterior distribution involves an intractable integral which is a function of the data $\mathbf{x}$. To facilitate posterior sampling, an arbitrary proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ is chosen to facilitate a diffusion process to explore the posterior state space. The standard procedure of MCMC sampling is then the MH algorithm given in Al-

gorithm 2.

Algorithm 2: Metropolis-Hasting Algorithm

Input: Initial $\boldsymbol{\theta}$, number of iterations T;
for $t = 1, ..., T$ **do**
  Propose $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})$;
  Calculate $r = \dfrac{f(\mathbf{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}$;
  Accept $\boldsymbol{\theta}'$ with probability $a = \min(1, r)$;
end

However, Algorithm 2 works under the assumption that the probability density $f(\cdot|\cdot)$ is tractable for all possible $\boldsymbol{\theta}$. This is not the case for the probability density of a GPP model. In general, a GPP model can be written as

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{h(\mathbf{x}|\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta})}, \tag{3.16}$$

where $h(\mathbf{x}|\boldsymbol{\theta})$ is the part that we can define and $\mathcal{Z}(\boldsymbol{\theta})$ is an intractable normalizing constant which is a function of the parameters $\boldsymbol{\theta}$. f in this case is called a doubly-intractable distribution (Murray, Ghahramani, and MacKay, 2006).

For a doubly-intractable distribution, the problem arises when we calculate the MH ratio

$$r = \frac{f(\mathbf{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{f(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})} = \frac{h(\mathbf{x}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')\mathcal{Z}(\boldsymbol{\theta})}{h(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})\mathcal{Z}(\boldsymbol{\theta}')} \tag{3.17}$$

where the ratio

$$\frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')}$$

is unknown. This makes the acceptance ratio in a MH-update unavailable to us and normal MCMC sampling cannot proceed.

The first method proposed to tackle the issue is called the single auxiliary variable method (SAVM) by Møller et al. (2006). Instead of sampling from the original posterior distribution, Møller et al. (2006) proposed to extend the state space of the

posterior distribution with an auxiliary variable $\mathbf{y}$ living in the same space as the data $\mathbf{x}$, and it has conditional density $g(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})$. In a GPP context, this means that $\mathbf{y}$ is another point pattern. Now the augmented posterior distribution becomes

$$\pi(\boldsymbol{\theta}, \mathbf{y}|\mathbf{x}) \propto g(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})h(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})/\mathcal{Z}(\boldsymbol{\theta}). \tag{3.18}$$

Marginalization of the augmented posterior distribution over $\mathbf{y}$ will then return it back to the original posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$.

Now with the new posterior distribution, a change in the proposal distribution is also needed. However, the proposal distribution is still arbitrary as in the case for MH algorithm. In this case, the proposal for $\boldsymbol{\theta}'$ can stay as $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$. The ingenious part of the SAVM algorithm is the choice for the proposal of $\mathbf{y}'$ where Møller et al. (2006) set it as

$$q(\mathbf{y}'|\boldsymbol{\theta}', \boldsymbol{\theta}, \mathbf{y}) = f(\mathbf{y}'|\boldsymbol{\theta}') = \frac{h(\mathbf{y}'|\boldsymbol{\theta}')}{\mathcal{Z}(\boldsymbol{\theta}')}. \tag{3.19}$$

This means that if one is able to simulate $\mathbf{y}'$ perfectly, then the SAVM algorithm is a valid MCMC algorithm that will produce a Markov chain with stationary distribution equal to the augmented posterior distribution. Then the original posterior distribution we desire can be obtained through simple marginalization.

Now if we compute the MH ratio, it is then

$$r_{\text{SAVM}} = \frac{g(\mathbf{y}'|\boldsymbol{\theta}', \mathbf{x})h(\mathbf{x}|\boldsymbol{\theta}')h(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{g(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})h(\mathbf{x}|\boldsymbol{\theta})h(\mathbf{y}'|\boldsymbol{\theta}')p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}. \tag{3.20}$$

We can see that now the unknown ratio between the normalizing constants cancel and every term in the MH ratio can be computed. The only term left to be specified is the conditional density of $g(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})$ of $\mathbf{y}$. Møller et al. (2006) suggests that it is best to mimic the distribution of the original probability model, i.e.,

$$g(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x}) = f(\mathbf{y}|\tilde{\boldsymbol{\theta}}),$$

i.e., choose a fixed parameter $\tilde{\boldsymbol{\theta}}$ that is approximately the mode of the posterior distribution. However, this is rather difficult to do since finding the posterior mode is in a sense what the MCMC algorithm tries to do. This is one of the main issue of SAVM algorithm since choosing $g(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})$ is essential in ensuring a reasonable performance of the algorithm (Park and Haran, 2018).

On the other hand, the SAVM algorithm is in fact using a two-sample importance sampling scheme to approximate the unknown ratio through $\mathbf{y}$ and $\mathbf{y}'$:

$$\frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')} \approx \frac{g(\mathbf{y}'|\boldsymbol{\theta}', \mathbf{x})h(\mathbf{y}|\boldsymbol{\theta})}{g(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})h(\mathbf{y}'|\boldsymbol{\theta}')}. \tag{3.21}$$

Murray, Ghahramani, and MacKay (2006) later suggested that the importance sampling scheme can be much simpler through a one-sample importance sampling estimate:

$$\frac{\mathcal{Z}(\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta}')} \approx \frac{h(\mathbf{y}|\boldsymbol{\theta})}{h(\mathbf{y}|\boldsymbol{\theta}')} \tag{3.22}$$

where $\mathbf{y} \sim f(\cdot|\boldsymbol{\theta}')$, and they proposed a new MCMC algorithm called the exchange algorithm by slightly altering the augmented posterior distribution as follow:

$$\pi(\boldsymbol{\theta}, \boldsymbol{\theta}', \mathbf{y}|\mathbf{x}) \propto p(\boldsymbol{\theta})\frac{h(\mathbf{x}|\boldsymbol{\theta})}{\mathcal{Z}(\boldsymbol{\theta})}q(\boldsymbol{\theta}'|\boldsymbol{\theta})\frac{h(\mathbf{y}|\boldsymbol{\theta}')}{\mathcal{Z}(\boldsymbol{\theta}')}. \tag{3.23}$$

Now the augmented posterior distribution is added with two new variables, $\{\boldsymbol{\theta}', \mathbf{y}\}$, with $\mathbf{y} \sim f(\cdot|\boldsymbol{\theta}')$. Again, marginalizing over $\{\boldsymbol{\theta}', \mathbf{y}\}$ will return it back to the original posterior distribution. Now at each MH step, the algorithm assumes that $\mathbf{x} \sim f(\cdot|\boldsymbol{\theta})$. The chain is then updated through a swapping proposal

$$q_s(\{\boldsymbol{\theta}'^*, \boldsymbol{\theta}^*\}|\{\boldsymbol{\theta}', \boldsymbol{\theta}\}) = \delta(\boldsymbol{\theta}'^* - \boldsymbol{\theta})\delta(\boldsymbol{\theta}^* - \boldsymbol{\theta}'), \tag{3.24}$$

where $\delta(\cdot)$ is the Dirac delta function. Note that this proposal is inherently symmetric, therefore, the MH ratio is simply computed as the ratio of the augmented posterior

distribution between the swapped settings:

$$r = \frac{\cancel{q_s(\{\boldsymbol{\theta}', \boldsymbol{\theta}\}|\{\boldsymbol{\theta}'^*, \boldsymbol{\theta}^*\})} h(\mathbf{x}|\boldsymbol{\theta}')h(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\cancel{q_s(\{\boldsymbol{\theta}'^*, \boldsymbol{\theta}^*\}|\{\boldsymbol{\theta}', \boldsymbol{\theta}\})} h(\mathbf{x}|\boldsymbol{\theta})h(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}. \tag{3.25}$$

A pseudo-code for the exchange algorithm is given in Algorithm 3.

Algorithm 3: Exchange Algorithm

Input: Initial $\boldsymbol{\theta}$, number of iterations T;
for $t = 1, ..., T$ **do**
  Propose $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})$;
  Generate auxiliary variable $\mathbf{y} \sim h(\cdot|\boldsymbol{\theta}')/\mathcal{Z}(\boldsymbol{\theta}')$;
  Calculate $r = \dfrac{h(\mathbf{x}|\boldsymbol{\theta}')h(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{h(\mathbf{x}|\boldsymbol{\theta})h(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}$;
  Accept $\boldsymbol{\theta}'$ with probability $a = \min(1, r)$;
end

The exchange algorithm is also a valid MCMC algorithm but it is much simpler and easier to implement than the SAVM algorithm. A more heuristic understanding of the algorithm is that it is comparing the preference of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ towards the real data $\mathbf{x}$ and the auxiliary variable $\mathbf{y}$ (Murray, Ghahramani, and MacKay, 2006; Park and Haran, 2018). At each update, the chain would propose to the current state $\boldsymbol{\theta}$ a new $\boldsymbol{\theta}'$ to give up the data $\mathbf{x}$ and move to the proposal. If $h(\mathbf{x}|\boldsymbol{\theta}')/h(\mathbf{x}|\boldsymbol{\theta}) > 1$, then this indicates that the proposal $\boldsymbol{\theta}'$ fits the data $\mathbf{x}$ better than $\boldsymbol{\theta}$. We also need to consider the other side of the story and see which of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ fits the auxiliary variable $\mathbf{y}$ better. Hence, the ratio $h(\mathbf{y}|\boldsymbol{\theta})/h(\mathbf{y}|\boldsymbol{\theta}')$ represents this preference.

Despite the simplicity, the exchange algorithm still poses a serious implementation problem in that the generation of $\mathbf{y}$ requires exact/perfect sampling to ensure the algorithm is asymptotically exact. Now this requirement is highly restrictive since it is extremely difficult or almost impossible to generate data that perfectly follows $f(\cdot|\boldsymbol{\theta}')$, especially if $f(\cdot|\boldsymbol{\theta}')$ defines a sophisticated model (Liang, 2010; Park and Haran, 2018).

Liang (2010) proposed a double Metropolis-Hasting (DMH) algorithm based on the exchange algorithm by simulating the auxiliary variable through a standard MCMC algorithm to relax the perfect sampling restriction so that the computation becomes feasible. Algorithm 4 provides the pseudo-code for DMH algorithm.

Algorithm 4: Double Metropolis-Hasting (DMH) Algorithm

Input: Initial $\boldsymbol{\theta}$, number of iterations T, number of iterations M of BDMH algorithm for the auxiliary variable;

for $t = 1, ..., T$ **do**

Propose $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta})$;

Generate auxiliary variable $\mathbf{y} \sim h(\cdot|\boldsymbol{\theta}')/\mathcal{Z}(\boldsymbol{\theta}')$ through M-step BDMH algorithm;

Calculate $r = \dfrac{h(\mathbf{x}|\boldsymbol{\theta}')h(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{h(\mathbf{x}|\boldsymbol{\theta})h(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}$;

Accept $\boldsymbol{\theta}'$ with probability $a = \min(1, r)$;

end

Since there is a standard MH run for the auxiliary variable in each MH update for the parameters, the algorithm is called the double Metropolis-Hasting (DMH) algorithm. This algorithm is the easiest to construct and computationally one of the most feasible among all existing algorithms that deal with doubly intractable distributions (Park and Haran, 2018). However, there is a trade-off as DMH is an asymptotically inexact algorithm due to the imperfect sampling of the auxiliary variable obtained by a MH run and the resulting posterior estimates may be biased. This problem can be mitigated by running the Markov chain for the auxiliary variable long enough at a cost of computational expense. It is recommended to run the Markov chain for $10m \sim 20m$ steps where m is the number of points in the pattern. However, for complex models with a high number of points in the pattern, this is still relatively computationally costly. Nonetheless, modern day high performance computing super-cluster such as Shared Hierarchical Academic Research Computing Network (SHARCNET) can easily carry out threaded parallel computing with thousands of cores, hence, the computation for model inference is feasible.

Adaptive MCMC

To facilitate speedy convergence of MCMC algorithms, a standard and fixed normal proposal distribution is rarely sufficient. For posterior distributions that are relatively high dimensional and exhibit high correlation between components, standard independent proposal from a normal distribution will perform extremely poorly.

A simple and powerful solution to speed up the convergence is the adaptive MCMC scheme (Haario, Saksman, and Tamminen, 2001; Roberts and Rosenthal, 2009; Rosenthal, 2011). The idea is to let the MCMC algorithm self-learn the covariance structure of the posterior distribution and repeatedly update the proposal distribution using the past samples in the chain. Experiments indicates that the Markov chain can indeed learn the structure of the posterior distribution and its performance is much more superior than the standard independent normal proposal distribution (Rosenthal, 2011).

Assuming the proposal distribution comes from the normal distribution family, the simplest construction of adaptive MCMC proceeds as follow: for the first m iteration, to ensure that the chain has a defined covariance matrix, we use a fixed proposal distribution:

$$\boldsymbol{\theta}'|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{D})$$

where $\mathbf{D}$ is a user-defined covariance matrix. Usually, $\mathbf{D}$ is a diagonal matrix with small diagonal components. After the initial m iteration, we let the proposal distribution to be the following:

$$\boldsymbol{\theta}'|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_n + \mathbf{D})$$

where Σ_n is the covariance matrix of the first n ($n \geq m$) samples in the chain. the extra component of $\mathbf{D}$ is to ensure the covariance matrix is invertible. There are various scaling mechanisms to optimize the choice of Σ_n and $\mathbf{D}$, but as it is not the main purpose of this research, I will leave them out and readers can refer to Roberts

and Rosenthal (2009) and Rosenthal (2011) for details.

One important thing to note, however, is that the sampler in adaptive MCMC is no longer a Markov chain since the chain employs information from all previous states of the chain. This may causes the chain to not have the posterior distribution as its stationary distribution. However, as noted in Roberts and Rosenthal (2007), as long as the chain satisfies the *Diminishing Adaptation* condition and *Bounded Convergence* condition, the chain will still have the posterior distribution as the stationary distribution.

Bounded convergence condition is a technical condition satisfied by almost all reasonable adaptive schemes (Rosenthal, 2011), which includes the adaptive scheme where the estimated covariance is employed. The diminishing adaptation condition is a little bit more tricky in that it requires the information change in the adaptation goes to 0 as $n \to \infty$ (Rosenthal, 2011). Fortunately, this is satisfied by the adaptive scheme using the covariance matrix of all past samples as the information change in the this scheme is $\mathcal{O}(1/n)$ since it is an empirical average of the past samples and it goes to 0 as $n \to \infty$ (Rosenthal, 2011).

3.5 Model Criticism

In this section, I introduce the methodology used for model criticism of GPP models. As mentioned in Chapter 2, the most basic validation tools include the empirical PCF if the correlation summary is used, while NND distribution is generally employed if the spacing summary is concerned. However, difficulties arise when the point pattern exhibits heterogeneous trend and interpoint interactions which are essentially what GPP models try to capture simultaneously. One naive strategy would be comparing the summary statistics, such as PCF, of real data to that of the simulated data after obtaining the fitted model parameters. However, these only serve as a second order summary statistics of the model and it does not provide information on model fit in

terms of general intensity of the model.

To provide a well-rounded tool set of model criticism for GPP models, I introduce the methodologies developed by Baddeley et al. (2005) which enables systematic diagnostic of intensity for GPP models fitted by arbitrary inference algorithms. It concerns the residuals analysis for spatial point processes which I will define below.

3.5.1 GNZ Formula

A very important characterization of GPP models which is useful for model diagnostic is given by the Georgii-Nguyen-Zessin (GNZ) formula (Georgii, 1976; Xanh and Zessin, 1979; Baddeley et al., 2005). It is a crucial equation that characterizes the integral properties of the GPP models which is highly useful for various purposes as we will see in the next section.

Theorem 3.5.1. *(GNZ formula) Suppose a GPP* $\mathbf{X}$ *is finite and hereditary defined on a compact set* S*. Let* $\lambda(\cdot,\cdot)$ *denote its conditional intensity with respect to a unit rate Poisson process. Then for any function* $g : S \times \mathcal{N}^f \to \mathbb{R}^+$*, the following equation holds:*

$$\mathbb{E}\left[\sum_{x\in\mathbf{X}} g(x, \mathbf{X}\setminus\{x\})\right] = \int_S \mathbb{E}[g(\xi,\mathbf{X})\lambda(\xi,\mathbf{X})]d\xi. \tag{3.26}$$

An important identity we can obtain from equation (3.5) is the following: let $g(x, \mathbf{X}\setminus\{x\}) = \mathbf{1}[x \in A]$ for some $A \subset S$, we have

$$\mathbb{E}[n(\mathbf{X}\cap A)] = \int_A \mathbb{E}[\lambda(\xi,\mathbf{X})]d\xi. \tag{3.27}$$

Equation 3.27 provides important theoretical groundwork for conducting model diagnostic for GPP models since the equation holds for arbitrary subset A of S. This means this equation can be used for checking the fit of intensity of the model. This is because the left hand side of the equation can be thought of as the intensity of the data in A while the right hand side is the intensity in A obtained from the model.

Significant difference between these two terms will then indicate there is a lack of fit on the model. I will discuss the details on model diagnostic in the next section.

3.5.2 Residuals of Point Processes

Suppose we employ a parametric model for a spatial point process $\boldsymbol{X}$ defined by a probability distribution $f_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta}$ is the parameter vector. Further assume that $f_{\boldsymbol{\theta}}$ satisfies the hereditary condition. Then the innovation process (Baddeley et al., 2005) of $f_{\boldsymbol{\theta}}$ is defined as

$$I_{\boldsymbol{\theta}}(A) = n(\boldsymbol{X} \cap A) - \int_A \lambda_{\boldsymbol{\theta}}(\xi, \boldsymbol{X}) d\xi \tag{3.28}$$

for any $A \subset W$. $\lambda_{\boldsymbol{\theta}}(\xi, \boldsymbol{X})$ is the Papangelou conditional intensity defined in the previous section. Furthermore, we assume that

$$\lambda_{\boldsymbol{\theta}}(\xi, \boldsymbol{X}) = \frac{f_{\boldsymbol{\theta}}(\boldsymbol{X} \cup \{\xi\})}{f_{\boldsymbol{\theta}}(\boldsymbol{X})}$$

if $\xi \notin \boldsymbol{X}$ while

$$\lambda_{\boldsymbol{\theta}}(\xi, \boldsymbol{X}) = \lambda_{\boldsymbol{\theta}}(\xi, \boldsymbol{X} \setminus \{\xi\})$$

if $\xi \in \boldsymbol{X}$. Innovation process is essentially the discrepancy between the intensity of $\boldsymbol{X}$ and that of the model. If the model is "correct", then the innovation process will be a spatial white noise process and it is analogous to errors in simple linear models (Baddeley et al., 2005). The estimator for the innovation process is the raw residual process (Baddeley et al., 2005). It is obtained by using a plug-in estimator of the conditional intensity, i.e.,

$$\widehat{\lambda_{\boldsymbol{\theta}}}(\xi, \mathbf{x}) = \lambda_{\hat{\boldsymbol{\theta}}}(\xi, \mathbf{x})$$

where $\hat{\boldsymbol{\theta}}$ is the estimated model parameters. The raw residual process is then given as

$$R_{\hat{\boldsymbol{\theta}}}(A) = n(\mathbf{x} \cap A) - \int_A \lambda_{\hat{\boldsymbol{\theta}}}(\xi, \mathbf{x}) d\xi. \tag{3.29}$$

Raw residual process is then analogous to residuals in a linear model, and it can serve as a diagnostic for analyzing the fit of spatial trend of the model. Any deviance from white noise will then indicate a lack of fit in the intensity. (Baddeley et al., 2005) also considered various scaled residuals such as inverse residuals and Pearson residuals. They did this by employing the GNZ formula through different functional forms of $g(\cdot, \cdot)$ in equation 3.26. However, the models constructed in this research is not suitable for the employment of scaled residuals mentioned above due to hard-core component in the interaction term. Furthermore, due to the Bayesian paradigm adopted in this research, it is not clear how a scaled residuals should be computed through the posterior predictive simulation. Therefore, we only consider using the raw residuals.

3.5.3 Computation of Residuals

To compute the residuals, a naive approach suggested by Baddeley et al. (2005) is to divide the observation window W into m rectangular regions, $A_1, \ldots, A_m$ and calculate $R_{\hat{\boldsymbol{\theta}}}(A_k, h, \lambda)$ for each $k = 1, \ldots, m$. A better approach is to compute a smoothed version of residuals using a kernel $k(\cdot)$ to obtain a smoothed residual field (Baddeley et al., 2005). The smoothed residual field at location ξ is then

$$s(\xi) = e(\xi) \left[\sum_{x_i \in \mathbf{x}} k(\xi - x_i) h_{\hat{\boldsymbol{\theta}}}(x_i, \mathbf{x} \setminus \{x_i\}) - \int_W k(\xi - \eta) h_{\hat{\boldsymbol{\theta}}}(\eta, \mathbf{x}) \lambda_{\hat{\boldsymbol{\theta}}}(\xi, \mathbf{x}) d\eta \right] \tag{3.30}$$

where $e(\xi)$ is used for edge correction where $e(\xi)^{-1} = \int_W k(\xi - \eta) d\eta$. Simply eye-balling the behavior of the obtained smoothed residual fields will provide information on the fit of the model intensity, i.e., negative residuals indicate underestimation in

the intensity and vice versa.

3.5.4 Residual Computation under the Bayesian Paradigm

The above residual analysis techniques is proposed under the MLE paradigm, and there exists slight differences in the implementation of the residual analysis under the Bayesian paradigm.

As noted in Leininger and Gelfand (2017), the Bayesian residual analysis of point patterns should consider the posterior predictive simulation of the point pattern rather than directly compute the conditional intensity using the fitted model parameters as in the MLE approach.

To assess the residuals in region A_k, one needs to consider the following as an estimator for the residuals:

$$N(A_k) - N_{sim}(A_k) \tag{3.31}$$

where $N(A_k)$ is the number of points in A_k in data and $N_{sim}(A_k)$ is the number of points in A_k through posterior predictive simulation.

To obtain a smoothed residual field, one would first draw n samples from the posterior distribution, and simulate point pattern $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ by using the selected posterior sample $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_n$. Then we need to compute the smoothed residual field using certain kernel density for each $i = 1, ..., n$ and consider their average. The resulted average then serves as an estimated Bayesian residual field.

Chapter 4

Gibbs Point Process Models for Objects in the Star Formation Complexes of M33

In this chapter, I will focus on the construction of GPP models for objects in the star formation complexes of M33. The objects concerned are the carbon monoxide (CO) filament structure, giant molecular clouds (GMCs), and young star cluster candidates (YSCCs).

4.1 Preliminary

It is important to note that CO is not the main ingredient for star formation. Rather it is molecular hydrogen, H_2, which serves as the main source for star formation. However, detecting H_2 from GMCs in extragalactic environments is not possible as H_2 is too cold ($10 \sim 20$ K) for detection. But CO is easily excited and can be used as a tracer for H_2. A general assumption is made that the X-factor (which is the ratio, H_2/CO) is constant and CO can hence be used as a proxy for H_2. Furthermore, CO traced H_2 generally forms in filamentary structures. Due to the uneven distribution

of gas and dust in these filament structures, clumps of gas and dust start to coalesce and once the clump reaches enough mass, it becomes a GMC and star formation will then commence. For a detailed review of formation, structure, detection, and its role played in star formation, see McKee and Ostriker (2007).

I choose M33 for our analysis since it is one of the few low-inclination galaxies with a relatively complete catalog of GMCs. Three sets of data are used in the analysis, on the CO filament structure, the GMCs, and the YSCCs. The CO filament data and GMCs data are obtained from IRAM 30-m observations of CO(2-1) emission published in Druard et al. (2014). The CO filamentary structure is extracted from the CO emission map[1] using the method described in Koch and Rosolowsky (2015). The GMCs are also identified by Corbelli et al. (2017) using the IRAM 30-m observations of CO(2-1) emission by Druard et al. (2014) and the YSCCs are identified using the Spitzer 24-μm data, published by Sharma et al. (2011) and Corbelli et al. (2017). The data consist of the positions, galactocentric distance, effective radius, velocity dispersion, gas mass, and virial mass of 566 identified GMCs and the positions, size, and incomplete estimates of age and mass of 630 identified YSCCs. Both confirmed and candidate young stellar clusters (YSCs) are considered since there are only around 400 confirmed YSCs (with estimation of mass and age). Furthermore, the 630 candidate YSCs are what was analysed in Corbelli et al. (2017) and it is appropriate to also use the candidates catalog for drawing comparison.

Figure 4.1 and 4.2 show the overlay plots of GMCs on the CO filament structures and YSCCs on GMCs. Note that the coordinates of the objects are transformed from astronomical right ascension and declination (α, δ: longitude and latitude equivalent) to two-dimensional projected M33-centric Cartesian coordinates, accounting for the inclination of M33 with respect to the line of sight. The inclination is set as $i = 53^{\circ}$ (Magrini, Stanghellini, and Villaver, 2009), the distance to M33 is set as $D =$

[1]With permission from Eric Koch and Erik Rosolowsky

840 kpc (Bonanos et al., 2006; Magrini, Stanghellini, and Villaver, 2009) and the position angle (PA) of the major axis is $\theta = 22^\circ$ (Magrini, Stanghellini, and Villaver, 2009). Assuming the equatorial coordinates of a source is (α, δ), the procedure for transformation is then (Cioni, 2009):

- convert (α, δ) to angular coordinates (x, y);
- rotate the coordinate through

$$x_1 = x\sin(\theta) - y\cos(\theta) \tag{4.1}$$

$$y_1 = y\sin(\theta) + x\cos(\theta) \tag{4.2}$$

- deproject:

$$y_2 = y_1/\cos(i) \tag{4.3}$$

- calculate angular distance and convert into kpc through

$$d_{\rm ang} = \sqrt{x_1^2 + y_2^2} \tag{4.4}$$

$$d_{\rm kpc} = D\tan(d_{\rm ang}) \tag{4.5}$$

- obtain the 2D projected coordinate as

$$x^* = d_{\rm kpc}\frac{x_1}{d_{\rm ang}} \tag{4.6}$$

$$y^* = d_{\rm kpc}\frac{y_2}{d_{\rm ang}} \tag{4.7}$$

From Figure 4.1 and 4.2, it is astonishing how the CO filament structure, GMCs and YSCCs are strongly correlated with each other. However, a sensitive quantitative investigation is needed to describe the spatial distribution for deriving further physical

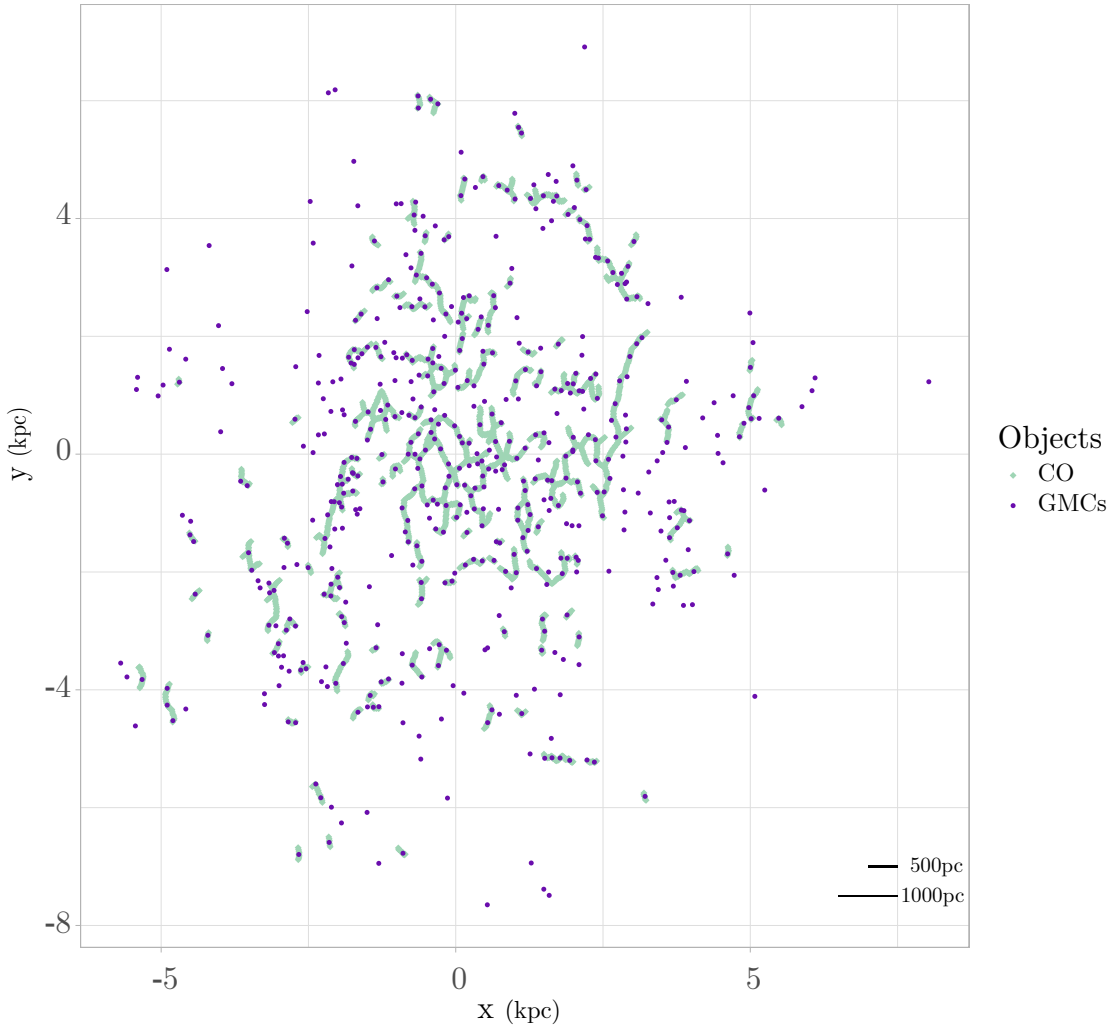


Figure 4.1: Overlay plot of the CO filament structure and GMCs. Green network is the CO filament structure. Purple dots are GMCs. It is striking how the distribution of GMCs follows the network of CO filament structure.

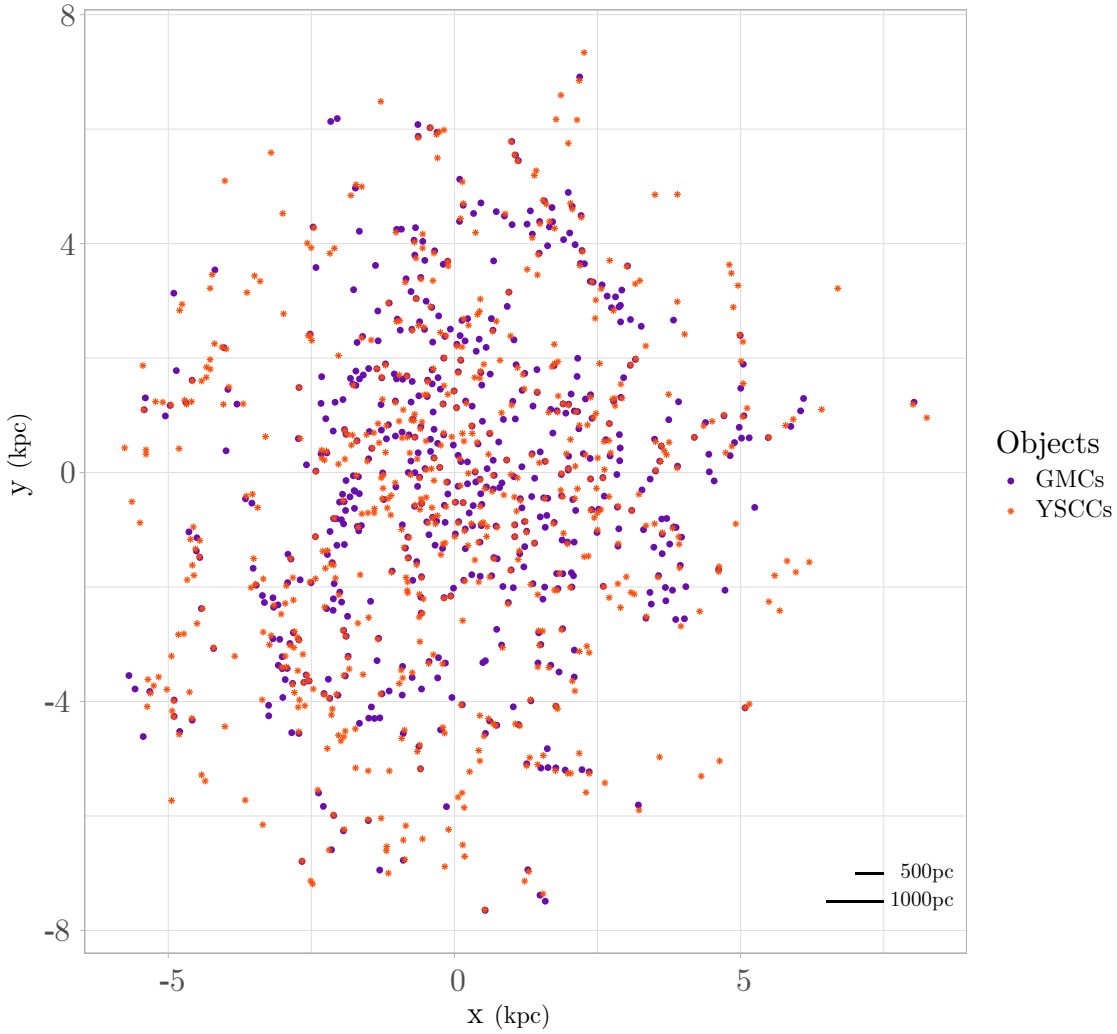


Figure 4.2: Overlay plot of GMCs and YSCCs. Red stars are YSCCs while purple dots are GMCs. It is clear that there is a significant positive correlation between GMCs and YSCCs.

implications.

4.2 Model for CO Filaments and GMCs

In this section, I introduce the GPP model that models the distribution of GMCs in M33 (CO-GMC model). From a modelling standpoint, the CO-GMC model is the high-level process in a hierarchical framework between GMC and YSCCs. On another note, it can also be regarded as a "low-level" process in that it is the lower-level process in a hierarchical framework between the CO filament structure and GMCs. However, the CO filament structure is not the modelling focus here.

Since the CO filament structure can also be regarded as a high-level process for the GMCs, an interpretation for the CO filament structure can be that it is an underlying spatial covariate for the GMCs. From a physical point of view, it has the foundation that GMCs are generally believed to have formed from these filament structures, i.e., interstellar medium (ISM) rich with gas and dust to fuel star formation. This means the first order intensity of the CO-GMC model is dependent on the CO filament structure.

Note that for simplicity, only the CO filament structure is considered to contribute to the first order intensity of GMCs. There may be other underlying variables that affect the first order intensity, such as the large scale intensity variation of matter in the galaxy. However, from Figure 4.1, the association of GMCs and CO filament structure is striking, with only a few GMCs scattered around the filament structure. It is, therefore, reasonable to assume a first order intensity dependent only on the CO filament structure.

Denote the pattern of the CO filament structure by $\mathcal{L}$ and denote the pattern of

GMCs as $\mathbf{x}_G$. Then the likelihood function of the CO-GMC model is given as

$$\boldsymbol{\ell}(\boldsymbol{\theta}_G|\mathbf{x}_G;\mathcal{L}) = f(\mathbf{x}_G|\boldsymbol{\theta}_G,\mathcal{L}) \propto \prod_{i=1}^{n(\mathbf{x}_G)} \lambda_{\mathcal{L}}(x_{i,G})\phi_G(\mathbf{x}_G), \tag{4.8}$$

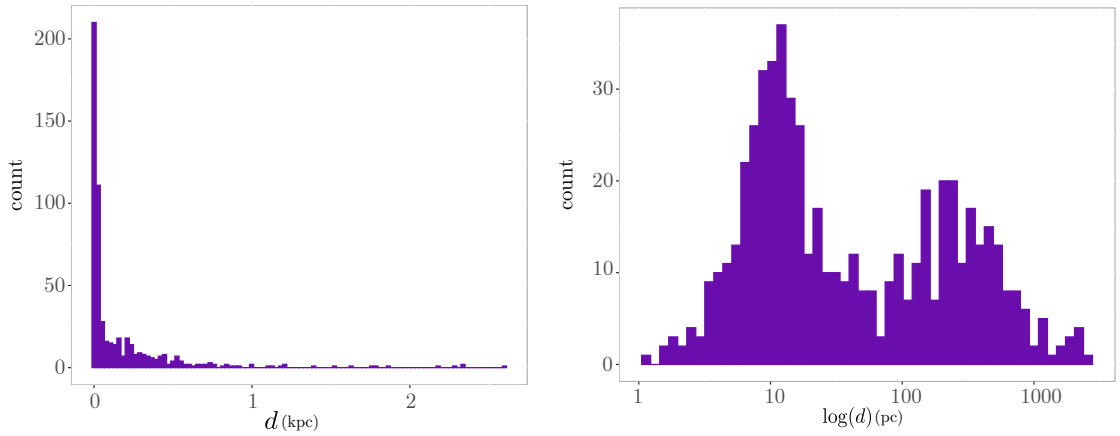
where $\lambda_{\mathcal{L}}(x_{i,G})$ is the first order intensity at the location of the i-th GMC, which also depends on the CO filament structure. $\phi_G(\mathbf{x}_G)$ is the pairwise interaction term among GMCs.

To specify a model for the first order intensity of GMCs that depends on the CO filament structure, a natural procedure is to consider the distance from a GMC to its closest point on the CO filament structure, i.e., the nearest neighbor distance (NND). Figure 4.3 shows the histograms of the NND from GMCs to the CO filament in both the normal scale (Figure 4.3 (a)) and the log-scale (Figure 4.3 (b)). It seems that the NND distribution follows a simple power law from Figure 4.3 (a). However, after transformation to the log-scale as shown in Figure 4.3 (b), the NND distribution in fact has two sub-populations, with a major population being extremely close to the filament structure and a minor population being relatively far away. Modelling the first order intensity through a simple power law structure with respect to the NND is not able to capture the sub-population features.

To capture the multi-modal form of intensity variation, for simplicity I choose the following formulation for the first order intensity:

$$\lambda_{\mathcal{L}}(x_{i,G}) = \lambda(d_i) = \begin{cases} \theta\left(1+\dfrac{d_i}{\sigma_0}\right)^{-\alpha}, & 0 < d_i \le R_c, \\ \beta\left(1+\dfrac{(d_i-R_c)^2}{\sigma^2}\right)^{-1}, & d_i > R_c. \end{cases} \tag{4.9}$$

Note that in this model, $\theta > 0$, $\alpha > 0$, and $\sigma_0,\ \sigma > 0$. d_i is the NND from the i-th GMC to the CO filament. θ controls the strength of the first order intensity, σ_0 controls the characteristic scale at which the major sub-populations of GMCs



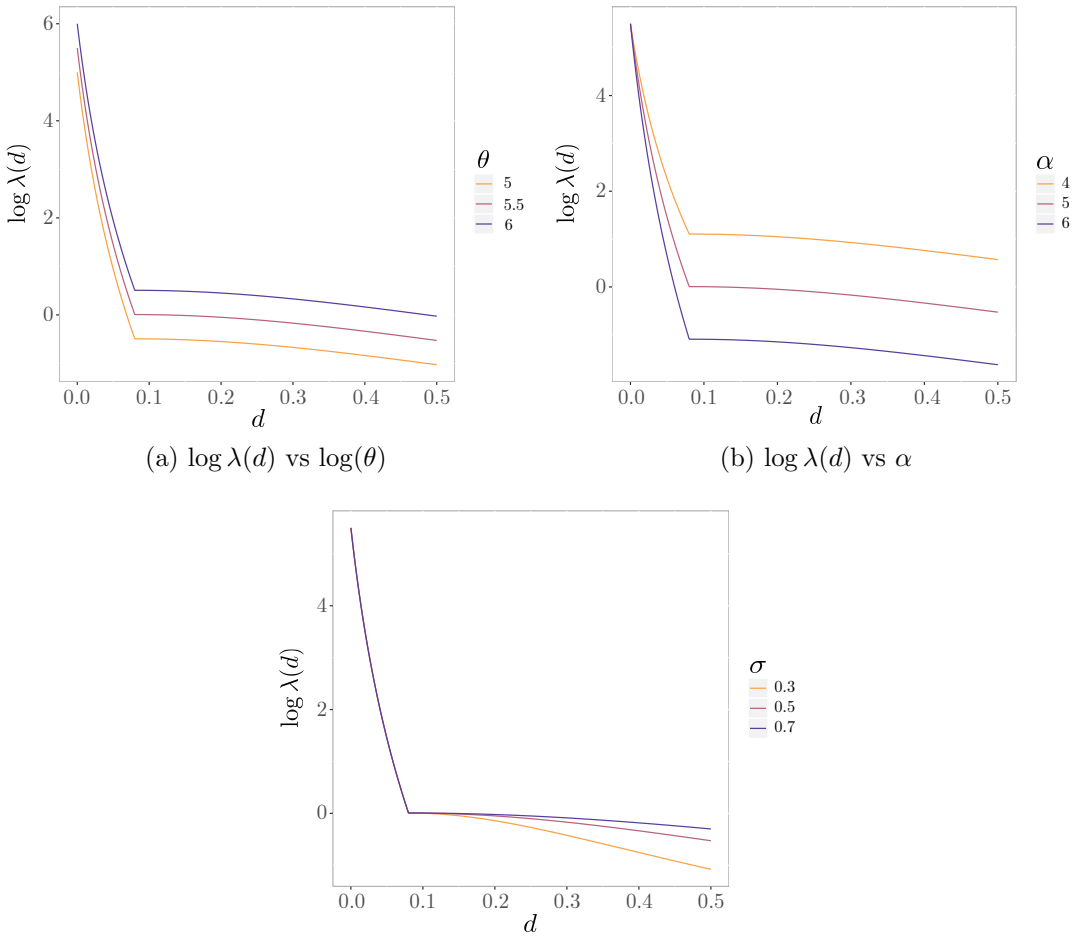
(a) Histogram of NND from GMC to CO filament (b) Histogram of log-NND from GMC to CO filament

Figure 4.3: Histogram of nearest neighbor distance (NND) from GMC to the CO filament: (a) histogram with unscaled NND; (b) histogram with log-transformed NND. It is not so obvious in (a) that the NND distribution has two sub-populations but it becomes apparent in (b) when NND is shown in log-scale.

distribute around the CO filament structure. α is the power law coefficient governing the distribution of the major sub-population of GMCs. β is chosen so that $\lambda(\cdot)$ is continuous at $R_c > 0$. R_c is the cutoff boundary for the two populations. For simplicity, R_c is determined by visually inspecting the NND histogram and found to be approximately 84 pc. Furthermore, directly fitting this parameter can lead to potential numerical issues due to the piecewise structure of the intensity function. σ is the characteristic scale controlling the distribution of the minor sub-population of GMCs.

The effects of the parameters are visually demonstrated in Figure 4.4. I set $\theta = \exp(5.5)$, $\alpha = 5$, $R_c = 2\sigma_0 = 0.08$, and $\sigma = 0.5$ as the reference parameters and see how $\lambda(\cdot)$ is affected by the change in these parameters. For better visualization, I transform the function value to log-scale. From Figure 4.4(a), θ governs the general magnitude of the strength of intensity as across all values of NND. In a sense, it controls, on average, how many GMCs there are in the span of the galaxy. Figure 4.4(b) shows that α controls the rate at which the intensity of GMCs decays in the close vicinity of the CO filament structure. Figure 4.4(c) shows that σ determines the

asymptotic behavior of $\lambda(d)$ or the rate of intensity decay for the minor sub-population of GMCs.



(a) $\log \lambda(d)$ vs $\log(\theta)$

(b) $\log \lambda(d)$ vs α

(c) $\log \lambda(d)$ vs σ

Figure 4.4: Parameter effects on $\lambda(d)$: (a) θ controls the overall intensity of GMCs across galaxy disk; (b) α controls the intensity decrease of the major population as a function of distance; (c) σ controls the asymptotic behavior of intensity.

For the pairwise interaction term among GMCs, I employ a simple modified very soft-core process specified as below

$$\phi_G(x_i, x_j) = \begin{cases} 0, \ 0 < d_{ij} \leq R_G, \\ 1 - \exp\left(-\dfrac{(d_{ij} - R_G)^2}{\delta^2}\right), \ d_{ij} > R_G. \end{cases} \tag{4.10}$$

d_{ij} is the distance between the i-th and j-th GMCs. R_G is the smallest distance be-

Table 4.1: Model parameters for CO-GMC model

Parameters	Meaning	Domain
θ	Overall intensity of GMCs across galaxy disk	$(0, \infty)$
α	Power law governing the decay of GMC intensity in CO vicinity	$(0, \infty)$
σ_0	Characteristic scale of GMC distribution near CO	$(0, \infty)$
σ	Scale parameter controlling asymptotic intensity of GMCs	$(0, \infty)$
δ	Scale parameter controlling repulsive scale of second order interaction of GMCs	$(0, \infty)$

tween any two GMCs. Adding this modification to the very soft-core process prevents over-clustering near the CO filament structure. Furthermore, GMCs generally have physical sizes which are denoted by R_G. The use of R_G also embodies the physical reality that two GMCs tend to separate from each other due to gravitational collapse. Chevance et al. (2019) also found that in nine nearby spiral galaxies, the mean separation distance between star formation complexes, i.e., GMCs with star formation activities, is approximately 100–300 pc. This justifies the short range repulsive structure among GMCs. $\delta > 0$ determines the range of the repulsive scale. Now at greater pairwise distance, this interaction term essentially behaves like a Poisson process. Although this might not be the true spatial distribution of GMCs, we can obtain information on the behavior of GMCs through model criticism.

Table 4.1 provides a summary of the model parameters for reference.

4.3 Model for GMCs and YSCCs

In this section, I introduce a new model to probe the distribution of YSCCs assuming a hierarchical structure from GMCs to YSCCs, while simultaneously accounting for the large scale variation of the intensity of YSCCs across the galaxy disk as well as the effect of properties of GMCs on the distribution of YSCCs.

Under the hierarchical GPP model framework, the point pattern of GMCs is treated as given. Denoting the point pattern of GMCs as $\mathbf{x}_G$ and the point pattern of YSCCs as $\mathbf{x}_S$, the general form of the likelihood function then follows from Chapter

3:

$$\boldsymbol{\ell}(\boldsymbol{\theta}_S|\mathbf{x}_S;\mathbf{x}_G) = f_{\mathbf{x}_G}(\mathbf{x}_S;\ \boldsymbol{\theta}_S) = \alpha_S(\mathbf{x}_G)\prod_{j=1}^{n(\mathbf{x}_S)}\lambda_S(x_{j,S})\phi_S(\mathbf{x}_S)\prod_{i=1}^{n(\mathbf{x}_G)}\prod_{j=1}^{n(\mathbf{x}_S)}\phi_{GS}(x_{i,G},x_{j,S}). \tag{4.11}$$

As in Chapter 3, $\boldsymbol{\theta}_S$ is the vector of model parameters. $\lambda_S(x_{j,S})$ is the first order intensity at the location of the j-th YSCC, $\phi_S(\mathbf{x}_S)$ is the pairwise-interaction term for YSCCs, and $\phi_{GS}(x_{i,G},x_{j,S})$ is the correlation term between the i-th GMC and the j-th YSC. $\alpha_S(\mathbf{x}_G)$ is the unknown normalising constant dependent on the parameters and $\mathbf{x}_G$. We now give the parametric structure for each term.

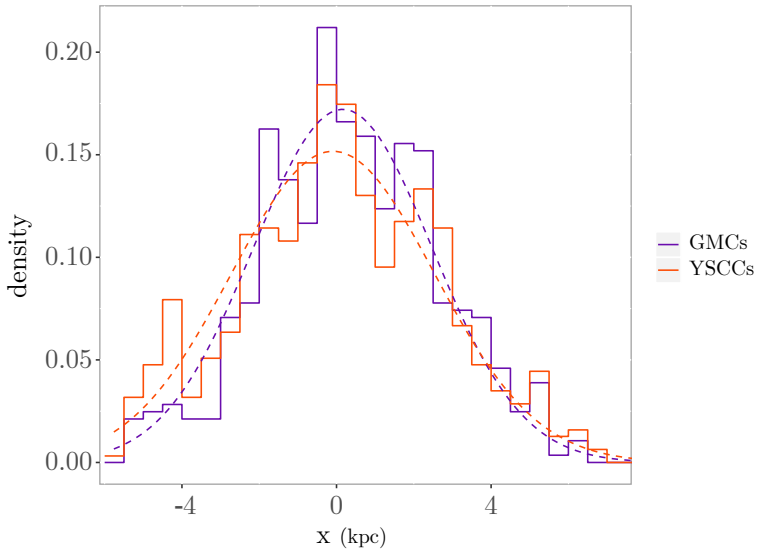
Since the general large scale distributions of GMCs and YSCCs are both approximately normal centred around the galaxy centre as shown in the histograms in Figure 4.5, the overlapping large-scale distribution of GMCs and YSCCs will be a lurking variable that can undermine the investigation of the actual relationship GMCs and YSCCs. Therefore, this will be accounted for in the first-order potential term as a large-scale spatial trend:

$$\lambda_S(x_{j,S}) = \exp\left(P_2(x_{j,S};\boldsymbol{p})\right), \tag{4.12}$$

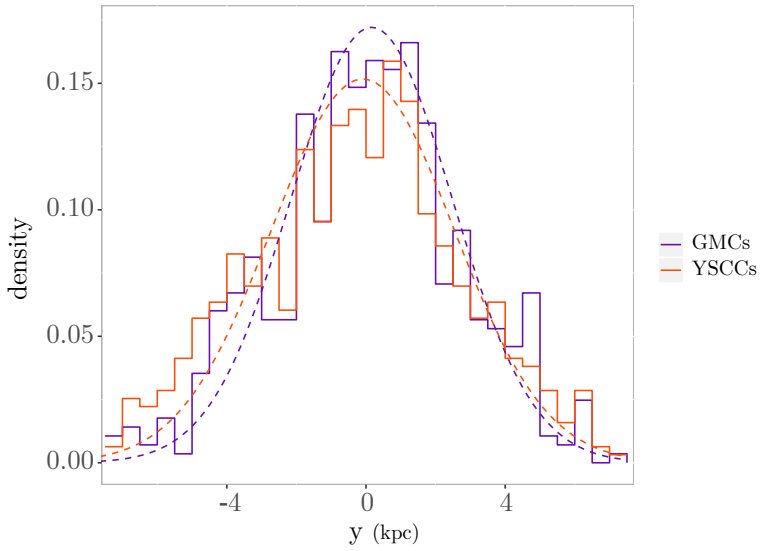
where $P_2(x_{j,S};\boldsymbol{p})$ is a second-order polynomial in terms of the distance from the j^{th} YSCC to the galactic centre. To make the model as simple as possible, we assume the following form for $P_2(\cdot;\cdot)$:

$$P_2(x_S;\rho,R_{s,c}) = -\left(\frac{r_{s,c}}{R_{s,c}}\right)^2 + \rho, \tag{4.13}$$

where $r_{s,c}$ is the distance from YSCC x_S, to the galaxy centre. $R_{s,c}$ is the characteristic scale of the distribution of YSCCs in the galaxy disc, and ρ is an offset parameter controlling the large scale intensity. For the correlation between the GMCs and



(a) x-coordinate



(b) y-coordinate

Figure 4.5: (a) Histogram of the x-coordinates of GMCs and YSCCs; (b) Histogram of the y-coordinates of GMCs and YSCCs. Purple solid lines are histograms for GMCs while red solid lines are histograms for YSCCs. Purple dashed lines are fitted Gaussian density for GMCs and red dashed lines are fitted Gaussian density for YSCCs. We can see that both GMCs and YSCCs are generally Gaussian distributed with centers at the galaxy center and both distributions overlap significantly.

YSCCs, we choose the following parametric form:

$$\phi_{GS}(x_{i,G}, x_{j,S}) = \exp\left[\psi_i \left(1 + \frac{r_{ij}^2}{\sigma_{GS}^2}\right)^{-\frac{5}{2}}\right]. \tag{4.14}$$

In this model, ψ_i controls the correlation strength between the i-th GMC and all YSCCs. The greater the value of ψ_i, the greater the correlation between GMCs and YSCCs. r_{ij} is the distance between the i^{th} GMC and the j^{th} YSC. σ_{GS} is a characteristic scale parameter controlling the correlation scale between GMCs and YSCCs. The notion of correlation here is in terms of both distance and number since it is a smoothly decaying function with respect to the inter-type distance r_{ij}. Notice that if $\psi_i = 0$, it then suggests that there is no correlation between GMCs and YSCCs. I assume the distribution of YSCCs around each YSCC follows a Plummer (5,2) power law (Plummer, 1911; Dejonghe, 1987) for simplicity. Moreover, a preliminary analysis on the cross-type 2PCF/PCF between GMCs to YSCCs shows a similar power law shape as indicated in Figure 4.6. Note that the scale of the cross-type PCF is in log-scale for better visualization. The computation of cross-type PCF is carried out in the same fashion as in Corbelli et al., 2017 by dividing the the galaxy disc into three zones based on the galactocentric distance (zone 1: $D < 1.5$ kpc; zone 2: 1.5 kpc $\leq D < 4$ kpc; zone 3: $D \geq 4$ kpc). We can see that the correlation scale is generally the same in all zones with zone 3 slightly greater than zone 1 and zone 2. The correlation strength, however, is drastically different, indicating there is a relationship between correlation strength and galactocentric distance. I will incorporate this information in the model later as a mark information on the GMCs. It is also noteworthy that for zone 3, the cross-type PCF is always above unity when distance increases. However, zone 1 and zone 2 approaches unity at approximately the same speed. This indicates that there is indeed first-order inhomogeneity in zone 3 that is not accounted for. A visual investigation of the YSCCs distribution in zone 3 reveals that the first-order

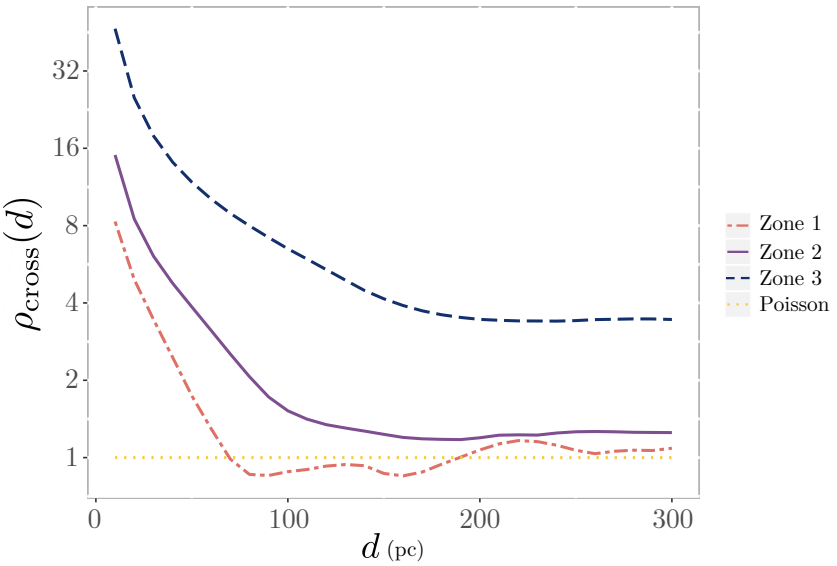


Figure 4.6: Cross-type PCF between GMCs and YSCCs; zone 1: $D < 1.5$ kpc; zone 2: 1.5 kpc $\leq D < 4$ kpc; zone 3: $D \geq 4$ kpc where D is the galactocentric distance. d is the distance between GMCs and YSCCs. The strong positive correlation between GMCs and YSCCs is certain as observed and it increases with respect to D. However, the cross-type PCF for zone 3 consistently being above 1 and showing no sign of decreasing means there exists strong overlap of inhomogeneity in the intensity of GMCs and YSCCs.

intensity is much higher closer to the galaxy center while it drops off drastically as galactocentric distance increases. Therefore, assuming a homogeneous intensity of YSCCs in zone 3 is inappropriate, further justifying the necessity of our approach.

On another note, since we are considering all possible pairings between GMCs and YSCCs, choosing such a formulation circumvents the problems in rudimentary analysis where YSCCs are assigned an associated GMC by nearest neighbour distance. This eliminates the potential bias introduced by wrongful nearest neighbour assignment.

There are certainly various forms of parameterization of the cross-type PCF one can choose beside the Plummer model. One example is the simple power law structure

$$\log(\phi(r)) = \theta \left(\frac{r_c}{r}\right)^{\alpha}, \tag{4.15}$$

which is the one proposed by Peebles (1980) and used by Grasha et al. (2015), Grasha et al. (2017), and Grasha et al. (2019). In this model, θ controls the strength/amplitude of the correlation, r_c is the characteristic scale of the correlation, and α is the governing power law coefficient. Looking at the cross-type PCF in Figure 4.6, the simple power law model seems to fit better than the Plummer model. However, the issue with this model under the GPP framework is that it is highly numerically unstable. Since this model is fitted directly to the empirical PCF in the work of Peebles (1980) and Grasha et al. (2015), Grasha et al. (2017), and Grasha et al. (2019), it does not necessarily pose a computational issue as that work is essentially fitting a regression model. However, under a GPP framework where simulation is required, the distance r in the denominator can severely undermine the computation. In fact, this model is not even bounded, directly violating the GPP model assumption.

The Plummer model was originally conceived to model the distribution of stars in globular clusters. Although YSCCs do not clump around a GMC as stars do around the globular cluster center, we can imagine that the Plummer model is capturing the ensemble distribution of YSCCs around GMCs. The power of the Plummer model can in fact vary. I set it to the simplest (5,2) configuration to reduce computational complexity. Furthermore, a crude estimate based on the empirical cross-type PCF in Figure 4.6 suggests that the (5,2) configuration is reasonable.

For the pairwise-interaction term, assuming stationarity, we employ the following model:

$$\phi_S(d_{ij}) = \begin{cases} 0, \ 0 < d_{ij} \leq R_S, \\ \dfrac{4}{3}\left(\dfrac{d_{ij} - R_S}{\sigma_S}\right)^2 \left(1 - \left(\dfrac{d_{ij} - R_S}{\sqrt{3}\sigma_S}\right)^2\right), R_S < d_{ij} \leq R_P, \\ 1, \ d_{ij} > R_P, \end{cases} \tag{4.16}$$

where $R_P = \sqrt{3/2}\sigma_S + R_S$. d_{ij} is the distance between the i-th YSCC and the j-

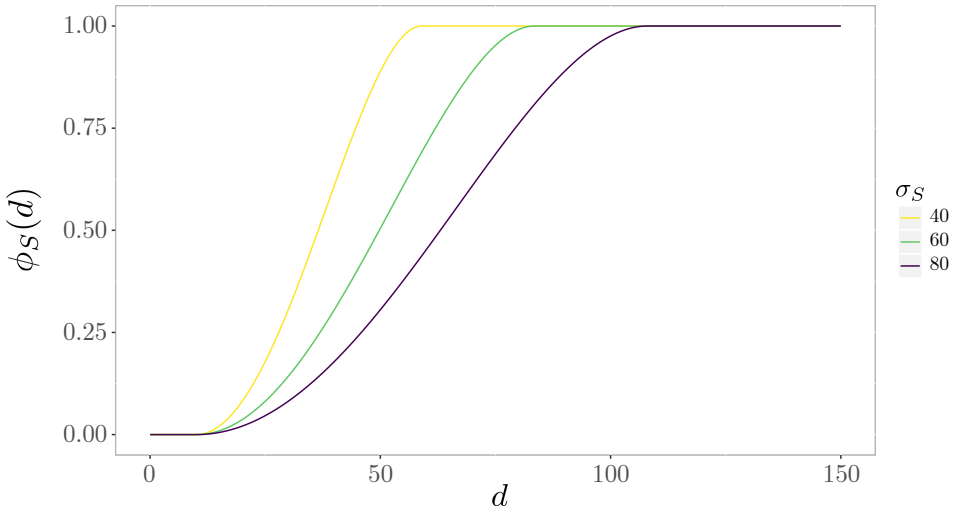


Figure 4.7: Plot of $\phi_S(d)$ with different σ_S values. Increasing value of σ_S increases the repulsive range of second order behavior.

th YSCC. σ_S is a characteristic scale that determines the range of repulsive effect between two YSCCs. However, R_p here is the actual parameter representing the repulsive scale. Figure 4.7 shows the shape $\phi_S(d)$ with different choices of σ_S.

We choose this model since the empirical PCF can no longer be used to determine the actual second-order property for the YSCCs due to the obvious inhomogeneity of the YSCC distribution. The justifications for the choice of this form of pairwise-interaction term are the following: (a) it is easy to implement and has guaranteed numerical stability. Furthermore, the second-order potential is smooth and differentiable at all scales; (b) YSCCs all have physical sizes denoted by R_S. If two YSCCs are at the same location, they will eventually be identified as one YSCC, and as noted, we do not consider cases where there exist coincidental points. Therefore, I incorporate a hard-core component in the pairwise-interaction term; (c) at very short scales, the distribution of YSCCs should be repulsive since there exists competition for the star formation fuel. Furthermore, the stellar feedback can blow away surrounding gas in the molecular clouds and regulate star formation rate (Grasha et al., 2019; Chevance et al., 2019). This is also demonstrated in the simulation by Rogers and Pittard

(2013). The stellar feedback and blowouts by SCs of their surrounding molecular gas in fact corresponds to a form of "competition" for star forming resources. Grasha et al. (2019) also suggests that the formation of SCs from GMCs is sequential, rather than a simultaneous clustering formation. This means that in a small and compact region, it is unlikely for two YSCCs to exist. Although it might happen that two YSCCs can become gravitationally bound with each other and proceed towards a merger, the probability of it happening and being observed should be very small.

One important thing to note is that, for pairwise distance within R_P, it does not mean there cannot be more than one YSCC. It only means that the chance of finding two YSCCs within this distance is less than that of a Poisson process and the chance of this happening goes to zero as the pairwise-distance approaches the hard-core scale R_S. Now at larger scales, the distribution of YSCCs might not be Poisson-like, but we can infer their behaviour at larger scales from model criticism. Any discrepancy between the data and model can be easily interpreted since the model, as a reference, is a Poisson process at the greater range.

4.3.1 Interaction as a Function of Marks

The correlation strength parameter ψ in equation 4.14 is indexed by i to emphasize the dependence on the i-th GMC. It is interesting to see how the properties/marks of GMCs affect their interaction/correlation with YSCCs. The most difficult part in the modelling procedure is to parameterize the interaction as a function of continuous marks as we do not know the shape of the function; we only know its domain and range. Picard et al. (2009) proposed to model the second-order interaction parameter in an area interaction process (Baddeley and Lieshout, 1995) as a sigmoid function of the mark:

$$\gamma(m) = \gamma_0 + \gamma_1 \tanh\left(\frac{m - s}{\delta}\right). \tag{4.17}$$

γ_0 and γ_1 are to be estimated where s and δ are determined by users through summary statistics. The sigmoid function is chosen since it is bounded and for the GPP to be defined, the second-order statistics have to be bounded.

For this study, instead of modelling the second-order interaction parameter, I shift the focus to the cross-type correlation function between GMCs and YSCCs and relate the correlation strength parameters as a function of marks of GMCs. For simplicity, I only assume that the correlation strength parameter, ψ_i, is a function of the marks of GMCs. I do not consider the marks of the YSCCs since various marks of YSCCs are unavailable and in the cases where the marks are available, the estimation is poor. Therefore, conclusions obtained from these estimated marks may not have representative power. It is noteworthy since we are employing a hierarchical point process between GMCs and YSCs, the marks of GMCs are no longer considered marks as they are treated as given. We here abuse the terminology of marks for the sake of simplicity.

Since the functional relationships between the parameters and the marks are generally unknown, I employ a simple linear relationship between the correlation strength parameter and the marks:

$$\psi_i = \theta_0 + \sum_{j=1}^{M} m_{i,j}\theta_j, \ i = 1, ..., n(\mathbf{x}_G), \tag{4.18}$$

where M is the number of marks of GMCs. $m_{i,j}$ denotes the j-th mark value of the i-th GMC. Note that, similarly to a simple linear model, θ_0 represents the baseline value of the correlation strength. θ_j denotes the effect of the j-th mark of a GMC. If $\theta_j = 0$, then the corresponding mark has no effect on the correlation strength. If $\theta_j < 0$, then the j-th mark will have a negative effect on the correlation strength between GMCs and YSCCs and vice versa.

The marks being considered here include (1) the galactocentric distance of GMC,

Table 4.2: Model parameters for GMC-SC model

Parameters	Meaning	Domain
$R_{s,c}$ (kpc)	Characteristic scale of the large scale variation of YSCCs across the galaxy disc	$(0, \infty)$
ρ	Log-intensity of YSCCs at the centre of the galaxy	$\mathbb{R}$
θ_0	Baseline correlation strength between GMCs and YSCCs	$\mathbb{R}$
θ_D	Effect of galactocentric distance of GMCs on correlation strength between GMCs and YSCCs	$\mathbb{R}$
θ_M	Effect of mass of GMCs on correlation strength between GMCs and YSCCs	$\mathbb{R}$
θ_{gc}	Effect of distance from GMCs to CO filament on correlation strength between GMCs and YSCCs	$\mathbb{R}$
σ_{GS} (pc)	Characteristic scale of correlation between GMCs and YSCCs	$(0, \infty)$
σ_S (pc)	Characteristic scale of repulsive structure among YSCCs	$(0, \infty)$

D, which is already shown in Figure 4.6 to have an effect on the correlation; (2) the log-mass of a GMC, $\log_{10}(M/M_\odot)$; (3) the log-NN distance from a GMC to the CO filament, $\log_{10}(d_{gc})$. The correlation strength parameter is then the following:

$$\psi_i = \theta_0 + \theta_D D_i + \theta_M \log_{10}(M_i/M_\odot) + \theta_{gc} \log_{10}(d_{i,gc}). \tag{4.19}$$

Note that when conducting model fitting, the marks are standardized for better comparison between the effects of different properties on the correlation strength. In this case, the baseline θ_0 also represents the average correlation strength of a randomly chosen GMC with YSCCs.

Note that we do not need to ensure the functional forms of the marks are bounded in our case since (a) the marks themselves are bounded. No marks can reach a value of infinity; (b) the hierarchical assumption will treat the GMCs as fixed, therefore, the correlation strength can be any finite real value and the model would still be well-defined. To summarize, Table 4.2 gives a summary of the parameters of the GMC-SC model for reference.

4.4 Analysis of Simulated Data

Before conducting data analysis on the real data, we need to confirm that the DMH algorithm can indeed recover the information from the data through the constructed model. This is done through conducting inference on simulated data. I consider ten

Table 4.3: Chosen parameters for CO-GMC model simulation

Sets	$\log(\theta)$	α	σ (pc)	δ (pc)
1	5	4.5	300	100
2	6	4	200	150
3	6.5	5	250	200

Table 4.4: Chosen parameters for GMC-SC model simulation

Sets	$R_{s,c}$ (kpc)	ρ	θ_0	θ_D	θ_M	θ_{gc}	σ_{GS} (pc)	σ_S (pc)
1	4.65	0.5	4	0.5	0.5	0	89	54
2	4.65	1	4	1	0	-0.5	146	89
3	4.65	0.7	4.5	0	1	0.5	54	89

sets of simulated data from the birth-death MH algorithm for both the CO-GMC and GMC-SC models. Three sets of parameters are chosen for each model and given in Table 4.3 and Table 4.4. For the CO-GMC model, the parameter σ_0 is set as $\sigma_0 = R_c/2 = 42$ pc for simplicity. The prior distribution for the CO-GMC model is set as $\mathcal{N}(\mathbf{0}, 100^2\mathbf{I})$ where $\mathbf{I}$ is the identity matrix. Note that the parameters are reparameterized into log-scale to ensure that all parameters have positive support. For the GMC-SC model, the same prior distribution is chosen with all positive parameters reparameterized to log-scale.

Figure 4.8, 4.9, and 4.10 show the results of the three parameter sets for the CO-GMC model, respectively. Figures 4.11, 4.12, and 4.13 are the results for the GMC-SC model, respectively. The thick red line segments denotes the 50% credible intervals of bias against the true parameters obtained through the posterior distributions while the thin red lines are the 95% credible intervals. The red circles are the estimated posterior mean biases. The light red triangles and the horizontal black solid lines are the reference baseline of zero bias. The dotted black lines are the average bias obtained from all posterior samples.

For the CO-GMC model, all of Figures 4.8, 4.9, and 4.10 show similar results.

Almost all of the 95% posterior distributions cover the true parameters. However, few of the posterior distributions produce means that are close to the true values. There are several potential reasons for this to occur. First, this could be due to random discrepancies from the simulation of point patterns from the BDMH algorithm. The BDMH algorithm can only approximately simulate a point pattern that follows the specified GPP models. The simulation error introduced from BDMH algorithm will inevitably cause discrepancies between the true parameter values and the values corresponding to the simulated pattern. Secondly, simulating one point pattern is similar to generating one value from a standard normal random variable. Generating one value from a standard normal distribution will not necessarily give us a value that is close to the mean 0. It may happen that the generated value is, say, 2.1, and validating the inference algorithm using this value has no representative statistical power. Lastly, the prior distribution may have some effect on the resulting posterior distribution. Since several parameters are strictly positive, the posterior mean can be heavily affected by the choice of prior distributions. Therefore, an ensemble assessment of the DMH algorithm is to consider the average of all posterior mean represented by the dotted black lines in the Figures. We can see that the average of all 10 posterior means are very close to the true parameters, which indicates a correct implementation of the DMH algorithm.

In terms of the length of the credible intervals, they are generally similar for each parameters across all simulations. However, the lengths of the intervals differ a lot for different parameter sets. For example, the average credible interval lengths for σ for parameter set 1 is approximately 150 pc but only 80 pc for parameter set 2. This is mostly likely due to the effect from other parameters. Since for parameter set 1, $\log(\theta)$ is lower than that of parameter set 2. Furthermore, α for parameter set 1 is higher than that of parameter set 2. This means that the general first-order intensity of point process under parameter set 1 is lower than set 2. This means that

the average number of points far away from the CO filament structure is much lower for parameter set 1 than that of parameter set 2. This will result in a smaller sample available for estimating σ for parameter set 1 and a wider credible interval.

For the GMC-SC model, results from Figures 4.11, 4.12, and 4.13 paint a similar picture as the CO-GMC model. Almost all 95% credible intervals cover the true parameter and the average of posterior means across simulations are all very close to the true parameters. The reasons for posterior mean bias and fluctuation in estimation are essentially the same as that for CO-GMC model and I will not repeat it here.

From the above analysis on simulated data, we can confirm a reasonably good performance of the DMH algorithm in retrieving information in the data through the constructed model. We can now proceed to conduct data analysis on the real data using the constructed model and the DMH algorithm.

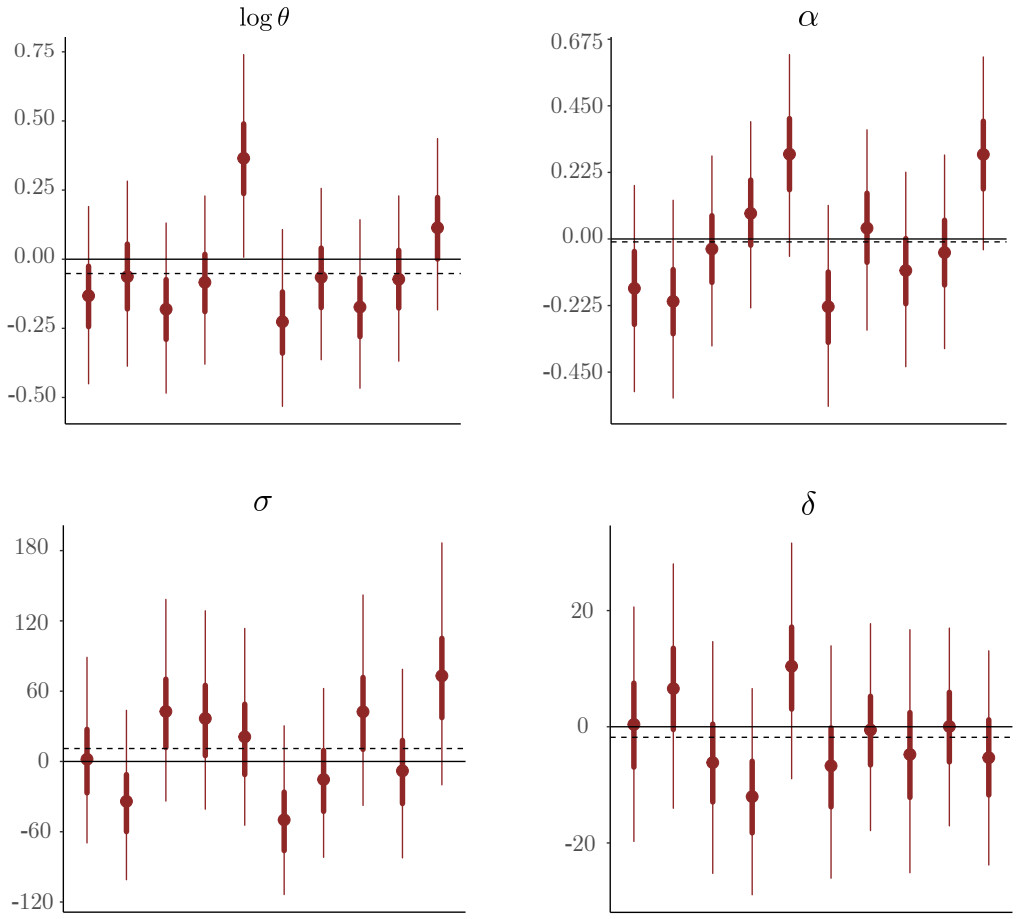


Figure 4.8: Plot of bias-adjusted posterior samples inferred from 10 simulated data sets for CO-GMC model under parameter set 1 ($\log(\theta) = 5$, $\alpha = 4.5$, $\sigma = 300$, $\delta = 100$). The thick red line segments denotes the 50% credible intervals of bias while the thin red lines are the 95% credible intervals. The red circles are the estimated posterior mean biases. The horizontal black solid lines are the reference baseline of zero bias. The dotted black lines are the average bias obtained from all posterior samples.

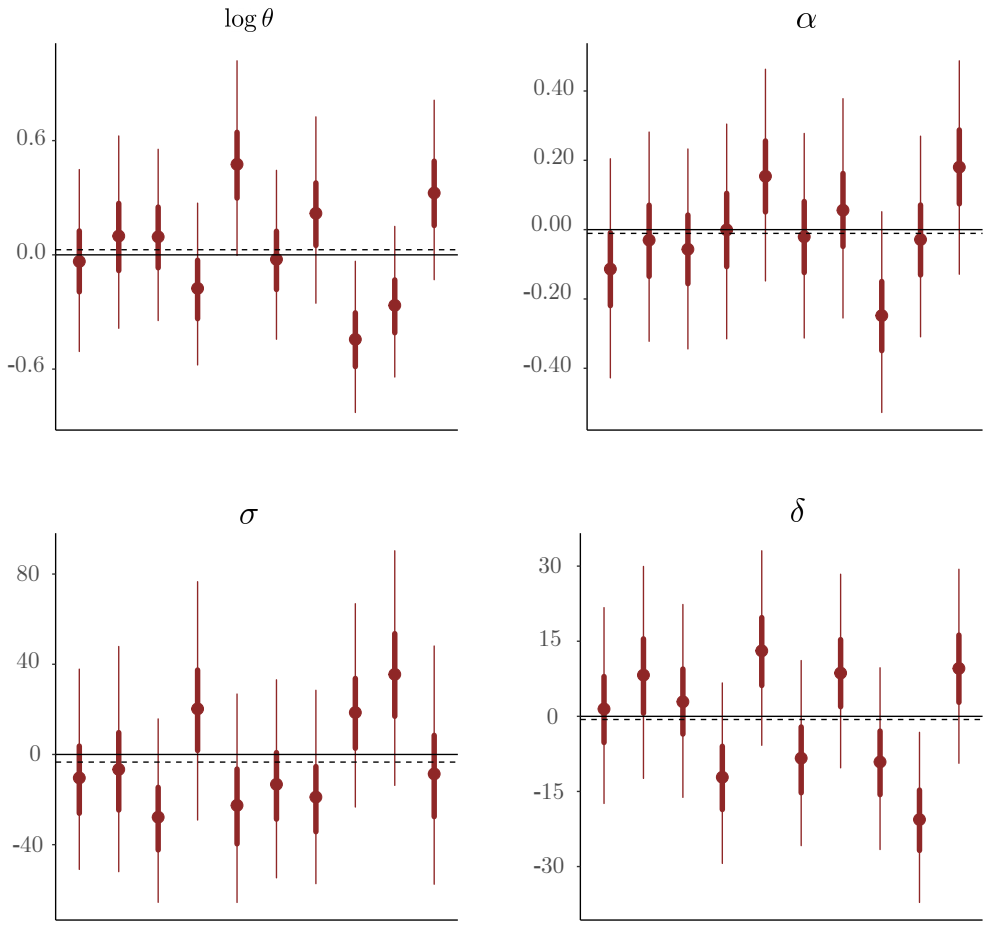


Figure 4.9: Plot of bias-adjusted posterior samples inferred from 10 simulated data sets for CO-GMC model under parameter set 2 ($\log(\theta) = 6$, $\alpha = 4$, $\sigma = 200$, $\delta = 150$). The thick red line segments denotes the 50% credible intervals of bias while the thin red lines are the 95% credible intervals. The red circles are the estimated posterior mean biases. The horizontal black solid lines are the reference baseline of zero bias. The dotted black lines are the average bias obtained from all posterior samples.

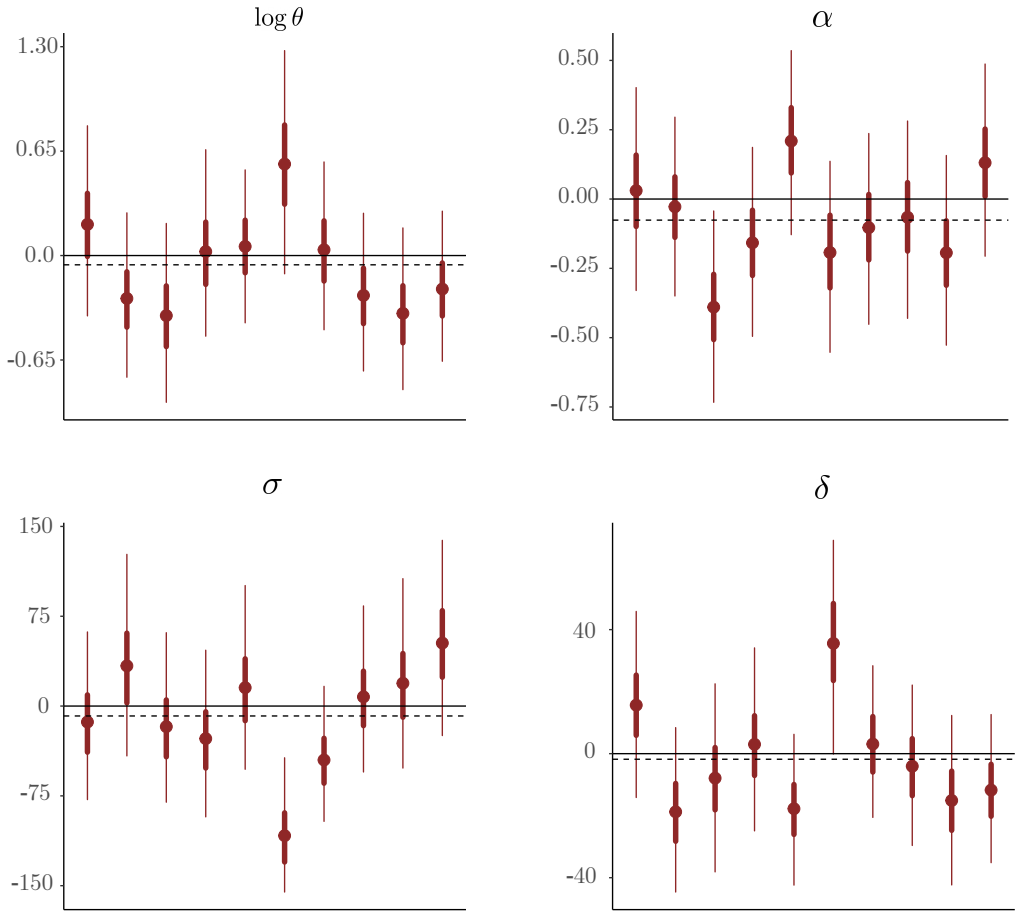


Figure 4.10: Plot of bias-adjusted posterior samples inferred from 10 simulated data sets for CO-GMC model under parameter set 3 ($\log(\theta) = 6.5$, $\alpha = 5$, $\sigma = 250$, $\delta = 200$). The thick red line segments denotes the 50% credible intervals of bias while the thin red lines are the 95% credible intervals. The red circles are the estimated posterior mean biases. The light red triangles and the horizontal black solid lines are the reference baseline of zero bias. The dotted black lines are the average bias obtained from all posterior samples.

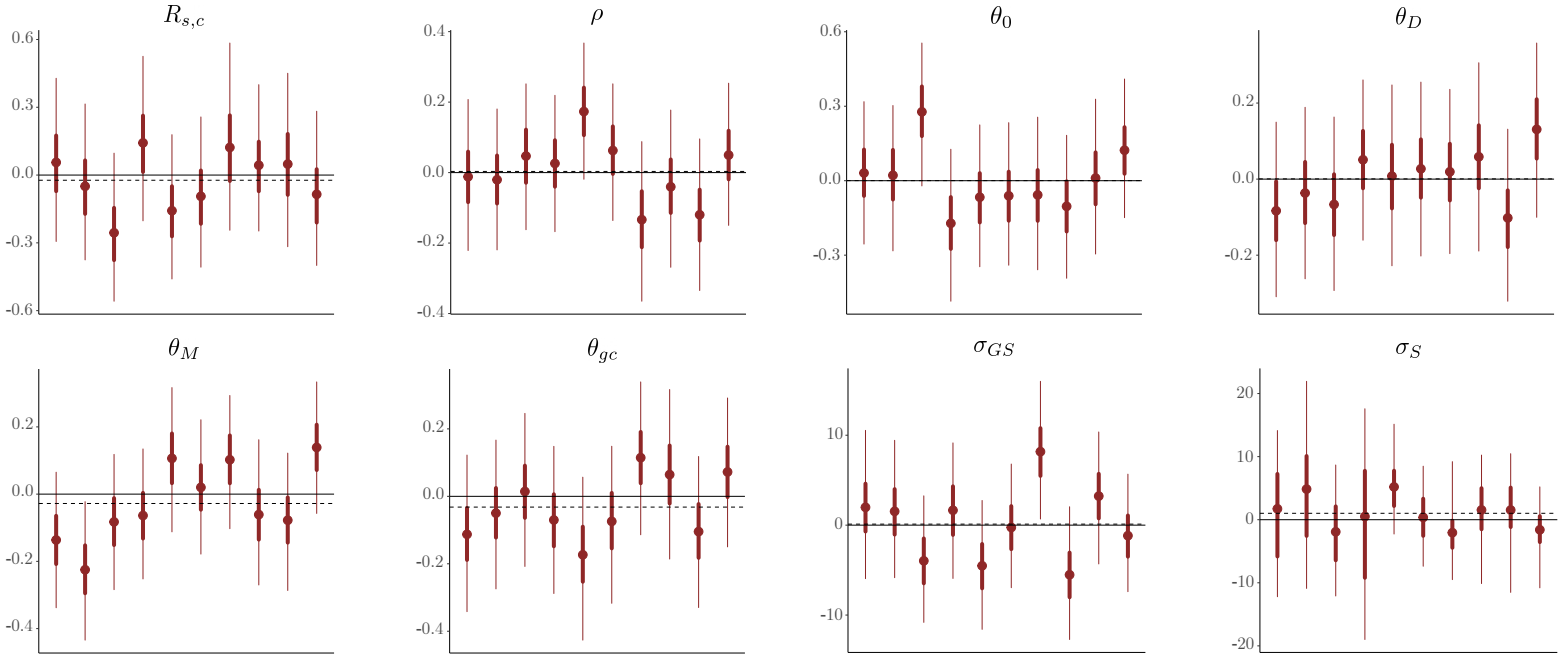


Figure 4.11: Plot of bias-adjusted posterior samples inferred from 10 simulated data sets for GMC-SC model under parameter set 1 ($R_{s,c} = 4.65$, $\rho = 0.5$, $\theta_0 = 4$, $\theta_D = 0.5$, $\theta_M = 0.5$, $\theta_{gc} = 0$, $\sigma_{GS} = 89$, $\sigma_S = 54$). The thick red line segments denotes the 50% credible intervals of bias while the thin red lines are the 95% credible intervals. The red circles are the estimated posterior mean biases. The horizontal black solid lines are the reference baseline of zero bias. The dotted black lines are the average bias obtained from all posterior samples.

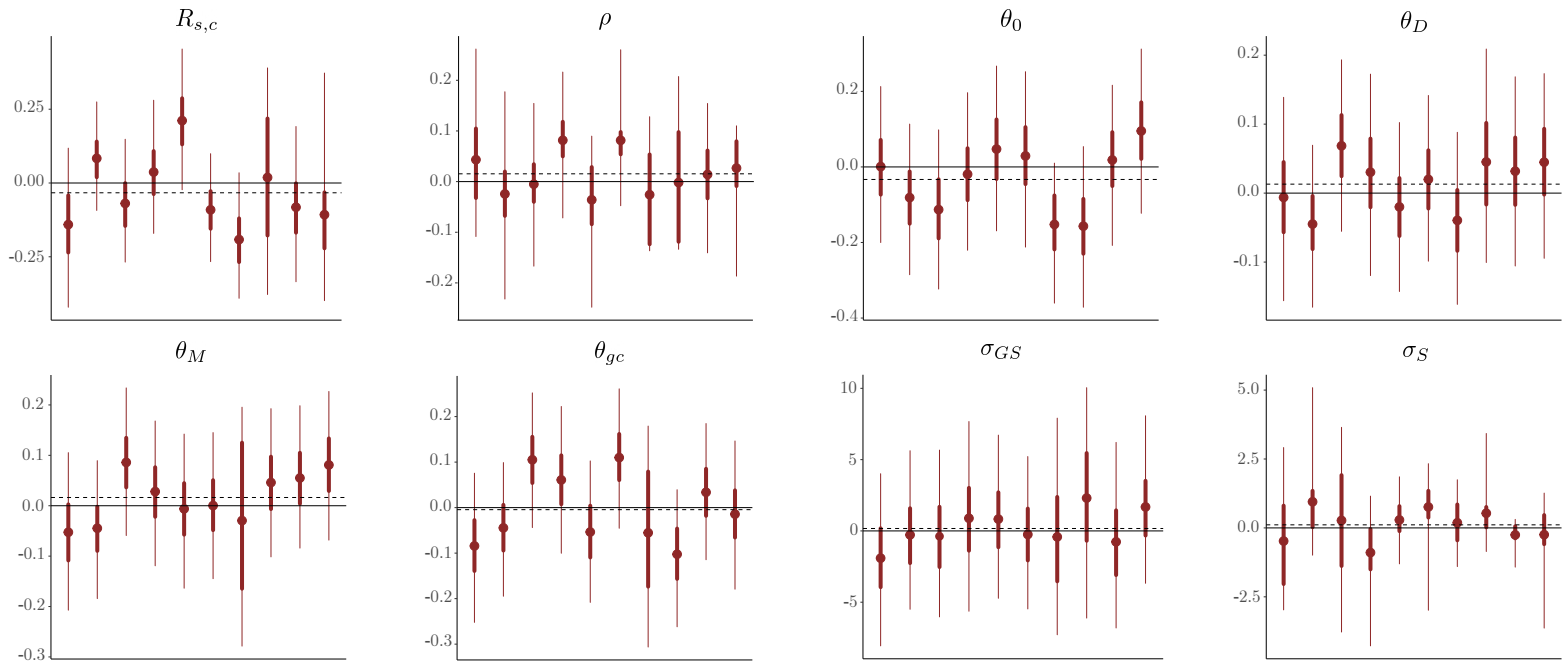


Figure 4.12: Plot of bias-adjusted posterior samples inferred from 10 simulated data sets for GMC-SC model under parameter set 2 ($R_{s,c} = 4.65$, $\rho = 1$, $\theta_0 = 4$, $\theta_D = 1$, $\theta_M = 0$, $\theta_{gc} = -0.5$, $\sigma_{GS} = 146$, $\sigma_S = 89$). The thick red line segments denotes the 50% credible intervals of bias while the thin red lines are the 95% credible intervals. The red circles are the estimated posterior mean biases. The horizontal black solid lines are the reference baseline of zero bias. The dotted black lines are the average bias obtained from all posterior samples.

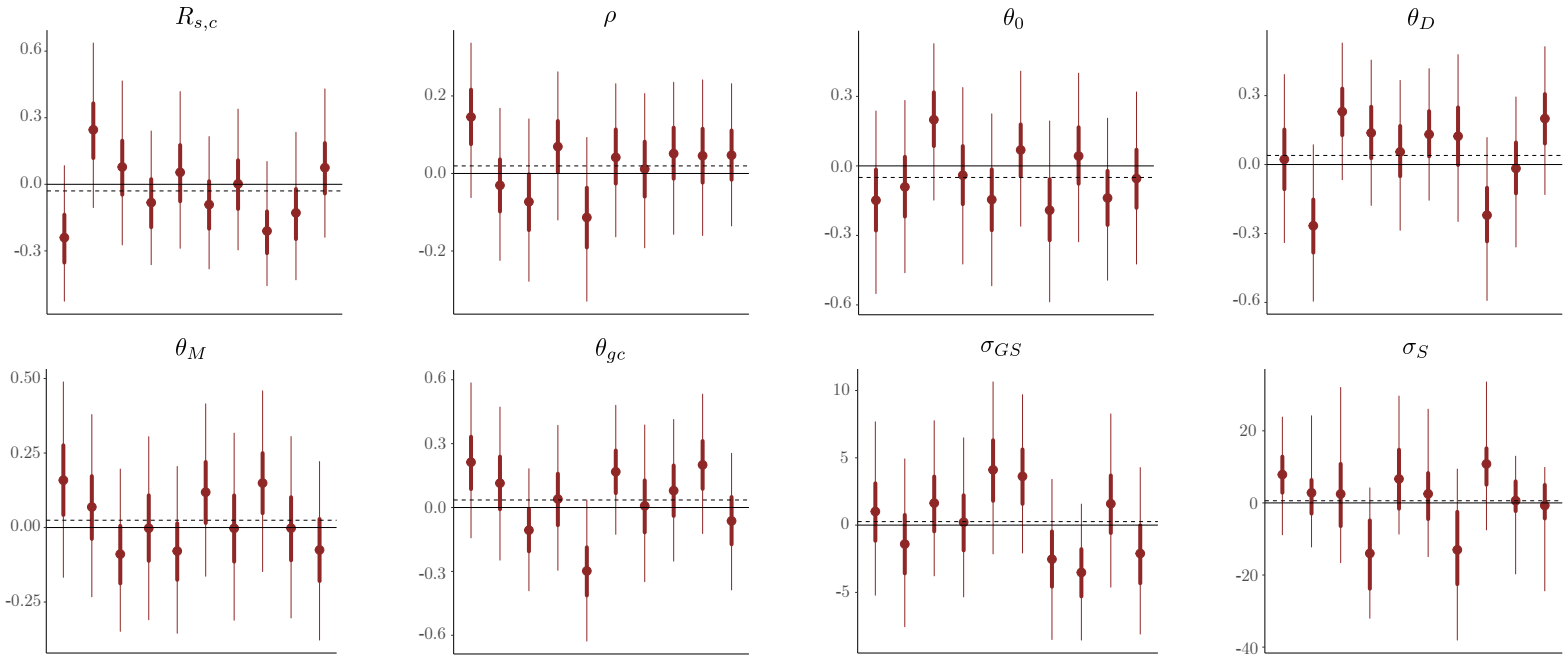


Figure 4.13: Plot of bias-adjusted posterior samples inferred from 10 simulated data sets for GMC-SC model under parameter set 3 ($R_{s,c} = 4.65$, $\rho = 0.7$, $\theta_0 = 4.5$, $\theta_D = 0$, $\theta_M = 1$, $\theta_{gc} = 0.5$, $\sigma_{GS} = 54$, $\sigma_S = 89$). The thick red line segments denotes the 50% credible intervals of bias while the thin red lines are the 95% credible intervals. The red circles are the estimated posterior mean biases. The horizontal black solid lines are the reference baseline of zero bias. The dotted black lines are the average bias obtained from all posterior samples.

Chapter 5

Data Analysis for Objects in M33

In this chapter, I will provide the results for the inferred model parameters as well as the model diagnostics. From these, I will illustrate the potential physical implications and insights on the star formation process in M33.

5.1 CO-GMC Model

5.1.1 Results

In this section, I present the fitted results for the high-level CO-GMC model. Note that the purpose in this research of the CO-GMC model is to serve as a preliminary demonstration of the performance of the GPP model. I will focus on how to interpret the fitted results and most importantly how to obtain critical information from model diagnostics. Since previous work on the distribution of GMCs is scarce due to the difficulty in obtaining high-resolution observation of GMCs, it is difficult to draw comparisons and obtain potential physical implications. I will instead put more focus on physical implications on the GMC-SC models as previous studies on distribution

of SCs are more numerous.

Table 5.1: Estimated posterior mean, MCMC standard error, and 95% highest posterior density (HPD) intervals for parameters in the CO-GMC model. 95% HPD intervals are calculated using **coda** package in R.

Parameters	**Posterior Mean (PM)**	**MCMC s.e. of PM**	95% **HPD Intervals**
$\log(\theta)$	6.1946	0.0072	(5.832, 6.556)
α	5.1340	0.0057	(4.797, 5.455)
σ (pc)	310.4500	1.6000	(237.98, 384.49)
δ (pc)	129.8700	0.3213	(113.40, 144.76)

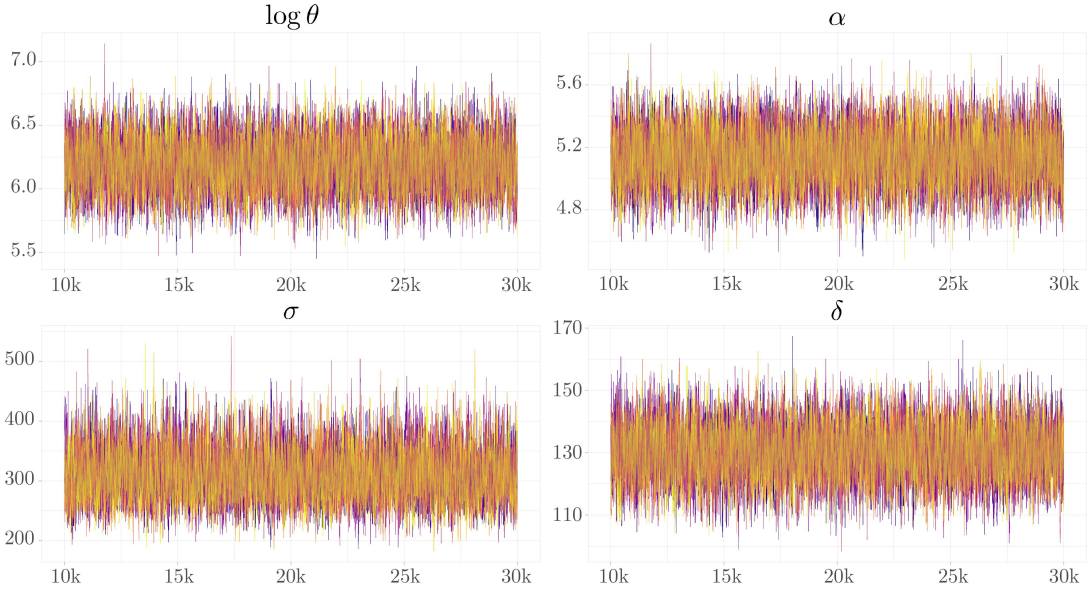


Figure 5.1: Traceplot of each parameter in the CO-GMC model obtained from ten MCMC runs for 30k iterations. The plot only shows the last 20k iterations for improved visualization.

There are a total of ten independent MCMC runs with 30,000 iterations. Since all parameters are strictly positive, the inference is carried out in the log-space. The prior distribution for each parameter is set to $\mathcal{N}(\mathbf{0}, 100^2\mathbf{I})$ where $\mathbf{I}$ is the identity matrix. Note that, similar to the simulation study, the parameters σ_0 and R_c are set to $R_c = 2\sigma_0 = 84$ pc. The reason for choosing $R_c = 84$ pc is provided in Chapter 4. Setting $\sigma_0 = R_c/2$ is to reduce computational complexity. Furthermore, due to

the close distances ($< R_c = 84$ pc) from the major population of GMCs to the CO filament structure, setting the characteristic scale to a fixed small value will not affect the resulting fitted model by much. The size of GMCs is set to a fixed value of 10 pc obtained from the median value of the GMC sizes.

For the hyperparameters of BDMH algorithm for simulating point patterns, they are specified as follow:

$$p_b = p_d = \frac{1}{2},$$

$$b(\mathbf{X}_t; \xi) = c_{0.01} \left(1 + \frac{d^2(\xi, y)}{0.01^2}\right)^{-1},$$

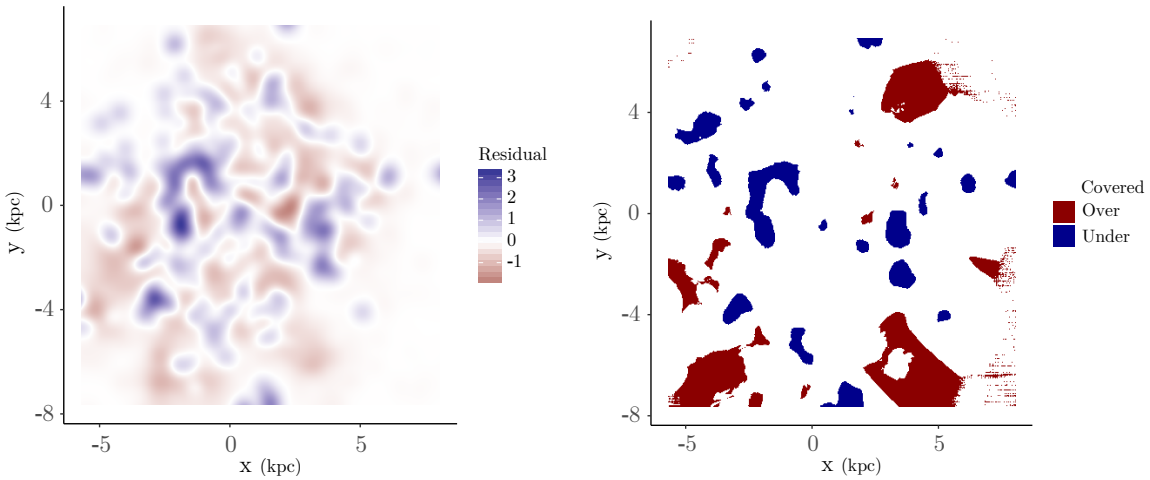
$$d(\mathbf{X}_t; \xi) = \frac{1}{n(\mathbf{X}_t)}.$$

Note that for the parameter h in the birth density $b(\mathbf{X}_t; \xi)$, a value of 0.01 produces a result of rejection samples that reasonably resembles the intensity variation of GMCs in the data. $c_{0.01}$ is obtained through a simple numerical integration over a fine grid on the observation window by

$$c_{0.01}^{-1} = \int_W \left(1 + \frac{d^2(\xi, y)}{0.01^2}\right)^{-1} d\xi.$$

The summary of the posterior distribution is given in Table 5.1. The MCMC convergence diagnostics are shown in Figure 5.1 which give the traceplots of the last 20,000 iterations of each chain. The plots clearly indicate convergence. A quantitative convergence test is conducted using the Gelman-Rubin statistic (Gelman and Rubin, 1992; Brooks and Gelman, 1998) on the ten independent chains. If the test statistic is close to 1, it indicates the convergence of the Markov chain. The resulting test statistic from the ten chains is 1 with the upper bound of the 99% confidence interval at 1. This strongly indicates the successful convergence of the DMH algorithm.

From Table 5.1, the fitted log-intensity parameter $\log(\theta)$ due to the CO filament structure is ~ 6.2, indicating that the CO filament structure has an extremely strong



(a) Mean posterior raw residuals of CO-GMC model (b) 95% pointwise credible intervals coverage

Figure 5.2: (a) Raw residuals obtained from kernel density estimations of the intensity of data and the intensity of 200 posterior simulation. (b) 95% pointwise credible intervals coverage; dark red shows the regions where the 95% credible intervals of raw residuals are below 0, i.e., model overestimates the intensity; dark blue shows the regions where the 95% credible intervals of raw residuals are above 0, i.e., model underestimates the intensity; white shows the regions where the 95% credible intervals of raw residuals cover 0.

effect on the distribution of GMCs. Numerically, this means that the presence of the CO filament structure will on average increase the intensity of GMCs by $\exp(6.2) = 492$ times compared to a unit rate Poisson process. This is in line with the general understanding that GMCs form from these filament structures. The posterior mean of the power law parameter α is ~ 5.13. This shows that the intensity of GMCs decreases drastically as one moves away from the CO filament structure. Combining the fitted results of θ and α, we can conclude that the GMC distribution is predominantly determined by the CO filament structure.

Now for the minor population of GMCs that are relatively far away from the CO filament structure, the posterior mean of the characteristic scale σ is approximately 310 pc. This shows that for the minor population of GMCs, the intensity decreases much more slowly as one moves away from the CO filament structure. A numerical conclusion we can obtain from the fitted results is that at 1 kpc away from the CO

filament structure, the average intensity of GMCs is less than 20 percent of a unit rate Poisson process. The scale parameter δ for the repulsive structure among GMCs is approximately 130 pc, indicating a rather strong repulsive structure among the GMCs.

5.1.2 Model Criticism

Figure 5.2 show the continuous posterior mean residual field between the data and model, obtained from 200 posterior predictive simulations. The continuous map of residuals are computed using a 400×400 grid using the package **spatstat** in R. A radial basis function is used as the smoothing kernel and the bandwidth is chosen to be 510 pc, which is selected through cross-validation using the built-in function from **spatstat**. Figure 5.2 (a) shows the posterior mean residual field. We can see that in general, the residual field is very close to a 2D white noise across the observation window, indicating a good fit of model to the data. Figure 5.2 (b) shows the 95% pointwise credible intervals coverage. This is done by considering the residual values from all 200 posterior predictive simulations for each grid point and constructing the 95% credible intervals from these residuals. Then it is determined whether zero falls below, above, or within each credible interval. We can see from Figure 5.2(b) that there are regions with consistent overestimation and underestimation, this may suggest that there may exist certain levels of misfit of the model. To pinpoint the cause, we will have to look at the fit of the model in terms of the fit of the covariance effect and the second-order structure.

For the fit of the covariance effect, Figure 5.3 shows the comparison between the cumulative distribution functions for log-NND from GMCs to the CO filament for the data and the model. The black line is the distribution function for data and the red line is the posterior mean distribution from 200 posterior predictive simulations. The light and dark red bands are the pointwise 50% and 95% credible

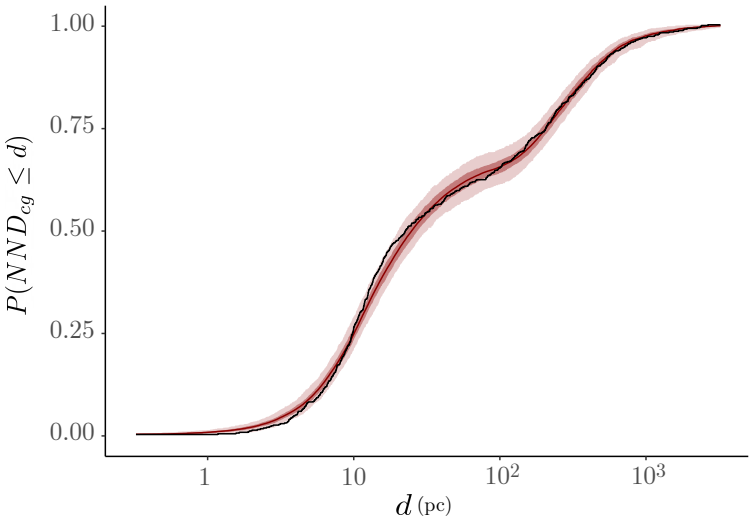


Figure 5.3: Distribution function of log-NND from GMCs to the CO filament structure for data and model: dark red line is the posterior mean distribution from 200 posterior predictive simulations; black line is the empirical distribution of data; dark red band is the pointwise 50% credible intervals while light red band is the 95% credible intervals.

intervals respectively. We can see that the model fits the data almost perfectly; this means that the covariance effect is sufficiently accounted for.

For the second order characteristics, Figure 5.4 and Figure 5.5 show the comparison of PCF and G-function between the data and model respectively. Similar to Figure 5.3, the black line is the statistics of the data while the red line is the estimated posterior mean statistics from 200 posterior predictive simulations. The dark and light red bands are pointwise 50% and 95% credible intervals respectively.

From Figure 5.4, we see that the empirical PCF from the data is generally within the 95% credible intervals. however, it is well beyond the 50% credible intervals and above the the posterior mean PCF. On the other hand, Figure 5.5 indicates that the model fits the data very well in terms of the G-function.

The analysis of Figure 5.3, 5.4, and 5.5 seems to suggest that the discrepancy between the model and data is due to an underestimation in the second-order intensity observed in Figure 5.4. However, an interesting observation of Figure 5.4 provides a clue that it might be due to another less obvious reason. As we can see from

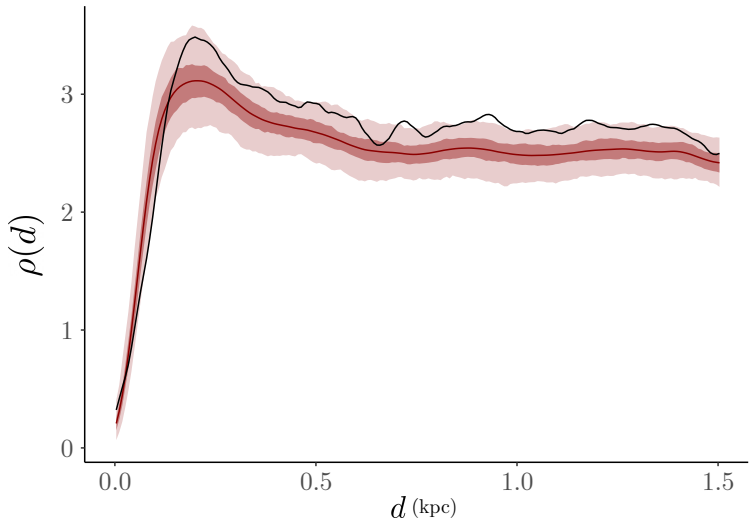


Figure 5.4: PCF comparison between data and CO-GMC model: black line is the PCF obtained from data; dark red line is the estimated mean PCF under the model obtained through 200 posterior simulation; dark red band is the grey band is the pointwise 50% credible intervals of the PCF at each d under the model while light red band is the 95% credible intervals.

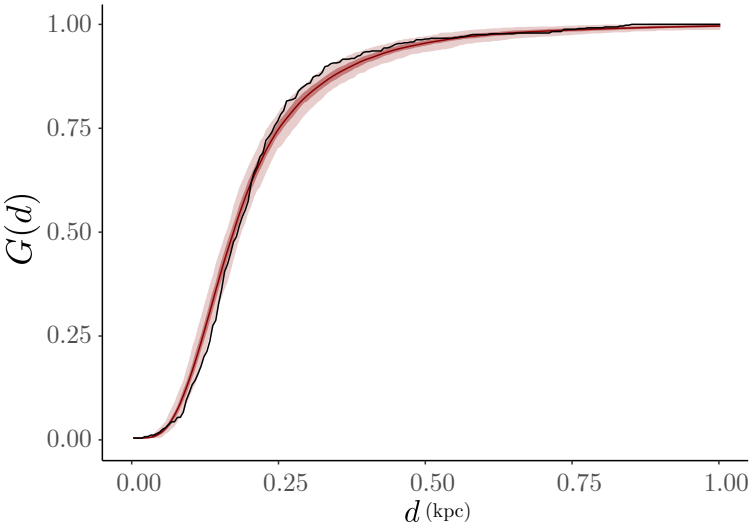


Figure 5.5: G-function for GMCs: dark red line is the estimated mean G-function from 200 posterior simulations; black line is the G-function estimated from data; dark red band is the estimated pointwise 50% credible intervals of the G-function for the model while light red band is the 95% credible intervals.

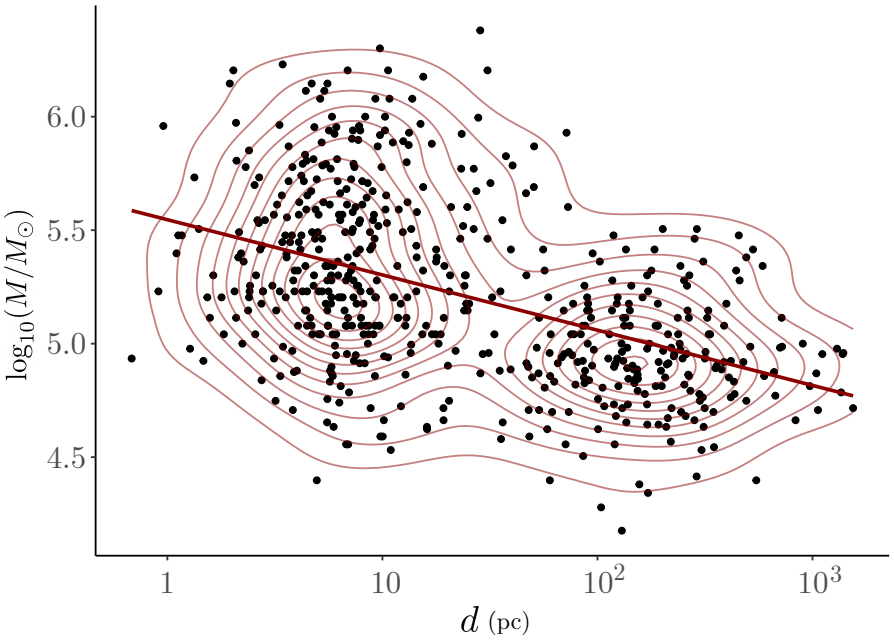


Figure 5.6: Mass of GMCs vs NND from GMCs to the CO filament structure in log-scale.

Figure 5.4, the empirical PCF of the data is always above the posterior mean PCF of the model from 180 pc all the way to 1500 pc. This usually happens not because of underestimation of second order clustering but rather due to a misfit in the first-order intensity, as it is very unlikely for second order clustering behavior to persists over such a wide range of scale. Realistically, it does not make physical sense for there to exist second-order clustering between two GMCs at over 1 kpc scale. However, as we have seen in Figure 4.3, the first order intensity exhibits an almost perfect fit between model and the data. So what could be the issue here? The culprit here is most likely the non-stationarity of the contribution to the first order intensity from the CO filament structure. In the model, it is assumed that the intensity parameter θ is the same at any point on the CO filament structure. However, this is unlikely to be the case since in reality there exists inhomogeneity of the CO intensity at different points on the CO filament structure. This will likely lead to a varying first-order intensity for GMCs depending on the position of a point relative to the filament structure.

Furthermore, there might exist other formation mechanism of GMCs in that they do not necessarily all form from the CO filament structure and then separate from their natal environment. Besides the CO filament structure, there could be interstellar medium permeating other regions of space. Although the abundance of ISM can be much lower compared to the CO filament structure, it can still potentially form GMCs due to accumulation (Corbelli, Braine, and Giovanardi, 2019). The distribution of ISM in these regions is highly unlikely to be homogeneous, which can lead to the misfit of the model to data shown in Figure 5.2(b).

The evidence for the above claim that certain portion of GMCs might not originate from the CO filament is two fold. First, from Figure 4.3(b), we have observed two sub-populations of GMCs based on the distance from GMCs to the CO filament structure. If all GMCs form from the CO filament and then drift away from their natal environment, the distribution of the distance from GMCs to the CO filament structure should not exhibit the clear bimodal distribution observed in Figure 4.3(b). Second, Figure 5.6 shows the scatter plot of the log-mass of GMCs and the log-NND from GMCs to the CO filament structure. We can clearly see there is a negative relationship between the two variables with an estimated slope of the linear regression line at -0.24. Furthermore, the major population of GMCs are 3 times more massive than the minor population on average. This suggests that the GMCs farther away from the CO filament structure might form out of the field of ISM with an intensity much lower than the filament structure. The less ISM rich environment also corresponds to the fact that the number of GMCs is much fewer in the less massive population.

5.2 GMC-SC Model

5.2.1 Results

In this section, I present the results of the model for the distribution of GMCs and YSCCs. For the hyperparmeters of BDMH algorithm, they are the same as the ones used for the CO-GMC model. There are a total of ten independent MCMC runs and 100k iterations are carried out for each run. Again, the parameters whose domain is strictly positive are transformed into log-scales. The prior distribution for each parameter is set to $\mathcal{N}(\tilde{\boldsymbol{\theta}}, 100^2\mathbf{I})$ where $\tilde{\boldsymbol{\theta}}$ is a crude estimate based on the MPLE approach and $\mathbf{I}$ is the identity matrix. Irregular parameters that cannot be inferred by the MPLE approach, such as σ_{GS}, are estimated based on summary statistics: $R_{s,c}$ is based on the histograms in Figure 4.5; σ_{GS} is estimated using the median nearest neighbor distance between YSCCs and GMCs; σ_S is set to the generally accepted scale for star formation complexes (Chevance et al., 2019). For parameters governing the relationship between marks of GMCs and the correlation strength of GMCs and YSCCs, I set them to 0 as it is not clear how to obtain a crude estimate. Table 5.2 shows the crude estimate $\tilde{\boldsymbol{\theta}}$.

The MCMC convergence diagnostics are shown in Figure 5.7. For better visualization, only the last 20k iterations of each chain are shown in Figure 5.7. The plots clearly indicate the convergence of the chains. The resulting Gelman-Rubin test statistic from the ten chains is 1 with the upper bound of the 99% confidence interval at 1, indicating convergence of the chain.

Table 5.3 lists the summary of the posterior sample of the model parameters. I discard the first 50k iteration of the chains as burn-in and choose the chain with the highest effective sample size as the representative sample of the posterior distribution.

Table 5.2: Crude estimate of GMC-SC model parameters

$\log(\tilde{R}_{s,c})$ (kpc)	$\tilde{\rho}$	$\tilde{\theta}_0$	$\tilde{\theta}_D$	$\tilde{\theta}_M$	$\tilde{\theta}_{gc}$	$\log(\tilde{\sigma}_{GS})$ (pc)	$\log(\tilde{\sigma}_S)$ (pc)
log(5)	0.8	4	0	0	0	log(76)	log(100)

Table 5.3: Estimated posterior mean, MCMC standard error, and 95% highest posterior density (HPD) intervals for parameters in the GMC-SC model. 95% HPD intervals are calculated using the **coda** package in R.

Parameters	**Posterior Mean (PM)**	**MCMC s.e. of PM**	95% **HPD Intervals**
$R_{s,c}$ (kpc)	4.7992	0.0067	(4.478, 5.137)
ρ	0.6876	0.0035	(0.509, 0.855)
θ_0	4.4915	0.0048	(4.239, 4.744)
θ_D	0.8513	0.0034	(0.641, 1.073)
θ_M	0.6635	0.0039	(0.442, 0.873)
θ_{gc}	-0.0456	0.0048	(-0.267, 0.185)
σ_{GS} (pc)	84.6900	0.1231	(78.40, 90.78)
σ_S (pc)	79.6200	0.2075	(69.03, 89.53)

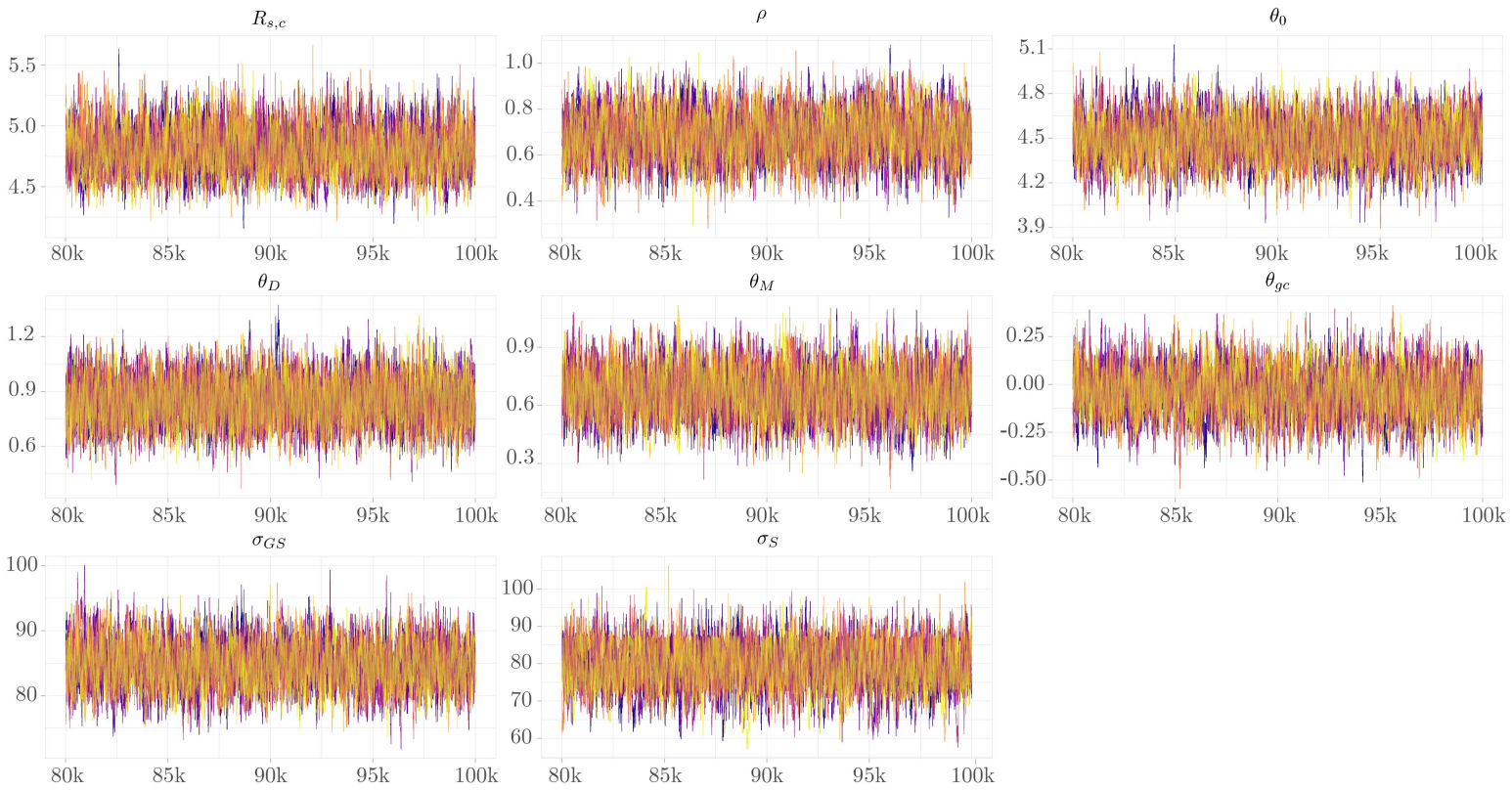


Figure 5.7: Traceplot of each parameter in GMC-SC model obtained from ten MCMC runs for 100k iterations. The plot only shows the last 20k iterations for improved visualization.

We see that the characteristic scale of the YSCs in the galactic plane, represented by $R_{s,c}$, is $\sim$ 4.8 kpc from the center of the galaxy. This coincides well with the mean of the prior distribution for $R_{s,c}$ at 5 kpc. The central intensity, ρ, controlling the galaxy-wide first-order log-intensity of the distribution of the YSCs is only about 0.68. This means at the center of the galaxy, the first-order intensity contributed by the large-scale intensity is approximately $\exp(0.68) = 1.97$ kpc^{-2}. This can be explained as approximately 2 YSCCs per kpc^2 at the galaxy center being not due to the presence of GMCs, rather the general intensity variation across the galaxy disc. This number will then drop as one moves away from the galaxy center. Now at the immediate surroundings of a GMC, the baseline correlation strength parameter θ_0, or the first-order log-intensity contributed by an average GMC is around 4.5. This means that at the same galactocentric distance, the increase in the intensity from a region with no GMC to the center of an average GMC is a walloping $\exp(4.5) = 90$ times. This indicates that the GMCs have a huge impact on the distribution of the YSCs and serves as ample evidence to the claim that GMCs are the birthplace of YSCs.

Note that, however, $\rho = 0.68$ does not equate to saying the overall intensity contributed by the large scale first-order intensity is 2 YSCCs per kpc^2 at the galaxy center. Rather, we do not know the overall intensity as it is also governed by the second-order intensity as well. However, the increase in the overall intensity from regions with no GMC to the vicinity of an average GMC is indeed 90 times.

The value of $\theta_{\text{dist}}, \theta_{\text{mass}}, \theta_{\text{gc}}$ indicate interesting effects from the properties of the GMCs on the correlation strength between GMCs and YSCCs. The effect of distance from the galactic center to GMC, represented by θ_{dist}, indicates that if the distance increases by 1 standard scale, the correlation strength between GMCs and YSCCs increases by 85%, while 1 standard log-scale increase in the mass of the GMC leads to a correlation strength increase by about 66%. The effect from the distance between

GMCs and the CO filament structure, however, does not seem to have a significant effect on the correlation strength.

The characteristic scale, σ_{GS}, of the distribution of YSCs around GMCs is ∼85 pc, which means that the effect of GMCs on the intensity of YSCs only has a very limited range.

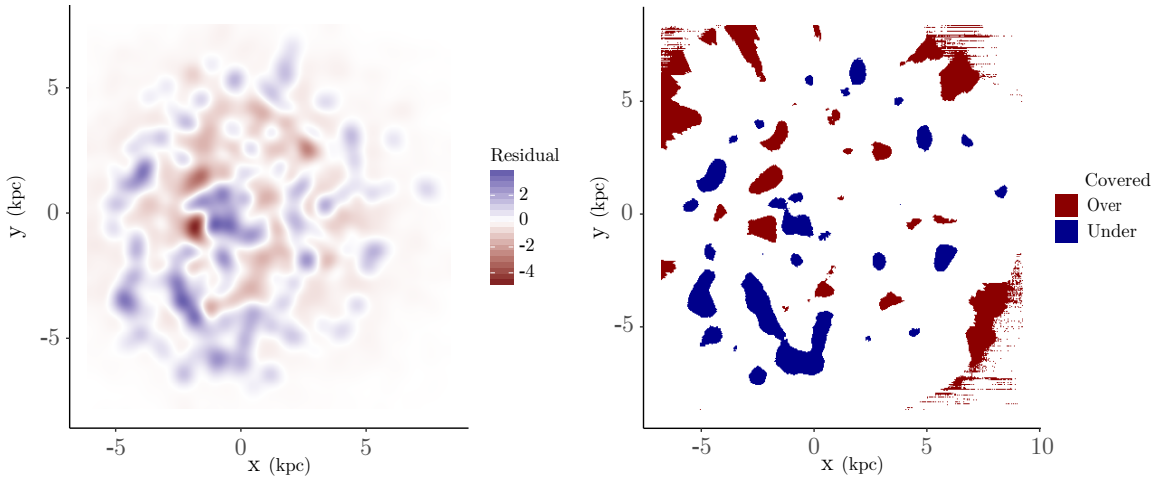
For the second-order intensity, the characteristic scale σ_S is $\sim$ 79 pc. According to the model, this means that, on average, the interpoint interaction between two YSCs disappears, i.e., the point pattern becomes (inhomogeneous) Poisson, once the distance is greater than $\sim$ 105 pc.

I will defer the detailed interpretation of these parameters to later sections.

5.2.2 Model Criticism

Now for model criticism, Figure 5.8 shows the intensity residuals obtained by comparing the data and simulation from the fitted model using 200 posterior samples. The intensity residuals are obtained through the methods described in Chapter 3.

The continuous residual map is given in Figure 5.8. The procedure is similar to that for constructing Figure 5.2. The kernel bandwidth is chosen to be 420 pc determined by cross-validation. Figure 5.8 (a) shows the posterior mean residual field. We can see that in general, the residual field is close to 2D white noise across the observation window, although there seems to exist certain structure that traces out the spiral structure of the galaxy. Nevertheless, it does indicate an overall reasonably good fit of the model to data. To further pinpoint the fit of the model, Figure 5.8 (b) shows the 95% pointwise credible intervals coverage. Note that the consistent overestimation in the corner regions of the observation window should be ignored since it poses no meaningful physical implications. This is due to the fact that the large scale intensity of the model has a positive probability of point occurrence in these regions while in the real data, there is no point observation. This eventually



(a) Mean posterior raw residuals of GMC-SC model (b) 95% pointwise credible intervals coverage

Figure 5.8: (a) Raw residuals obtained from kernel density estimations of the intensity of data and the intensity of 200 posterior simulation. (b) 95% pointwise credible intervals coverage; dark red shows the regions where the 95% credible intervals of raw residuals are below 0, i.e., the model overestimates the intensity; dark blue shows the regions where the 95% credible intervals of raw residuals are above 0, i.e., the model underestimates the intensity; white shows the regions where the 95% credible intervals of raw residuals cover 0.

leads to the perceived overestimation of the intensity. In fact, if we look at Figure 5.8(a), the posterior mean residuals in these regions are very close to zero, therefore, the overestimation of intensity in these regions is not of concern. However, we do see from Figure 5.8 (b) an interesting result in that the intensity in the outer region is getting consistently underestimated, denoted by the large blue blocks in the plot. This can potentially have multiple explanations and we will need other model diagnostics to pinpoint the possible cause.

To further our diagnostics, Figure 5.9 shows the comparison of the empirical PCF obtained from data and that of the model. The black line in the plot shows the empirical PCF inferred from real data and the red line is the posterior mean PCF obtained using the 200 posterior predictive simulations. The dark and light red bands are the pointwise 50% and 95% credible interval respectively, obtained through the 200 simulations. From the plot we can see that the empirical PCF is within the credible

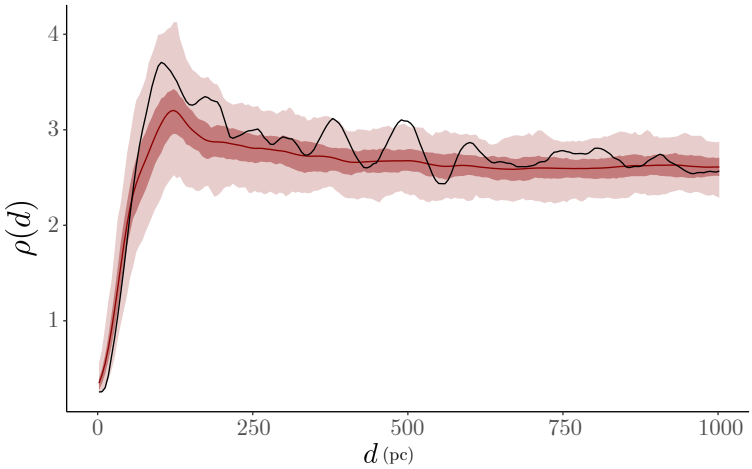
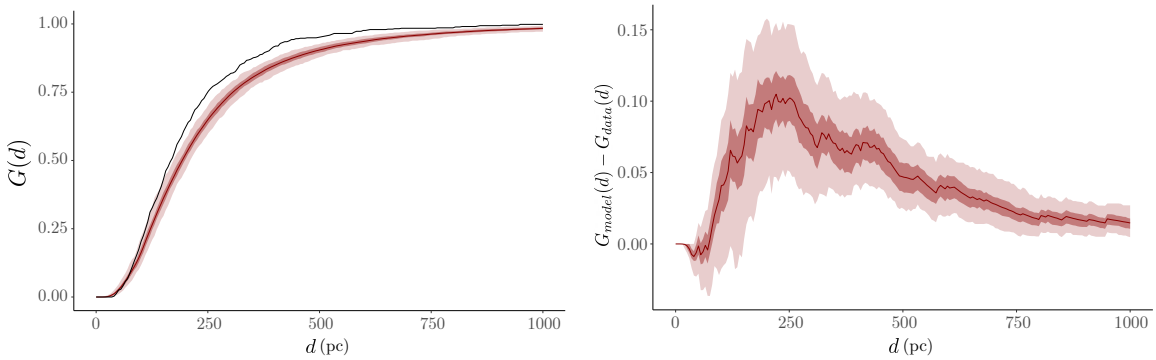


Figure 5.9: PCF comparison between data and GMC-SC model: black line is the PCF obtained from data; red line is the estimated mean PCF under the model obtained through 200 posterior simulation; grey band is the pointwise 95% credible intervals of the PCF at each r under the model.

band at almost all distance. However, we do see that the empirical PCF from data has a large deviation from the model mean at $r = 125$ pc and it mostly remains above the model mean all the way to around 500 pc. This deviation may correspond to the underestimation of the intensity in the outer region of the galaxy disc as shown in Figure 5.8. However, it seems to contradicts the fact that the empirical PCF from data is within the credible bands for almost all distance. As mentioned in Chapter 2, the PCF itself is not sufficient to fully characterize a point pattern as it has potential blind spots. To have a well-rounded view of the second-order characteristics, the G-function (NND distribution) is also plotted in Figure 5.10.

Figure 5.10 (a) shows the G-function between the real data (black line) and the posterior mean (red line) from the same 200 posterior simulated data used for Figure 5.9. Figure 5.10 (b) shows the difference between the G-functions of data and model. We see that in the short range ($r < 100$ pc), the G-functions of data and model match reasonably well. However, starting from approximately 150 pc, the point pattern from the data becomes more clustered than the model, peaking at around 250 pc with a difference of 0.1, i.e., on average, a YSC from the data has an excess of 10 percent



(a) G-function for YSCCs (b) Differences of G-functions for YSCCs

Figure 5.10: (a) G-function for YSCCs: red line is the estimated mean G-function from 200 posterior simulations; black line is the G-function estimated from data; grey band is the estimated pointwise 95% credible intervals of the G-function for the model. (b) Differences in the G-functions between data and model: red line is the mean difference; grey band is the pointwise 95% credible interval.

chance to that of the model of finding another YSC as its neighbor within 250 pc. This discrepancy of clustering behavior declines but persists all the way to over 600 pc. We see that this significant discrepancy is not reflected by simply comparing the PCF of data and model. This is most likely due to the fact that PCF is an averaged statistic where all pairwise distances are taken into account while the G-function is only considering nearest neighbor distance. In the sense of local structure, the G-function can be much more sensitive than the PCF.

It is important to note that since the inferred repulsive range $R_P \approx 105$ pc, the clustering feature is indeed with respect to a Poisson process. Combining the findings from Figure 5.8 (b), we can conclude that this discrepancy originates from the underestimated blocks in the outer region of the galaxy. However, there are three potential causes for this underestimation: (1) the underestimation of the large-scale inhomogeneity in the outer region; (2) the underestimation of the effect from GMCs; (3) second-order clustering not accounted for by the model. To determine the cause, we carry out two other analyses.

To see the general estimates of the large scale effect, Figure 5.11 shows a count

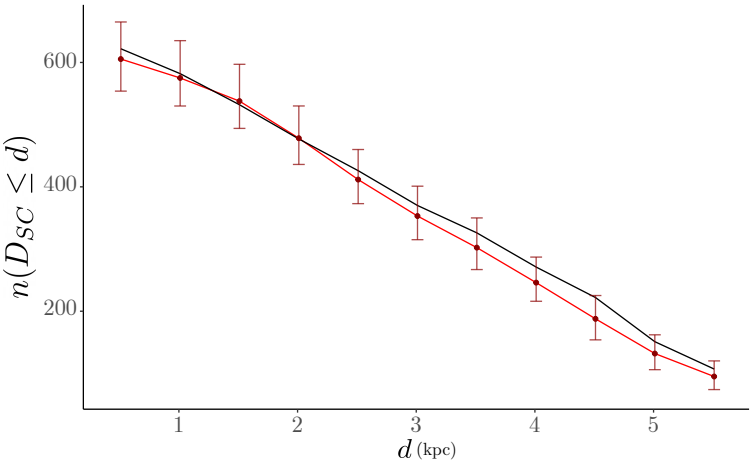


Figure 5.11: Count of points that are at least distance d away from the galactic center with d increasing from 0.5 kpc to 5.5 kpc; red line and dots are the mean count obtained from simulated data using 200 posterior samples; dark red vertical lines are 95% credible intervals of the count at each distance of d where the count is calculated; black line is the true count at each d where the count is calculated.

comparison between the data and the model with respect to the distance from the center of the galaxy to its outer rim. We do this by counting the number of points in the region that is distance r away from the galaxy center, where r ranges from 0.5 kpc to 5.5 kpc, in 0.5 kpc increments. We compare the statistics from the data to what is obtained from simulation of 500 posterior samples. Figure 5.11 shows that the data and the model are generally in good accordance with each other, meaning that the large scale inhomogeneity is indeed properly accounted for.

Figure 5.13 shows an overlay of GMCs and YSCCs on top of the residuals from Figure 5.8(a). Figure 5.13 shows that in the outer rim, the regions where the intensity is consistently underestimated in fact have no or disproportionately few GMCs in their vicinity. Note that we determine the vicinity by referencing the estimated characteristic scale σ_{GS} between GMCs and YSCCs which is only about 85 pc. We also marked the regions with no or few GMCs in their surroundings with ellipses for better visualization. It is easily seen that these ellipses corresponds to the regions where intensity is consistently underestimated in Figure 5.8(b).

Furthermore, using the same simulated data obtained for Figure 5.11, I plotted the

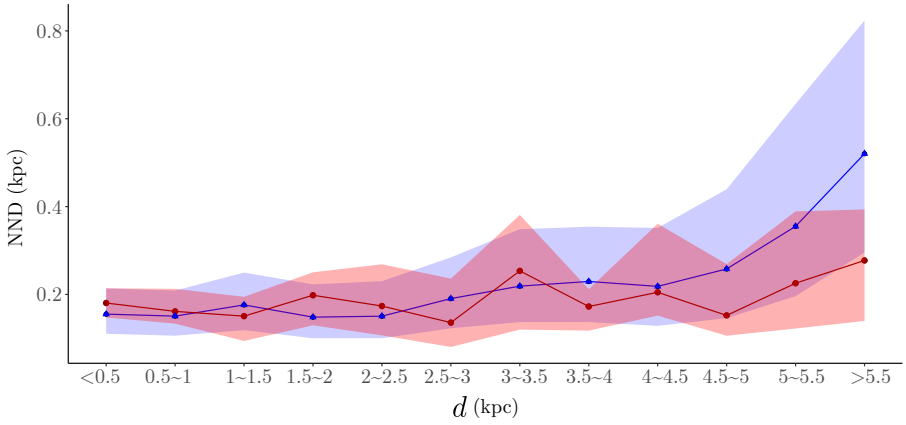


Figure 5.12: 50% credible intervals of nearest neighbor distances (NND) of YSCCs grouped by distance to the galaxy center. Red band denotes the central 50% confidence intervals of NND for each annuli obtained from data; red dots are the median NND from data; blue band denotes the central 50% credible intervals of NND for each annuli obtained from 200 posterior simulations; blue triangles are the median NND from the 200 posterior simulations;

comparison of the NND distribution of YSCCs in each annulus encircling the galaxy center. The result is shown in Figure 5.12. As shown in Figure 5.12, the discrepancy between the NND distribution in each annulus is reasonable until the annuli start to reach the outer region of the galaxy, at $r > 4.5$ kpc. Furthermore, the median NND distance of YSCCs in the outer region is generally close to 250 pc, which corresponds exactly to the distance at which the peak of discrepancy is reached in the G-function in Figure 5.10. This proves that the discrepancy between the data and the model indeed comes from the underestimation of intensity in the outer region, i.e., the blue blocks shown in Figure 5.8(b). However, this also shows that the underestimation is not due to the misfit of the large-scale intensity in the outer region.

To determine whether this discrepancy is due to the underestimation of effect from GMCs, I present the following figures.

For a more quantitative inspection, we also plot the bivariate density between the distance from a GMC to its nearest neighbor in YSCCs (R_{gs}) against the distance from that YSC to its nearest neighbor in YSCCs (R_{ss}). This is shown in Figure 5.14. We see here that there is a huge discrepancy between the data and the model when

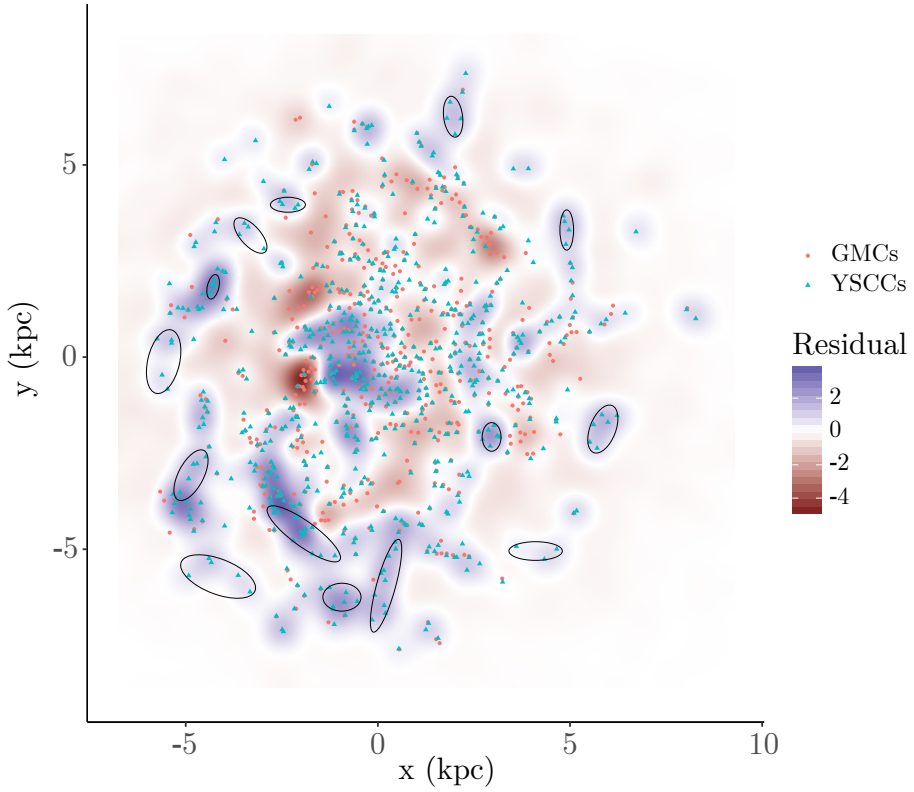
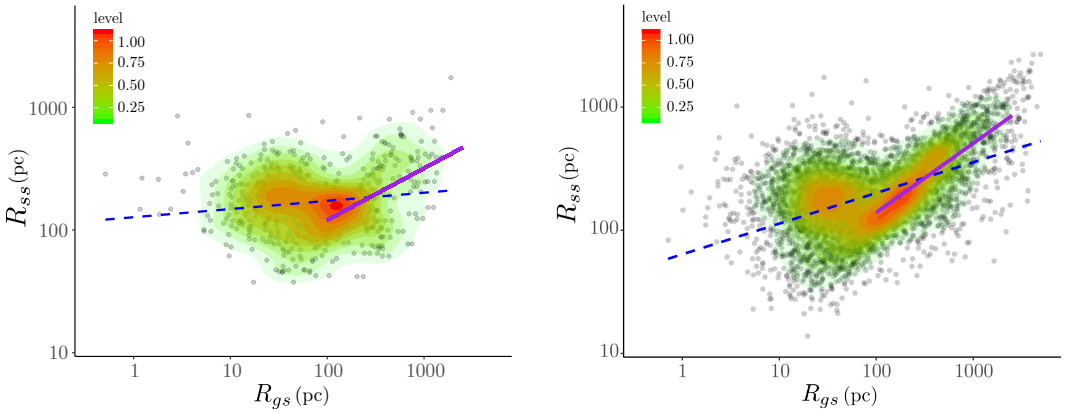


Figure 5.13: GMCs and YSCCs overlaid on raw residuals between data and model. Ellipses in the plot show the regions where the intensity is underestimated by the model and there are no or disproportionately few GMCs in the vicinity of YSCCs.

$R_{gs} > 100$ pc, however, there is not much discrepancy at $R_{gs} < 100$ pc. The blue-dashed lines in the plots are the fitted least-squares lines between R_{gs} and R_{ss}. The slope of the real data is ~ 0.06 while the slope of simulated data is ~ 0.25. The purple lines are fitted least-squares lines given $R_{gs} > 100$ pc. The slope for the real data is ~ 0.42 and the slope for simulated data is ~ 0.56. From this, we can determine that the point pattern in the data is in fact more clustered than the simulated data from the model when the YSCCs considered are far away from the GMCs. Furthermore, given that this discrepancy occur at range $R_{gs} > 100$ pc and peaks at 250 pc range, we can conclude that this discrepancy is not caused by underestimation of the correlation with GMCs. Simply from a physical sense and from the inferred value of σ_{GS}, the influence of GMCs on YSCCs should not extend to over 250 pc.



(a) Real R_{gs} vs R_{ss} (b) Posterior simulated R_{gs} vs R_{ss}

Figure 5.14: Density contours of distance from YSCs to nearest neighbor in GMCs (R_{gs}) against the nearest neighbor distance between YSCs (R_{ss}); (a) Plot obtained from real data; (b) Plot obtained from 200 posterior simulations; dashed blue lines are the fitted least squares lines between the two distances; solid purple lines are the fitted least squares lines between the two distances for $R_{gs} > 100$ pc. The plots are in log-log scale.

Therefore, combining all the results from the previous analysis, we see that there are indeed second-order clustering patterns unaccounted for by the model at 150–600 pc scales in the outskirts of the galaxy.

5.2.3 Comparison to Previous Studies & Physical Implications

First-Order Potential and Correlation Structure

The parameters governing the first-order potential provide some very interesting insights on the star formation process in M33. As mentioned, the central log-intensity for the large scale spatial trend of YSCCs is $\rho \approx 0.68$ compared to the baseline effect/correlation strength $\theta_0 \approx 4.5$ from an average GMC. This confirms that there indeed is a strong correlation between GMCs and YSCCs as suggested by Corbelli et al. (2017) and it provides rigorous proof that this correlation between GMCs and YSCCs is not simply due to the general overlapping distribution among them across the galaxy disc. This also provides sufficient indirect evidence that GMCs are indeed

birth places of YSCCs since for a correlation structure to be this strong, random alignments of GMCs and YSCCs are highly unlikely to be the cause.

On the other hand, the characteristic scale σ_{GS} of the correlation between GMCs and YSCs is about 85 pc; this matches well with the median distance of 76 pc from a GMC to its nearest neighbour in YSCCs. A slightly greater estimated value is largely due to the fact that we considered all possible assignments of a YSCC to a GMC. It is also similar to the general size of cloud-scale ($\lesssim$ 100 pc) star formation complexes (Chevance et al., 2019). However, compared to the correlation scale of 17 pc obtained by Corbelli et al. (2017), the difference is rather drastic. We suspect the difference might be due to (a) the fact that those authors did not adjust for the inclination of M33 in their analysis; (b) vastly different approaches in modelling framework; (c) those authors scaled the distance among GMCs and YSCCs to account for their large scale density variation when they fitted their model for the "positional correlation function" to describe the relationship between GMCs and YSCCs. However, they did not seem to account for the scaling when fitting the correlation length parameter. This seems rather unjustified and might have led to the drastic difference between our estimate and theirs. Nevertheless, a characteristic scale of 85 pc still shows a strong positive correlation between GMCs and YSCCs. Furthermore, it also means that the correlation strength between GMCs and YSCCs diminishes drastically as the separation distance increases.

For the slope parameters governing the effect of GMC properties on the correlation strength with YSCCs, we found that $\theta_D = 0.85$, $\theta_M = 0.66$, and $\theta_{gc} = -0.03$. For θ_D, the results show that the correlation strength increases by $\exp(0.85) = 2.3$ if the galactocentric distance of GMCs increases by 1 standard scale, which is about 1.55 kpc. This generally corresponds to the preliminary analysis on the cross-type PCF between GMCs and YSCCs obtained in Figure 4.6. To better compare our results to what Corbelli et al. (2017) obtained, we follow the procedure described by Corbelli

et al. (2017) and analyse the ratio between the "positional correlation function" of GMCs and YSCCs in the three zones constructed by Corbelli et al. (2017). We found that the maximum increase in the ratio is around 3 when moving from zone 1 ($D < 1.5$ kpc) to zone 2 (1.5 kpc $\leq D < 4$ kpc) and about 2 from zone 2 to zone 3 ($D \geq 4$ kpc). This is generally in line with what we have obtained, although differences in estimates diverge as the galactocentric distance increases. Again, this is likely due to the completely different approach in modelling since for simplicity, we considered the effect of galactocentric distance on the correlation strength as linear. This could be unrealistic across the galaxy disc. We will consider other forms of non-linear relationships in future work.

However, caution is needed in interpreting the physical meaning of θ_D since the GMCs and YSCCs also clump on the spiral arms and this might be a potential lurking variable that can influence the actual correlation between GMCs and YSCCs as noted by Corbelli et al. (2017). Nevertheless, the characteristic scale of the correlation at 85 pc still indicates strong evidence for the relationship between GMCs and YSCCs. We do not pursue the modelling of spiral arm structure since that can drive up model complexity and the model considered here already has eight parameters.

On another note, the strong positive effect of galactocentric distance on the correlation strength between GMCs and YSCCs leads us to make an important observation. As we have already seen in Figure 5.13, the outer region of the galaxy disc has a number of YSCC groups. Although we have pointed out that these groups do not have GMCs in their immediate surroundings (< 100 pc), a partial contribution to the high value of θ_D could come from the fact that these YSCCs groups all appear to be within 250–500 pc from GMCs. We note that this should not be caused by the crowding between GMCs and YSCCs in the spiral arms since (a) the scale at 250–500 pc is still relatively local for spiral arms to have any significant effect on the density variations of both GMCs and YSCCS; (b) YSCCs in the end need to have a birthplace and

they cannot show up out of nowhere simply because of the presence of spiral arms. The point of this observation is that these YSCCs groups not having GMCs in their surrounding at a distance similar on the order of σ_{GS} may have important physical implications for the formation and evolution of YSCCs. I will defer the discussion to section 5.2.3.

For θ_M, we see that the mass of GMCs also has a strong positive effect on the correlation strength between GMCs and YSCCs. Similar to the effect of the galactocentric distance, 1 standard scale ($2.1 \times \log_{10}(M_\odot)$) increase in the mass of a GMC can lead to a $\exp(0.66) = 1.9$ times increase in the correlation strength. This also corresponds to the finding by Corbelli et al. (2017) where they noted that 69% of the high-mass GMCs ($> 2 \times 10^5 M_\odot$) have a YSCC within 50 pc while only 44% of low-mass GMCs have an associated YSCC.

The distance from GMC to the CO filament structure may not seem to have any significant effect on the correlation strength between GMCs and YSCCs. However, the approximate posterior distribution of θ_M shows that 65% of the posterior samples are below 0. This, together with the estimated posterior mean at -0.045, shows that as GMCs break away from the CO filament structure, their correlation with the YSCCs tend to slightly decrease. This may indicate that the star formation activity is more fervent while GMCs are still part of the CO filament structure, although the effect might be minuscule.

Second-Order Potential

Based on the second-order potential and the results from model criticism, we confirm that there indeed exists repulsive behaviour between YSCs at short distances, as indicated by the matching of the NND distribution at short distances in Figure 5.10. The most important results we found are on the YSCC clustering behaviour in the outer region of the galaxy disc. As mentioned before, these groups of YSCCs are not

associated to any GMCs, but they are still generally within 200–500 pc from GMCs. This can involve several potential explanations that may shed light on the evolution of GMCs and YSCCs. Below I list three potential hypotheses that potentially explains the grouping behavior:

- There are undetected GMCs in the outer region of the galaxy
- YSCCs in the outer regions destroyed their natal GMCs
- YSCCs moved away from their natal GMCs

First, the grouping behavior of YSCCs in the outer region of the galaxy can serve as evidence for the hypothesis proposed by Corbelli et al. (2017). In their conclusion, they attributed the non-negligible disparity in the numbers of GMCs and YSCCs in the outer region to the presence of GMCs that are below detection limits, with some of the excess YSCCs born from these undetected GMCs. The detection of the grouping behaviour of YSCCs in the outer region in our analysis can support this hypothesis. If we assume similar levels of correlation between the undetected GMCs and YSCCs and some of these YSCCs are still associated with undetected GMCs, then these GMCs will strongly affect the position of the "unclaimed" YSCCs and these YSCCs will likely group around the undetected GMCs. However, due to the detection limit, these GMCs are not considered in the data, hence the model cannot account for their effect on the YSCCs, which is reflected by the grouping behaviour demonstrated in our analysis. Furthermore, the results from Figure 5.10 also seem to point in the direction of the undetected GMCs hypothesis. The recent study by Chevance et al. (2019) analysed the cloud-scale star formation complexes (including GMCs and associated SCs) in nine spiral galaxies. They found that the general mean separation distance between individual star formation complexes is roughly $\sim 100-300$ pc. This corresponds to the scale of 250 pc at which the peak of discrepancy occurs between NND distributions of the data and our model as shown in Figure 5.10. If these YSCCs

are indeed associated with undetected GMCs that are separated by 250 pc on average, then it explains the discrepancy in Figure 5.10.

To assess the plausibility of the hypothesis of undetected GMCs, we turn to the original paper of Druard et al. (2014) where the GMC observations are reported. As the GMCs are detected through the CO(2-1) emission line, a useful piece of information is the noise map of CO(2-1) observations presented in Figure 6 of Druard et al. (2014). Although there is noise variation across the galaxy disc, it is almost negligible and the noise map is in general quite homogeneous. If we compare the region with the highest noise level with the region with underestimated intensity in Figure 5.8, the high noise region does not have significant overlap with the blue blocks in Figure 5.8. Furthermore, the high noise region in fact has detected GMCs. If we assume that the CO intensity from GMCs is on a similar level in the outer region, the above comparison does not seem to support the hypothesis of undetected GMCs.

Another potential cause for the undetected GMCs is the variation of the "X-factor" between H_2 and CO mentioned in Chapter 4. In reality, this factor might not be constant and could vary with the galactocentric distance. In general, the X-factor is supposedly inversely proportional to the metallicity (abundance of elements heavier than hydrogen and helium), i.e., as metallicity increases, X-factor increases and we would infer the existence of more H_2 for the same level of CO intensity. This implies that GMCs in the outer region should in fact be more detectable, not less, than in the inner region, since it is generally accepted that there is a negative relationship between metallicity and the galactocentric distance. This again does not seem to support the hypothesis that there are undetected GMCs in the outer regions of the galaxy.

To eventually test the hypothesis of undetected GMCs, targeted high sensitivity observations of CO(2-1) emission line in the outskirts of M33 are needed. The residual field in Figure 5.13 in fact gives a map which can narrow down the region for the

pointed observations: they can simply be made at the regions with the most underestimation in the intensity of YSCCs. This is another demonstration of the power of GPP modelling.

As demonstrated in the previous arguments, the hypothesis of undetected GMCs does not seem to hold fast. In the case that the targeted observation for undetected GMCs turn out to be unsuccessful, other explanations are needed to explain the grouping behavior of YSCCs. I will provide the details of two hypotheses alternative to that of undetected GMCs.

Firstly, the grouping behavior of YSCCs can be caused by them destroying their natal clouds. Corbelli et al. (2017) concluded that GMCs in M33 tend to have a very short lifetime, around 14.2 Myr. Chevance et al. (2019) also estimated the lifetime of GMCs in nine nearby galaxies and found that they average $\sim 10 - 30$ Myr. They found that, in general, GMCs in these galaxies spent most of their lifetime ($\sim$75-90%) dormant but quickly disperse in $\sim 1 - 5$ Myr once the stars are formed, likely due to stellar winds. The study by Kruijssen et al. (2019) in NGC 300 found evidence of a rapid evolutionary cycle among GMCs and star formation, with GMCs destruction in less than 1.5 Myr by efficient stellar feedback.

A simple deduction can be made that if GMCs are of low mass, their destruction should be even more rapid. Corbelli, Braine, and Giovanardi (2019) analyzed the variation of mass of GMCs versus galactocentric distance and found that the mass of GMCs drops as galactocentric distance increases. They concluded that the presence of high mass GMCs in the inner disc of M33 ($D < 3.9$ kpc) is likely due to the supersonic rotation of the disc in the inner region where the gas is collected by the spiral arms and forms more massive clouds. However, this is not the case beyond the co-rotation region ($D > 4.7$ kpc) where the much slower rotation results in low mass GMCs. The co-rotation distance of 4.7 kpc corresponds to our observation of grouping of YSCCs beyond 4.5 kpc, and if we assume that GMCs in the outer region

belong to the low mass class ($\lesssim 10^5 M_\odot$), then a possible explanation for the absence of GMCs might be the formation of YSCCs and their efficient stellar feedback leading to the destruction of their low mass natal clouds.

Hollyhead et al. (2015) shows that young massive clusters in M83 generally break out of their natal clouds around 4 Myr. Corbelli et al. (2017) also analyzed the association between GMCs and another catalog of optically visible SCs by Fan and Grijs (2014) in M33 with a wider range of age estimates, ranging from 5 Myr to 10 Gyr. Although the correlations between these SCs and GMCs are much weaker than the ones found in this study, the correlations are still stronger than that of a Poisson process. This means that the time scale for SCs to disperse into a Poisson-like structure is much longer than the cloud life-time as suggested in previous studies. This indicates that the grouping behaviour of YSCCs in the outer region is potentially a result of YSCCs destroying their natal clouds before they have had time to disperse and appear Poisson-like. To test this hypothesis, we would need data on the age of these YSCCs. Age estimates are only available for 402 out of the 630 YSCCs with a mean estimate at ∼5 Myr. If using the results for GMC dispersal time ($1 \sim 5$ Myr) after star formation from previous studies (Chevance et al., 2019; Kruijssen et al., 2019), many GMCs might have just been destroyed by the newly formed SCs through stellar winds. This is even more probable if the destruction of low mass GMCs is more rapid than $\sim 1-5$ Myr. However, the age estimates of the YSCCs are rather imprecise and should not be used in general to draw any definitive conclusion.

Another potential process involved in the appearance of the clustering might be that numerous YSCCs are in fact generated by the same GMC. As these YSCCs break out and lose their association with their original GMCs, they might have similar velocity due to their common birthplace. Since they are all in the early stage of their evolution, they tend to move in the same direction before starting to disperse independently. Furthermore, analysis by Grasha et al. (2019) suggests that on average

SCs in the M51 galaxy that are not associated with any GMCs are much older ($\sim$ 50 Myr) compared to those that are associated with a GMC ($\sim$ 4 Myr). Assuming the star formation process is generally universal, this observation can be indirect evidence to support the hypothesis that the YSCCs in the outskirts are moving away from their natal clouds. This hypothesis also tends to explain the fact that most GMCs in the outer region of M33 tend to have low mass and the disparity in number between GMCs and YSCCs in the outer region, a potential indication that they may have produced enough YSCCs and are almost at the end of their life-cycle. However, to test this hypothesis, we would need more accurate estimates of the age of YSCCs to analyze the correlation between GMCs and YSCCs as a function of the age of YSCCs. If the association weakens, this would serve as evidence in support of the hypothesis.

In conclusion, the formation of SCs may be a combination of the processes mentioned above and further detailed study needs to be done to paint a clear picture. Nevertheless, the results we have obtained here clearly showcase the power of GPP modelling in its effectiveness and sensitivity on numerically identifying detailed structure and behavior exhibited by highly inhomogeneous point patterns. The identification of groups of YSCCs in the outer region would not be possible using the previous exploratory statistical tools of 2PCF/PCF and its variants, and has led to evidence for suggesting previous hypothesis and providing new possible hypotheses on the evolution of stellar populations.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, Gibbs point process models are constructed to provide a novel methodology for probing the spatial distributions of and relationships between objects, including the CO filament structure, GMCs, and YSCCs, in the star formation complexes of the M33 galaxy. These models provide a sensitive and rigorous approach to understand the highly inhomogeneous distribution of stellar populations. They also enable the investigation of multiple scientific questions in an integrated manner.

To investigate the spatial distribution and relationship among the CO filament structure, GMCs and YSCCs, a hierarchical Gibbs point process model structure is employed. The GMCs are assumed to be the high-level process in the hierarchy where YSCCs are considered the low-level process. This hierarchical structure instantiates the natural formation hierarchy among GMCs and YSCCs. Two univariate models stem from the hierarchical model where the CO-GMC model corresponds to the high-level process of GMCs and GMC-SC model for the low-level process of YSCCs.

For the CO-GMC model, we reached the following conclusions regarding the distribution of GMCs in M33:

- There is an approximately 492 times increase of intensity of GMCs in the presence of the CO filament structure compared to a unit-rate Poisson process (on average 1 GMC/kpc^2). This provides ample evidence on the formation origin of GMCs.

- There exist two sub-populations of GMCs with respect to the distance from GMC to the CO filament, with the main sub-population being tightly correlated with the CO filament structure and the minor sub-population being much less so.

- The second-order characteristic of the distribution of GMCs indicates that the typical separation between a pair of GMCs is approximately 130 pc, indicating a repulsive structure at the local scale. This also corresponds to the typical star formation complex separation distance found in spiral galaxies.

- From the model diagnostics, we conclude that the CO filament structure exerts an inhomogeneous effect on the intensity of GMCs. This inhomogeneous effect is two-fold. First, it can be due to the inhomogeneous CO intensity at different points on the CO filament and this eventually leads to an inhomogeneous distribution of GMCs. Second, it potentially reveals another formation mechanism for GMCs where the field of the interstellar medium away from the CO filament is fueling the formation of GMCs. This also tends to explain the significantly less massive GMCs in the minor sub-population that are much less correlated with the CO filament.

For the GMC-SC model, the following results and conclusions are obtained:

- GMCs have a significant impact on the distribution of YSCCs, where the presence of GMC will increase the intensity of YSCCs by 90 times on average.

However, the impact is rather limited as the characteristic correlation scale between GMCs and YSCCs is relatively local, with an estimate of 85 pc. This also corresponds to the typical cloud scale of $\lesssim 100$ pc.

- The intrinsic properties of GMCs also have strong effects on the distribution of YSCCs. We found that every 1.5 kpc increase in the galactocentric distance of GMCs leads to a 2.3 times of increase in the correlation strength between GMCs and YSCCs. Every $2.1 \times log_{10}(M_{\odot})$ increase in the GMC mass leads to a 1.9 times increase in the correlation strength. The distance from a GMC to the CO filament structure, however, does not have a significant impact on the correlation strength.

- The second-order behaviour of YSCCs shows that they are also repulsive at the local scale, with an estimated repulsive scale of approximately 80 pc. This corresponds to the stellar feedback from SCs that generally suppresses and regulates the star formation in their immediate surroundings.

- Model diagnostics provide interesting and crucial information on the formation process of YSCCs. We found that there exists second-order clustering of GMCs in the outer region of the galaxy disc ($D \geq 4.5$ kpc) that cannot be explained by the inhomogeneity in the first-order intensity. This can be attributed to three potential causes:

    1. There exist undetected GMCs that give rise to the unexplained second-order clustering behaviour. However, evidence suggests that this is less likely to be the case.

    2. The YSCCs destroyed their natal GMCs. However, due to their recent birth, YSCCs did not have enough time to diffuse and appear Poisson distributed.

3. Groups of YSCCs formed in the same clouds but started moving away from their natal clouds. Due to the same origin, they tend to have similar velocity and did not have enough time to disperse and appear Poisson-like.

The above three hypotheses can be confirmed by more resolved observations of GMCs and better data regarding the age and velocity of YSCCs.

In general, we can see the immense power demonstrated by Gibbs point process modelling which provides a rigorous method to obtain accurate numerical measurements on the distribution of investigated objects. This subsequently leads us to discover structures and propose new hypotheses that are otherwise impossible.

6.2 Future Work

Due to the limitation on available data, this study only considered an individual galaxy. For future work, we would like to consider applying the model and exploring its applicability to other galaxies.

Although Gibbs point process modelling is a highly flexible and extremely interpretable model for investigating spatial data, it is not without its own faults. As we have seen in this study, Gibbs point process modelling usually does not provide analytical solutions to the expected value of intensity of a point process, which can be restricting in many scenarios. Furthermore, if the parameter space becomes high-dimensional ($d > 10$), inference procedure can be extremely challenging. To curb the above issues, it would be interesting to see the applicability of log-Gaussian Cox process mentioned in Chapter 1. Certainly, this will bring corresponding challenges of interpretation. Another direction is to invent new inference algorithms that can enable efficient Bayesian computation for high-dimensional Gibbs point process. Last but not least, there can potentially be second-order non-stationarity in the point pat-

tern investigated. A direction for improvement is to explore a modelling approach which can account for second-order non-stationary processes.

References

[1] A. Baddeley et al. "Residual analysis for spatial point processes (with discussion)". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.5 (2005), pp. 617–666.

[2] A. J. Baddeley and M. N. M. van Lieshout. "Area-interaction point processes". In: *Annals of the Institute of Statistical Mathematics* 47.4 (Dec. 1995), pp. 601–619.

[3] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, Nov. 2015. ISBN: 978-1-4822-1020-0.

[4] Adrian Baddeley and Rolf Turner. "Practical Maximum Pseudolikelihood for Spatial Point Patterns". In: *Australian & New Zealand Journal of Statistics* 42.3 (2000), pp. 283–322.

[5] "Spatial Point Processes and their Applications". In: *Stochastic Geometry: Lectures given at the C.I.M.E. Summer School held in Martina Franca, Italy, September 13–18, 2004*. Ed. by Adrian Baddeley et al. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer, 2007, pp. 1–75. ISBN: 978-3-540-38175-4.

[6] R. B. Blackman and J. W. Tukey. "The Measurement of Power Spectra from the Point of View of Communications Engineering — Part I". In: *Bell System Technical Journal* 37.1 (1958), pp. 185–282.

[7] A. Z. Bonanos et al. "The First Direct Distance to a Detached Eclipsing Binary in M33". In: *Astrophysics and Space Science* 304.1 (Aug. 2006), pp. 207–209.

[8] Anders Brix and Jesper Moller. "Space-time Multi Type Log Gaussian Cox Processes with a View to Modelling Weeds". In: *Scandinavian Journal of Statistics* 28.3 (2001), pp. 471–488.

[9] Stephen P. Brooks and Andrew Gelman. "General Methods for Monitoring Convergence of Iterative Simulations". In: *Journal of Computational and Graphical Statistics* 7.4 (Dec. 1998), pp. 434–455.

[10] R. G. Carlberg and R. E. Pudritz. "Magnetic Support and Fragmentation of Molecular Clouds". In: *Monthly Notices of the Royal Astronomical Society* 247 (Dec. 1990), p. 353.

[11] Mélanie Chevance et al. "The lifecycle of molecular clouds in nearby star-forming disc galaxies". In: *Monthly Notices of the Royal Astronomical Society* (Dec. 2019).

[12] M.-R. L. Cioni. "The metallicity gradient as a tracer of history and structure: the Magellanic Clouds and M33 galaxies". In: *Astronomy & Astrophysics* 506.3 (Nov. 2009), pp. 1137–1146.

[13] Edvige Corbelli, Jonathan Braine, and Carlo Giovanardi. "Rise and fall of molecular clouds across the M 33 disk". In: *Astronomy and Astrophysics* 622 (Feb. 2019), A171.

[14] Edvige Corbelli et al. "From molecules to young stellar clusters: the star formation cycle across the disk of M 33". In: *Astronomy & Astrophysics* 601 (May 2017), A146.

[15] D. R. Cox. "Some Statistical Methods Connected with Series of Events". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 17.2 (1955), pp. 129–157.

[16] "Special Classes of Processes". In: *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*. Ed. by D. J. Daley and D. Vere-Jones. Probability and Its Applications. New York, NY: Springer, 2008, pp. 76–130. ISBN: 978-0-387-49835-5.

[17] Marc Davis, Amber Miller, and Simon D. M. White. "A Galaxy-weighted Measure of the Relative Peculiar-Velocity Dispersion". In: *The Astrophysical Journal* 490.1 (Nov. 1997), p. 63.

[18] Herwig Dejonghe. "A completely analytical family of anisotropic Plummer models". In: *Monthly Notices of the Royal Astronomical Society* 224.1 (Jan. 1987), pp. 13–39.

[19] C. Druard et al. "The IRAM M 33 CO(2–1) survey - A complete census of molecular gas out to 7 kpc". In: *Astronomy & Astrophysics* 567 (July 2014), A118.

[20] Bruce G. Elmegreen and John Scalo. "Interstellar Turbulence I: Observations and Processes". In: *Annual Review of Astronomy and Astrophysics* 42.1 (2004), pp. 211–273.

[21] Zhou Fan and Richard de Grijs. "Star Clusters in M33: Updated UBVRI Photometry, Ages, Metallicities, and Masses". In: *The Astrophysical Journal Supplement Series* 211.2 (Mar. 2014), p. 22.

[22] Christoph Federrath, Ralf S. Klessen, and Wolfram Schmidt. "The Fractal Density Structure in Supersonic Isothermal Turbulence: Solenoidal versus Compressive Energy Injection". In: *The Astrophysical Journal* 692.1 (Feb. 2009), pp. 364–374.

[23] Andrew Gelman and Donald B. Rubin. "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4 (Nov. 1992), pp. 457–472.

[24] Hans-Otto Georgii. "Canonical and grand canonical Gibbs states for continuum systems". In: *Communications in Mathematical Physics* 48.1 (1976), pp. 31–51.

[25] Charles J. Geyer. "Markov Chain Monte Carlo Maximum Likelihood". In: Interface Foundation of North America, 1991.

[26] Charles J. Geyer and Jesper Møller. "Simulation Procedures and Likelihood Inference for Spatial Point Processes". In: *Scandinavian Journal of Statistics* 21.4 (1994), pp. 359–373.

[27] Josiah Willard Gibbs. *Elementary Principles in Statistical Mechanics*. New York: Charles Scribner's Sons, Mar. 1902.

[28] Philipp Girichidis et al. "Importance of the initial conditions for star formation – III. Statistical properties of embedded protostellar clusters". In: *Monthly Notices of the Royal Astronomical Society* 420.4 (Mar. 2012), pp. 3264–3280.

[29] Jacqueline I. Goldstein et al. "An attraction-repulsion point process model for respiratory syncytial virus infections." In: *Biometrics* 71.2 (2014), pp. 376–385.

[30] Michel Goulard, Aila Särkkä, and Pavel Grabarnik. "Parameter Estimation for Marked Gibbs Point Processes through the Maximum Pseudo-Likelihood Method". In: *Scandinavian Journal of Statistics* 23.3 (1996), pp. 365–379.

[31] K. Grasha et al. "The Hierarchical Distribution of the Young Stellar Clusters in Six Local Star-forming Galaxies". In: *The Astrophysical Journal* 840.2 (May 2017), p. 113.

[32] K. Grasha et al. "The Spatial Distribution of the Young Stellar Clusters in the Star-Forming Galaxy NGC 628". In: *The Astrophysical Journal* 815.2 (Dec. 2015), p. 93.

[33] K. Grasha et al. "The spatial relation between young star clusters and molecular clouds in M51 with LEGUS". In: *Monthly Notices of the Royal Astronomical Society* 483.4 (Mar. 2019), pp. 4707–4723.

[34] Dávid Guszejnov, Philip F. Hopkins, and Mark R. Krumholz. "Protostellar feedback in turbulent fragmentation: consequences for stellar clustering and multiplicity". In: *Monthly Notices of the Royal Astronomical Society* 468.4 (July 2017), pp. 4093–4106.

[35] Heikki Haario, Eero Saksman, and Johanna Tamminen. "An adaptive Metropolis algorithm". In: *Bernoulli* 7.2 (Apr. 2001), pp. 223–242.

[36] K. Hollyhead et al. "Studying the YMC population of M83: how long clusters remain embedded, their interaction with the ISM and implications for GC formation theories". In: *Monthly Notices of the Royal Astronomical Society* 449 (May 2015), pp. 1106–1117.

[37] Philip F. Hopkins, Desika Narayanan, and Norman Murray. "The meaning and consequences of star formation criteria in galaxy models with resolved stellar feedback". In: *Monthly Notices of the Royal Astronomical Society* 432 (July 2013), pp. 2647–2653.

[38] H. Högmander and Aila Särkkä. "Multitype spatial point patterns with hierarchical interactions". In: *Biometrics* 55.4 (Dec. 1999), pp. 1051–1058.

[39] Valerie Isham. "Multitype Markov Point Processes: Some Approximations". In: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 391.1800 (1984), pp. 39–53.

[40] Ernst Ising. "Beitrag zur Theorie des Ferromagnetismus". In: *Zeitschrift für Physik* 31.1 (Feb. 1925), pp. 253–258.

[41] Don H. Johnson. "Point process models of single-neuron discharges". In: *Journal of Computational Neuroscience* 3.4 (Dec. 1996), pp. 275–299.

[42] Eric W. Koch and Erik W. Rosolowsky. "Filament identification through mathematical morphology". In: *Monthly Notices of the Royal Astronomical Society* 452.4 (Oct. 2015), pp. 3435–3450.

[43] J. M. Diederik Kruijssen et al. "Fast and inefficient star formation due to short-lived molecular clouds and rapid feedback". In: *Nature* 569.7757 (May 2019). arXiv: 1905.08801, pp. 519–522.

[44] Mark R. Krumholz. "The big problems in star formation: The star formation rate, stellar clustering, and the initial mass function". In: *Physics Reports*. The Big Problems in Star Formation: the Star Formation Rate, Stellar Clustering, and the Initial Mass Function 539.2 (June 2014), pp. 49–134.

[45] Michael Krumin and Shy Shoham. "Generation of spike trains with controlled auto- and cross-correlation functions". In: *Neural Computation* 21.6 (June 2009), pp. 1642–1664.

[46] Aleksandra Kuznetsova, Lee Hartmann, and Javier Ballesteros-Paredes. "Kinematics and structure of star-forming regions: insights from cold collapse models". In: *Monthly Notices of the Royal Astronomical Society* 473.2 (Jan. 2018), pp. 2372–2377.

[47] Charles J. Lada and Elizabeth A. Lada. "Embedded Clusters in Molecular Clouds". In: *Annual Review of Astronomy and Astrophysics* 41.1 (2003), pp. 57–115.

[48] David Lando. "On cox processes and credit risky securities". In: *Review of Derivatives Research* 2.2 (Dec. 1998), pp. 99–120.

[49] Thomas J. Leininger and Alan E. Gelfand. "Bayesian Inference and Model Assessment for Spatial Point Patterns Using Posterior Predictive Samples". In: *Bayesian Analysis* 12.1 (Mar. 2017), pp. 1–30.

[50] Ye Li et al. "Log Gaussian Cox processes and spatially aggregated disease incidence data". In: *Statistical Methods in Medical Research* 21.5 (Oct. 2012), pp. 479–507.

[51] Faming Liang. "A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants". In: *Journal of Statistical Computation and Simulation* 80.9 (Sept. 2010), pp. 1007–1022.

[52] M. N. M. Van Lieshout. *Markov Point Processes and Their Applications*. Google-Books-ID: e_tpDQAAQBAJ. World Scientific, 2000. ISBN: 978-1-86094-071-2.

[53] Finn Lindgren, Håvard Rue, and Johan Lindström. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4 (2011), pp. 423–498.

[54] Laura Magrini, Letizia Stanghellini, and Eva Villaver. "The Planetary Nebula Population of M33 and Its Metallicity Gradient: A Look into the Galaxy's Distant Past". In: *The Astrophysical Journal* 696.1 (Apr. 2009), pp. 729–740.

[55] Christopher F. McKee and Eve C. Ostriker. "Theory of Star Formation". In: *Annual Review of Astronomy and Astrophysics* 45.1 (2007), pp. 565–687.

[56] Dean E. McLaughlin and Ralph E. Pudritz. "The Formation of Globular Cluster Systems. I. The Luminosity Function". In: *The Astrophysical Journal* 457 (Feb. 1996), p. 578.

[57] Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. "MCMC for Doubly-intractable Distributions". In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. UAI'06. event-place: Cambridge, MA, USA. Arlington, Virginia, United States: AUAI Press, 2006, pp. 359–366. ISBN: 978-0-9749039-2-7.

[58] J. Møller et al. "An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants". In: *Biometrika* 93.2 (2006), pp. 451–458.

[59] Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. "Log Gaussian Cox Processes". In: *Scandinavian Journal of Statistics* 25.3 (1998), pp. 451–482.

[60] Jesper Møller and Rasmus Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, Sept. 2003. ISBN: 978-1-58488-265-7.

[61] Jesper Møller and Rasmus P. Waagepetersen. "Modern Statistics for Spatial Point Processes*". In: *Scandinavian Journal of Statistics* 34.4 (2007), pp. 643–684.

[62] Huu-Giao Nguyen, Ronan Fablet, and Jean-Marc Boucher. "Visual textures as realizations of multivariate log-Gaussian Cox processes". In: *CVPR 2011*. ISSN: 1063-6919, 1063-6919, 1063-6919. June 2011, pp. 2945–2952.

[63] Jaewoo Park and Murali Haran. "Bayesian Inference in the Presence of Intractable Normalizing Functions". In: *Journal of the American Statistical Association* 113.523 (July 2018), pp. 1372–1390.

[64] P. J. E. Peebles. *Principles of Physical Cosmology*. Princeton University Press, May 1993. ISBN: 978-0-691-01933-8.

[65] P. J. E. Peebles. "The Galaxy and Mass N-Point Correlation Functions: a Blast from the Past". In: *ASP Conference Proceedings*. Vol. 252. San Francisco: Astronomical Society of the Pacific, 2001, p. 201.

[66] P. J. E. Peebles. *The large-scale structure of the universe*. Princeton series in physics. OCLC: 6421704. Princeton, N.J: Princeton University Press, 1980. ISBN: 978-0-691-08239-4 978-0-691-08240-0.

[67] Nicolas Picard et al. "The Multi-scale Marked Area-interaction Point Process: A Model for the Spatial Pattern of Trees". In: *Scandinavian Journal of Statistics* 36.1 (2009), pp. 23–41.

[68] H. C. Plummer. "On the problem of distribution in globular star clusters". In: *Monthly Notices of the Royal Astronomical Society* 71 (Mar. 1911), pp. 460–470.

[69] Simon F. Portegies Zwart, Stephen L.W. McMillan, and Mark Gieles. "Young Massive Star Clusters". In: *Annual Review of Astronomy and Astrophysics* 48.1 (2010), pp. 431–493.

[70] T. Rajala, D. J. Murrell, and S. C. Olhede. "Detecting multivariate interactions in spatial point patterns with Gibbs models and variable selection". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.5 (2018), pp. 1237–1273.

[71] B. D. Ripley and F. P. Kelly. "Markov Point Processes". In: *Journal of the London Mathematical Society* s2-15.1 (1977), pp. 188–192.

[72] Mark D. Risser. "Review: Nonstationary Spatial Modeling, with Emphasis on Process Convolution and Covariate-Driven Approaches". In: *arXiv:1610.02447 [stat]* (Oct. 2016). arXiv: 1610.02447.

[73] Gareth O. Roberts and Jeffrey S. Rosenthal. "Coupling and Ergodicity of Adaptive Markov Chain Monte Carlo Algorithms". In: *Journal of Applied Probability* 44.2 (2007), pp. 458–475.

[74] Gareth O. Roberts and Jeffrey S. Rosenthal. "Examples of Adaptive MCMC". In: *Journal of Computational and Graphical Statistics* 18.2 (Jan. 2009), pp. 349–367.

[75] Alexandre Rodrigues and Peter J. Diggle. "Bayesian Estimation and Prediction for Inhomogeneous Spatiotemporal Log-Gaussian Cox Processes Using Low-

Rank Models, With Application to Criminal Surveillance". In: *Journal of the American Statistical Association* 107.497 (Mar. 2012), pp. 93–101.

[76] H. Rogers and J. M. Pittard. "Feedback from winds and supernovae in massive stellar clusters – I. Hydrodynamics". In: *Monthly Notices of the Royal Astronomical Society* 431.2 (May 2013), pp. 1337–1351.

[77] Jeffrey Rosenthal. "Optimal Proposal Distributions and Adaptive MCMC". In: *Handbook of Markov Chain Monte Carlo*. Ed. by Steve Brooks et al. Vol. 20116022. Chapman and Hall/CRC, May 2011. ISBN: 978-1-4200-7941-8 978-1-4200-7942-5.

[78] Håvard Rue and Leonhard Held. *Gaussian Markov Random Fields: Theory and Applications*. 1st ed. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, Feb. 2005. ISBN: 978-1-58488-432-3.

[79] David Ruelle. *Statistical mechanics: Rigorous Results.* New York, NY: New York : W. A. Benjamin, 1969. ISBN: 978-981-02-3862-9.

[80] Pantelis Samartsidis et al. "Bayesian log-Gaussian Cox process regression: applications to meta-analysis of neuroimaging working memory studies". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68.1 (2019), pp. 217–234.

[81] Laura Serra et al. "Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: the case of Catalonia, 1994–2008". In: *Environmental and Ecological Statistics* 21.3 (Sept. 2014), pp. 531–563.

[82] S. Sharma et al. "The population of young stellar clusters throughout the disk of M 33". In: *Astronomy & Astrophysics* 534 (Oct. 2011), A96.

[83] Shinichiro Shirota and Alan E. Gelfand. "Space and circular time log Gaussian Cox processes with application to crime event data". In: *The Annals of Applied Statistics* 11.2 (June 2017), pp. 481–503.

[84] D. Simpson et al. "Going off grid: computationally efficient inference for log-Gaussian Cox processes". In: *Biometrika* 103.1 (Mar. 2016), pp. 49–70.

[85] E. Tempel et al. "Bisous model—Detecting filamentary patterns in point processes". In: *Astronomy and Computing* 16 (July 2016), pp. 17–25.

[86] H. J. de Vega, N. Sánchez, and F. Combes. "Self-gravity as an explanation of the fractal structure of the interstellar medium". In: *Nature* 383.6595 (Sept. 1996), pp. 56–58.

[87] Rasmus Waagepetersen et al. "Analysis of multispecies point patterns by using multivariate log-Gaussian Cox processes". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 65.1 (2016), pp. 77–96.

[88] Benjamin D. Wandelt. "Gaussian Random Fields in Cosmostatistics". In: *Astrostatistical Challenges for the New Astronomy*. Ed. by Joseph M. Hilbe. Springer Series in Astrostatistics. New York, NY: Springer, 2013, pp. 87–105. ISBN: 978-1-4614-3508-2.

[89] Nguyen Xuan Xanh and Hans Zessin. "Integral and Differential Characterizations of the GIBBS Process". In: *Mathematische Nachrichten* 88.1 (1979), pp. 105–115.

CURRICULUM VITAE

Name	Dayi Li
Post-Secondary Education and Degrees:	Western University London, ON 2018 - 2020 M.Sc.
	Western University London, ON 2014 - 2018 H.B.Sc
Honours and Awards	Western Graduate Research Scholarship 2018-2020
	Western Gold Medalist in Financial Modelling 2018
	Western Continuing Scholarship 2014-2018
	Jane-Plas International Student Scholarship 2015
Related Work Experience	Teaching Assistant Western University 2018 - 2020