Spatial clustering of gravitational wave sources with k-nearest neighbour distributions

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Kaustubh Rajesh Gupta

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

May, 2024

Supervisor: Dr. Arka Banerjee

Certificate

This is to certify that this dissertation entitled **Spatial clustering of gravitational wave sources with k-nearest neighbour distributions** towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Kaustubh Rajesh Gupta at Indian Institute of Science Education and Research under the supervision of Dr. Arka Banerjee, Assistant Professor, Department of Physics , during the academic year 2023-2024.

Dr. Arka Banerjee

Committee:

Dr. Arka Banerjee

Dr. Diptimoy Ghosh

This thesis is dedicated to my family.

Declaration

I hereby declare that the matter embodied in the report entitled **Spatial clustering of gravitational wave sources with k-nearest neighbour distributions** are the results of the work carried out by me at the Department of Physics, Indian Institute of Science Education and Research (IISER) Pune, under the supervision of Dr. Arka Banerjee , and the same has not been submitted elsewhere for any other degree. Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Kaustubh Rajesh Gupta

Roll Number 20191179

Acknowledgements

I would like to thank my supervisor, Dr. Arka Banerjee, for being a wonderful mentor. I am extremely grateful to him for allowing me to work independently and with maximum flexibility while ensuring that I have sufficient guidance, and for his immense patience whenever I blunder. I was often overwhelmed by the myriad of challenges I faced in the project, but I could always count on his encouraging attitude and unwavering support to get me through the uncertainty and chaos that often results from research. I would also like to thank Dr. Diptimoy Ghosh for his valuable time and support as an expert supervisor.

I would like to thank Eishica, Kwanit, Vikhyat, Harrsh, Shubhankar and Yash for their thoughtful questions, discussions and insightful suggestions at the group meetings each week. I thank Dr. Aditya Vijaykumar, Dr. Shasvath Kapadia, Dr. Prayush Kumar, Aditya Sharma and Mukesh Kumar Singh for useful discussions, particularly for their guidance in creating the mock BBH catalogues. I would also like to thank Dr. Susmita Adhikari for useful discussions on generating mock data for the forecast study. I am grateful to Dr. Yuuki Omori for providing access to the simulation products from the Agora lightcone. The support and the resources provided by PARAM Brahma Facility under the National Supercomputing Mission, Government of India at the Indian Institute of Science Education and Research; Pune are gratefully acknowledged.

I am grateful to my two best friends, Ravish Mehta and Pranav Maheshwari, for always being there for me, through thick and thin. Your presence made my best experiences richer and your constant moral support made my worst moments bearable. My five years at IISER wouldn't have been the same without you. I am immensely grateful to Amogh, Soumil, Varun, Sugat and the entire Physics Mimamsa team; you were like a family to me during my 5 years at IISER Pune. This acknowledgement would not be complete without thanking my family, who made me the person I am today. I am thankful to my younger brother, Karan; I can always count on you to cheer me up whenever I feel down. Finally, I would like to thank my parents, although no words are

enough to describe my gratitude towards them. Without their constant encouragement to explore my dreams, however unconventional, a career in research would not be possible for me. Thank you for supporting me in my decision to take the road less travelled, and for always believing in me, even when I seem lost.

Abstract

This thesis presents a framework to quantify the clustering of gravitational wave (GW) transient sources and measure their spatial cross-correlation with the large-scale structure of the universe using k-nearest neighbour (kNN) distributions and two-point summary statistics. We extend the kNN formalism, initially developed to study 3D clustering in cartesian coordinates, to 2D clustering in angular coordinates. As a first application to data, we measure the nearest-neighbour distributions of 53 suitably selected Binary Black Hole (BBH) mergers detected in the first three observation runs of LIGO-Virgo-KAGRA and cross-correlate these sources with $\sim 1.7 \times 10^7$ galaxies and quasars from the WISE$\times$SuperCOSMOS all-sky catalogue. To determine the significance of the clustering signal while accounting for observational systematics in the GW data, we create 135 realisations of mock BBHs that are statistically similar to the observed BBHs but spatially unclustered. We find no evidence for spatial clustering or cross-correlation with large-scale structure in the data and conclude that the present sky localisation and number of detections are insufficient to get a statistically significant clustering signal. As a second application of our analysis framework, we investigate the feasibility of detecting the BBH-galaxy cross-correlation with future GW observing runs and stage-IV large-scale structure surveys. We forecast 10 years of GW observations with a network of 5 ground-based detectors consisting of 3 advanced LIGO detectors (Hanford, Livingston, India) operating at A+ sensitivity and Virgo, KAGRA operating at design sensitivity. The resulting BBH catalogue consists of $\sim 2.8 \times 10^4$ BBHs, of which $\sim 1.6 \times 10^4$ have a 68% credible sky localisation area less than 50 sq. deg. We cross-correlate these modestly well-localised BBHs with the simulated galaxy overdensity field of an LSST Y1-like survey and find that the second nearest neighbour distribution captures a nearly statistically significant cross-correlation signal at $\sim 1^\circ$ angular scales. We further show that this signal is not measured by the two-point cross-correlation function, demonstrating the ability of the nearest neighbour distributions to extract higher-order, non-Gaussian clustering from the small spatial scales accessible with upcoming GW observations and large-scale surveys that makes them more robust measures of spatial clustering than two-point clustering statistics that capture only the Gaussian clustering on all scales.

Contents

List of Tables

List of Figures

Declaration on Research Contribution

A part of the research conducted for this thesis has resulted in an arXiv pre-print titled *Spatial clustering of gravitational wave sources with k-nearest neighbour distributions* (Gupta & Banerjee, 2024). Some results from this thesis are part of upcoming research work (Gupta & Banerjee 2024 in prep.).

Chapter 1

Introduction

Large-scale surveys of the universe reveal a beautiful structure in the spatial distribution of galaxies (Adelman-McCarthy et al., 2006; Strauss et al., 2002; Stoughton et al., 2002; Falco et al., 1999; Geller & Huchra, 1989; Huchra et al., 1999; Davis et al., 1982), implying that the constituents of the universe are not distributed randomly, but are inherently clustered. The clustering properties of objects, such as galaxies, that trace this structure formation contain a wealth of information that can be used to test our understanding of cosmology and probe new physics beyond the standard model (see, e.g., Fumagalli, A. et al., 2024; Amon et al., 2023; DES Collaboration et al., 2023; Miyatake et al., 2023; Dvornik et al., 2023; DES Collaboration et al., 2022a,b,c). The detection of gravitational waves by the LIGO-Virgo-KAGRA (LVK) collaboration (LIGO Scientific Collaboration and Virgo Collaboration et al., 2016) has unveiled potential new tracers of structure formation in the form of merging binaries of compact stellar remnants such as black holes and neutron stars.

Gravitational waves allow a direct measurement of the luminosity distance to their sources (see, e.g., Holz et al., 2018; Holz & Hughes, 2005; Schutz, 1986) without the need for a hierarchical distance ladder or an empirical calibration process, with the only fundamental assumption being that general relativity is valid, making merging compact binaries 'standard sirens'. Hence, gravitational waves provide a mechanism to study the expansion history of our universe if the source redshifts can be estimated. This led Schutz (1986) to suggest that merging compact binaries can be used to constrain the Hubble-Lemaître parameter H_0, which characterises the present-day rate of expansion of the universe.

Since the redshift of a merging binary cannot be inferred directly from its gravitational waves,

many techniques have been developed in the literature to measure H_0 using additional astrophysical observations (see Mastrogiovanni et al. (2024) for a recent review). For example, the 'bright siren' method uses direct electromagnetic counterparts of the merger events to obtain redshifts (Abbott et al., 2017). In contrast, the statistical dark siren method (see Gair et al. (2023) for a review) uses galaxy surveys to identify potential hosts of the merger events inside the localisation volumes provided by gravitational wave observations (Alfradique et al., 2024; Abbott et al., 2021; Palmese et al., 2020; Soares-Santos et al., 2019). MacLeod & Hogan (2008) proposed a method that uses galaxy clustering to extract redshift information for a sample of merger events in a statistical sense to estimate H_0 without needing to identify host galaxies for individual merger events. Methods have also been proposed that try to estimate the redshift of gravitational wave sources by breaking the mass-redshift degeneracy in gravitational wave analyses; this can be achieved, for example, by constraining the neutron star tidal deformability or by combining features in the mass distribution and redshift evolution of merger rate (the so-called 'spectral siren method'), to measure the source masses (Abbott et al., 2023; Mancarella et al., 2022; Ezquiaga & Holz, 2022; Mastrogiovanni et al., 2021; Farr et al., 2019).

All of these methods, however, have various drawbacks (Mastrogiovanni et al., 2024). The bright siren method relies on detecting rare events accompanied by electromagnetic counterparts and possibly only applies to binary neutron star (BNS) mergers. The dark siren method suffers from difficulties due to large localisation volumes of gravitational wave events and is susceptible to potential biases in the inference of H_0 due to the incompleteness of galaxy catalogues (Trott & Huterer, 2022). The mass-redshift degeneracy method is model-dependent; uncertainties in the modelling of the neutron star equation of state, or a wrong model of the binary merger rate, can introduce systematic biases in the cosmological inference (see section 2.3.3 of Mastrogiovanni et al. (2024) and references therein).

If star-forming regions follow the fluctuations in the underlying cosmological matter field, merging compact binaries are expected to be inherently clustered and spatially correlated with other tracers such as galaxies and galaxy clusters (Scelfo et al., 2018). The strength of the cross-correlation between sources of gravitational waves and the large-scale structure of the universe is sensitive to cosmological parameters. It can, therefore, be used as an independent probe of the Hubble-Lemaître constant (Mukherjee et al., 2022, 2021; Bera et al., 2020; Oguri, 2016), after marginalizing over the other relevant parameters. As long as the gravitational wave sources and the galaxy sample trace fluctuations in the same underlying density field, a measurement of their spatial cross-correlation does not require uniquely identifying the source of each merger event (Fang et al.

(2020) have a similar discussion in the context of cross-correlations between high-energy neutrinos and large-scale structure). Therefore, the cross-correlation method of determining H_0 does not suffer from biases due to the incompleteness of the galaxy catalogues used (see Bera et al., 2020, for a systematic study). Moreover, this method does not require any direct assumptions about the population properties of the merging binaries. Therefore, it measures H_0 nearly independently of merging binary population models[1].

Since the various techniques of measuring H_0 using gravitational wave sources discussed above do not utilise information in the temperature fluctuations of the cosmic microwave background (CMB) (Aghanim et al., 2020; Ade et al., 2016) or cosmic distance ladder distance measurements from supernovae and other standard candles (Riess et al., 2022; Riess, 2020; Wong et al., 2020; Riess et al., 2019, 2018), measuring the cross-correlation between gravitational wave sources and the large-scale structure of the universe is important in the context of the so-called Hubble Tension (see Hu & Wang (2023) and Valentino et al. (2021) for comprehensive reviews). In addition to cosmology, these cross-correlation measurements can also be used to study the astrophysical origins and formation channels of gravitational wave sources (see, e.g., Gagnon et al., 2023; Adhikari et al., 2020; Scelfo et al., 2018; Raccanelli et al., 2016). With the third generation of gravitational wave detectors likely to bring in $\sim 10^5$ more detections per year (Iacovelli et al., 2022; Borhanian & Sathyaprakash, 2022), measuring the clustering of these objects and modelling it as a function of cosmological parameters will, therefore, play an essential role for both precision cosmology and compact binary astrophysics in the coming decade.

There have been a few attempts to measure the angular two-point correlation function (Zheng et al., 2023; Cavaglià & Modi, 2020) and the angular power spectrum (Zheng et al., 2023) of the currently detected LVK events, as well as attempts to measure their spatial cross-correlation with galaxy catalogues (Mukherjee et al., 2022), but a statistically significant detection of clustering has not yet been achieved. Studies using forecasts for future detectors have also been performed (Gagnon et al., 2023; Vijaykumar et al., 2023b; Balaudo et al., 2023; Libanore et al., 2022, 2021; Calore et al., 2020; Scelfo et al., 2018; Namikawa et al., 2016), primarily focusing on two-point summary statistics.

In this thesis, we present a framework to quantify the clustering of gravitational wave sources

[1]It is to be noted, however, that the clustering of gravitational wave sources is also expected to be a function of bias parameters that model the tracer-matter connection. One needs to marginalise over these parameters to obtain constraints on cosmological parameters (see, e.g., Peron et al., 2023; Banerjee et al., 2022, and references therein). Since the bias parameters are controlled by the source population properties (Peron et al., 2023; Raccanelli et al., 2016), an indirect dependence on population models is introduced in the measurement of H_0.

and their spatial cross-correlation with large-scale structure catalogues using the k-nearest-neighbour distributions (Banerjee & Abel, 2021a) as summary statistics. The nearest-neighbour measurements are sensitive to all N-point correlation functions of the tracers and hence are a much more powerful probe of clustering, compared to the two-point function, on scales where the underlying matter field is not well-approximated as a Gaussian random field, and the effect of gravitational nonlinearities cannot be neglected (Banerjee & Abel, 2023, 2021a,b). Application of these statistics could, in principle, lead to a detection of the clustering signal from the same datasets used in previous two-point analyses. To enable this new analysis, we extend the kNN formalism, originally presented for 3D clustering in cartesian coordinates, to angular clustering in the sky.

As a first application of our analysis framework to data, we compute the auto-correlation of a suitable subset of the binary black holes (BBHs) detected in the first three observing runs of LVK and their cross-correlation with the WISE$\times$SuperCOSMOS all-sky survey. We also compare the results of the two-point and nearest-neighbour analyses. As a second application, we investigate the feasibility of detecting the BBH-galaxy cross-correlation with future gravitational wave observing runs and stage-IV large-scale structure surveys. Although we focus on BBHs in this work, our framework can easily be extended to study the clustering of other gravitational wave transients like binary neutron stars and neutron star-black hole binaries.

The rest of the thesis is structured as follows. In chapter 2, we develop the mathematical formalism for clustering statistics and discuss how to compute them numerically. We present the clustering analysis using the current data in chapter 3. In this chapter, we describe the data used in this study in section 3.1 and discuss the application of the clustering formalism to this particular data in section 3.2. Section 3.3 outlines our procedure to determine the statistical significance of the clustering signal and discusses the hypothesis-testing framework used for this purpose. We describe the angular scales used in the clustering analysis in section 3.4. In section 3.5, we present an illustrative example that demonstrates the potential boost in the clustering signal of sparsely sampled tracers expected from the nearest-neighbour measurements on small spatial scales over the two-point summary statistics. We present our results in section 3.6 and conclude chapter 3 by discussing some interesting aspects of our findings in section 3.7. Chapter 4 presents the results of our forecast study. We describe the mock data created for the forecasts in section 4.1 and present our findings in section 4.2. Finally, we summarise, draw conclusions, and discuss possible future directions in chapter 5. Some additional material is presented in the appendices.

Chapter 2

Mathematical Formalism

In this chapter, we describe the summary statistics used to quantify clustering strength. For each statistic, we give a mathematical definition followed by a computational recipe to calculate the clustering strength for given data using that particular statistic. We discuss the auto-clustering of a set of discrete tracers in section 2.1, the cross-clustering between two different sets of tracers in section 2.2, and the cross-clustering between a set of discrete tracers and a continuous field in section 2.3.

2.1 Auto-clustering

Consider a set X of N_X discrete, point-like tracers or data points[1].

$$X \equiv \{(\delta_1, \alpha_1), (\delta_2, \alpha_2), ..., (\delta_{N_X}, \alpha_{N_X})\} \tag{2.1}$$

where δ_i, α_i represents the declination and right ascension of the i^{th} tracer in celestial equatorial (J2000) coordinates. In polar coordinates, the position of the i^{th} tracer is given by

$$\theta_i = \delta_i - \pi/2 \tag{2.2}$$

$$\phi_i = \alpha_i \tag{2.3}$$

[1] We use tracer and data point interchangeably.

To study clustering, we need a metric for the distance between two points $(\delta_1, \alpha_1), (\delta_2, \alpha_2)$ in the sky. A natural choice is the great-circle distance given by $d = \theta$, where θ is the central angle between the two points on the sphere. Note that we treat the sky as a unit sphere; hence, the radius term that usually multiplies the angle to get the distance is absent. The central angle between $(\delta_1, \alpha_1), (\delta_2, \alpha_2)$ is computed using the haversine formula (RIOS, 1795)

$$\text{hav}(\theta) = \text{hav}(\delta_2 - \delta_1) + \cos\delta_1 \cos\delta_2 \text{hav}(\alpha_2 - \alpha_1) \tag{2.4}$$

where $\text{hav}(\theta) \triangleq \sin^2(\theta/2)$ is the haversine function.

2.1.1 Two-point statistics

The sky positions of the tracers X can be used to define a number density field $n_X(\theta, \phi)$ in the sky, such that

$$\int_{\text{Allsky}} n_X(\theta, \phi) \sin\theta d\theta d\phi = N_X \tag{2.5}$$

This can be done numerically, for example, by dividing the sky into equal-area pixels using some pixelating scheme and counting the number of tracers in each pixel divided by the area of each pixel. Let the average tracer number density in the sky be $\bar{n}_X = \frac{N_X}{4\pi}$. The overdensity field is given by

$$\delta_X(\theta, \phi) = \frac{n_X(\theta, \phi)}{\bar{n}_X} - 1 \tag{2.6}$$

The overdensity field contains all the information about the auto-clustering of the tracer set A. By expanding $\delta_X(\theta, \phi)$ into spherical harmonics, one can derive the angular power spectrum $\mathscr{C}_\ell^{X,X}$, a widely used two-point summary statistic for clustering:

$$\delta_X(\theta, \phi) = \sum_{\ell m} \alpha_{\ell m}^X Y_{\ell m}(\theta, \phi) \tag{2.7}$$

$$\mathscr{C}_\ell^{X,X} = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{\ell} |\alpha_{\ell m}^X|^2 \tag{2.8}$$

Where ℓ goes from 0 to ∞ and m takes values from $-\ell$ to ℓ. In practice, the summation is cut off at some ℓ_{max} determined by the resolution of the numerical grid on which the field is defined. The power spectrum $C_\ell^{X,X}$ at a particular value of ℓ quantifies the clustering strength at an angular

scale corresponding roughly to $\theta \approx \pi/\ell$. In this study, we use the HEALPix[2] scheme (Górski et al., 2005) as implemented in the `python` library `healpy`[3] (Zonca et al., 2019) to compute the overdensity field on a grid of equal-area pixels in the sky. We use the `healpy`'s `anafast` routine to compute the power spectra. We choose the default high-ℓ cutoff of `3NSIDE - 1` in our analysis.

An equivalent measure of spatial clustering of a set of tracers is the two-point auto-correlation function, $w^{X,X}(\theta)$, which captures the excess probability of finding two data points separated by an angular distance of θ in the sky over finding two points drawn from a random (Poisson) distribution in the sky. Mathematically,

$$w^{X,X}(\theta) = \left\langle \delta_X(\hat{\Omega}_1)\delta_X(\hat{\Omega}_2) \right\rangle_{\hat{\Omega}_1\cdot\hat{\Omega}_2=\cos\theta} \tag{2.9}$$

where $\hat{\Omega}_1$ and $\hat{\Omega}_2$ are unit vectors separated by an angular distance θ in the sky such that $\hat{\Omega}_1 \cdot \hat{\Omega}_2 = \cos\theta$, and the angular brackets denote an average over all such configurations of $\hat{\Omega}_1$ and $\hat{\Omega}_2$. It can be shown that the angular power spectrum is related to the two-point function of the tracers in the following way:

$$w^{X,X}(\theta) = \frac{1}{4\pi}\sum_{\ell}(1+2\ell)\mathscr{C}_\ell^{X,X}P_\ell(\cos\theta) \tag{2.10}$$

Where $P_\ell(\cos\theta)$ denotes the Legendre polynomial of order ℓ and argument $\cos\theta$. Therefore, the two-point function and the angular power spectrum encode the same physical information about the clustering of the tracer set X; the angular power spectrum is a two-point clustering statistic.

In practice, the two-point function is numerically computed directly from the angular positions of the N_X data points and a set of N_r randomly distributed points using the Landy-Szalay estimator (Landy & Szalay, 1993)

$$\hat{w}^{X,X}(\theta) = \frac{(N_r^2)DD(\theta) - 2(N_rN_X)DR(\theta) + (N_X^2)RR(\theta)}{(N_X^2)RR(\theta)} \tag{2.11}$$

where DD, DR and RR refer to the number of data-data, data-random and random-random pairs separated by an angular distance θ. Typically, the number of randoms is chosen to be significantly larger than the number of data points ($N_r \gg N_X$).

[2] https://healpix.sourceforge.io/
[3] https://healpy.readthedocs.io/en/latest/

2.1.2 Nearest-neighbour distributions

The nearest-neighbour distributions as a measure of spatial clustering in 3D were introduced in Banerjee & Abel (2021a). Here, we briefly summarise the idea behind these statistics and extend the mathematical formalism to 2D clustering in the sky using angular coordinates.

The key idea that motivates the nearest-neighbour clustering framework is as follows: all the physical information about the clustering of a set of discrete tracers is contained in the distribution of their number counts, ie., the number of tracers enclosed inside a randomly chosen spatial region of a given spatial extent. The spatial regions can have an arbitrary geometrical shape, as long as there is a way of assigning a spatial extent to them. Since we are concerned with tracers in the sky, which is represented by the surface of a 3-sphere, we choose to work with spherical caps of area $A = 2\pi(1-\cos\theta)$ to study clustering at an angular scale θ[4].

Suppose we are given the positions of a set X of discrete tracers. Given the discussion above, the fundamental quantity that quantifies their clustering at spatial scale θ is the probability $\mathscr{P}_{k|A}$ of finding k data points of X in a randomly placed spherical cap of area A in the sky. $\mathscr{P}_{k|A}$ can be written in terms of a generating function as

$$P(z|A) \triangleq \sum_{k=0}^{\infty} \mathscr{P}_{k|A} z^k \tag{2.12}$$

or,

$$\mathscr{P}_{k|A} = \frac{1}{k!}\left[\left(\frac{d}{dz}\right)^k P(z|A)\right]_{z=0} \tag{2.13}$$

For the case of spherical caps, it can be shown that the generating function is given by (Banerjee & Abel, 2021a)[5]

$$P(z|A) = \exp\left[\sum_{k=1}^{\infty} \frac{\bar{n}_X^k (z-1)^k}{k!} \int_A \ldots \int_A d\hat{\Omega}_1 \ldots d\hat{\Omega}_k \omega^{(k)}\left(\hat{\Omega}_1, \ldots \hat{\Omega}_k\right)\right] \tag{2.14}$$

where $\omega^{(N)}$ are the N-point correlation functions of the underlying field of the tracers X, with $\omega^{(0)} = 0$ and $\omega^{(1)} = 1$ by definition[6]. Equation 2.14 shows the connection between the number

[4] Henceforth, we use θ and A interchangeably.

[5] See appendix A of Banerjee & Abel (2021a) or references therein for a derivation in the context of 3D clustering in cartesian coordinates.

[6] Note that the N-point functions are defined analogously to the two-point auto-correlation function defined earlier.

count distribution and the correlation functions that are usually employed to measure clustering.

An equivalent measure of clustering is the cumulative distribution of the tracer number counts, which represents the probability $\mathscr{P}_{>k|A}$ of having more than k tracers in a randomly chosen spherical cap of A. By definition, this is equal to the sum of the probabilities of having $k+1$, $k+2$, ... tracers in area A:

$$\mathscr{P}_{>k|A} = 1 - \sum_{m=0}^{k} \mathscr{P}_{m|A} \tag{2.15}$$

$\mathscr{P}_{>k|A}$ can also be expressed in terms of a generating function $C(z|A)$

$$C(z|A) \triangleq \sum_{k=0}^{\infty} \mathscr{P}_{>k|A} z^k \tag{2.16}$$

By plugging in equation 2.15 in equation 2.16 and simplifying the resulting expansion following Banerjee & Abel (2021a), we can express $C(z|A)$ in terms of $P(z|A)$ as

$$C(z|A) = \frac{1 - P(z|A)}{1 - z} \tag{2.17}$$

which allows us to determine $\mathscr{P}_{>k|A}$

$$\mathscr{P}_{>k|A} = \frac{1}{k!} \left[\left(\frac{d}{dz} \right)^k C(z|A) \right]_{z=0} \tag{2.18}$$

From equations 2.14 to 2.18, it is not clear how to compute these distributions without first computing all higher-order correlation functions. Following Banerjee & Abel (2021a), we now discuss another interpretation of the cumulative count distributions that allows us to compute $\mathscr{P}_{>k|A}$ directly from the positions of the tracers. The count distributions $\mathscr{P}_{k|A}$ can then be calculated trivially using $\mathscr{P}_{k|A} = \mathscr{P}_{>k-1|A} - \mathscr{P}_{>k|A}$.

Consider a set of N_r area-filling, randomly distributed query points in the sky, such that $N_r \gg N_X$. Each query point will have a data point in X that is nearest to it, a data point that is second-nearest to it and so on. The distributions of the distances to these neighbouring data points, over all query points in the sky, are directly connected to the count distributions discussed above. We argue that the cumulative distribution function (CDF) of the distances from the query points to the

In fact, $\omega^{(2)} \equiv w$. However, for $N > 2$, $\omega^{(N)}$ are defined as functions of N unit vectors $\{\hat{\Omega}_1, ..., \hat{\Omega}_N\}$ instead of a single angular separation θ, and the average is performed over all possible configurations preserving the polyhedron formed by the N vectors.

k-nearest-neighbour data point, or kNN-CDF, is precisely equal to $\mathscr{P}_{>k-1|A}$.

To understand this connection, let us examine the case of $k = 1$. Consider N_r spherical caps of area $A = 2\pi(1 - \cos\theta)$, the centres of which are distributed randomly in the sky. The fraction of such caps enclosing at least 1 data point is equal to that of cap centres with the angular distance to their nearest neighbour less than θ. The nearest-neighbour CDF at angular scale θ is the precise measure of the fraction of query points (equivalent to centres of the spherical caps) for which the nearest data point is at a distance less than θ. This argument can be easily generalised if we consider the k-nearest-neighbour instead of the first nearest-neighbour. Therefore, we conclude

$$\mathscr{P}_{>k-1|A} = \mathrm{CDF}_{k\mathrm{NN}}(\theta) \tag{2.19}$$

In practice, the kNN-CDFs are simple to calculate in a computationally efficient manner. We start by creating a HEALPix grid of query points with a sufficiently high value of `NSIDE`, such that the resolution of the query grid is much finer than the smallest angular scale at which we want to study spatial clustering. As noted in Banerjee & Abel (2021a), placing query points on a finely spaced grid gives the same results as randomly distributed query points, as long as the grid separation is much smaller than the mean interparticle separation of the data. Next, we compute the distances to the k-nearest-neighbour data point of each query point. The nearest-neighbour search is carried out very efficiently by constructing a Ball tree structure (Omohundro, 2009) on the data points. Once a tree is built, it can be used to calculate the distances to the first k neighbouring data points for all query points simultaneously. Sorting the computed distances for each neighbour index k immediately gives the empirical CDF of the k-nearest-neighbour distances over a range of spatial scales. The empirical CDF converges to the true kNN-CDF in the limit of a large number of query points. In this study, we utilise the `sklearn.neighbors.BallTree` routine from the library `scikit-learn`[7] (Pedregosa et al., 2012) with the haversine distance metric for our purposes.

It is evident from equation 2.14 that the count distributions $\mathscr{P}_{>k|A}$, and hence the kNN-CDFs, are formally sensitive to integrals of all N-point correlation functions of the underlying tracer field. This makes these summary statistics extremely powerful probes of clustering on small spatial scales where the higher-order correlation functions contribute significantly. The interested reader is referred to Banerjee & Abel (2021a) for a detailed study of the gain in clustering measurements as well as cosmological constraints achieved using the kNN-CDFs over two-point clustering statistics.

[7] https://scikit-learn.org/

2.2 Tracer-tracer Cross-clustering

So far, we have worked with a single set of tracers X. Consider now another set of discrete tracers Y. These could be (possibly biased) tracers of the same underlying field that the set X traces or of another field that is physically correlated to the field traced by X, in which case the positions of the tracers X and Y are expected to be cross-correlated. In this section, we describe the summary statistics that measure the extent of this correlation. Of course, it is also possible that X and Y trace completely independent fields. In that case, there would be no cross-correlation in the positions of X and Y. As we will see, the summary statistics defined below can be associated with a unique fiducial value that indicates the absence of correlations.

2.2.1 Two-point statistics

As discussed in section 2.1.1, we can define overdensity fields δ_X and δ_Y from the positions of the discrete tracers X and Y, and expand both δ_X and δ_Y in spherical harmonics

$$\delta_X(\theta,\phi) = \sum_{\ell m} \alpha^X_{\ell m} Y_{\ell m}(\theta,\phi)$$
$$\delta_Y(\theta,\phi) = \sum_{\ell m} \alpha^Y_{\ell m} Y_{\ell m}(\theta,\phi)$$

The cross angular power spectrum between X and Y is defined as

$$\mathscr{C}_\ell^{X,Y} = \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} \left\{\alpha^X_{\ell m}\right\}^* \alpha^Y_{\ell m} \tag{2.20}$$

We compute the cross angular power spectrum in a similar manner to the auto angular power spectrum, using `healpy`'s `anafast` routine.

Similarly, the two-point cross-correlation function is defined as

$$w^{X,Y}(\theta) = \left\langle \delta_X(\hat{\Omega}_1)\delta_Y(\hat{\Omega}_2) \right\rangle_{\hat{\Omega}_1\cdot\hat{\Omega}_2=\cos\theta} \tag{2.21}$$

where the angular brackets denote an average over configurations of unit vectors separated by a fixed central angle θ as before. In practice, the two-point cross-correlation function between X and

Y is computed using their angular positions in a very similar way to the two-point auto-correlation function of X. Here we need to create two sets of randoms (one for each tracer set) containing $N_r^1 \gg N_X$ and $N_r^2 \gg N_Y$ points. Using these, the Landy-Szalay estimator provides a measure of the two-point cross-correlation function as (Landy & Szalay, 1993)

$$\hat{w}^{X,Y}(\theta) = \frac{N_r^1 N_r^2 D_X D_Y(r) - N_r^1 N_Y D_1 R_2(r) - N_r^2 N_X D_2 R_1(r) + N_X N_Y R_1 R_2(r)}{N_X N_Y R_1 R_2(r)} \tag{2.22}$$

where $D_X D_Y$, $R_1 R_2$ and $D_i R_j$ refer to the number of data-data, random-random and data-random pairs separated by an angular distance θ[8].

If X and Y trace fields that are statistically independent, ie., if X and Y are spatially uncorrelated, then the angular power spectrum and the two-point cross-correlation function both are expected to be zero. However, in practice, their measured values can fluctuate from zero even for uncorrelated tracers due to finite-sampling noise. Therefore, to determine if a non-zero measurement of $\mathscr{C}_\ell^{X,Y}$ or $\hat{w}^{X,Y}$ actually represents a cross-clustering signal, it is important to characterize the errors in these statistics due to sample variance. We will discuss the procedure to do this in detail in chapter 3.

2.2.2 Nearest-neighbour distributions

The nearest-neighbour framework for measuring the spatial cross-correlations between two sets of tracers was introduced in Banerjee & Abel (2021b) for 3D cartesian coordinates. In this section, we briefly discuss the conceptual ideas and extend the mathematical formalism to 2D clustering in the sky using angular coordinates.

We discussed in section 2.1.2 that the physical information needed to characterize the clustering of a set of tracers X is contained in the distribution of number counts of X in randomly placed spherical caps in the sky. Similarly, the extent to which two discrete tracers X and Y are cross-correlated is characterized by the joint distribution of number counts of X and Y. More precisely, the quantity of interest is the joint probability $\mathscr{P}_{k_X,k_Y|A}$ of finding k_X data points of X and k_Y data points of Y in randomly placed spherical caps of area A in the sky. $\mathscr{P}_{k_X,k_Y|A}$ can be expressed in

[8]Note that the pairs considered in the above expression are always X-Y pairs, and never X-X pairs, since we are computing a cross-correlation.

terms of a generating function $P(z_x, z_y|A)$ as follows (Banerjee & Abel, 2021b)

$$P(z_X, z_Y|A) \triangleq \sum_{k_X=0}^{\infty} \sum_{k_Y=0}^{\infty} \left(\mathscr{P}_{k_X,k_Y|A}\right) z_X^{k_X} z_Y^{k_Y} \tag{2.23}$$

or,

$$\mathscr{P}_{k_X,k_Y|A} = \frac{1}{k_X!}\frac{1}{k_Y!}\left[\left(\frac{d}{dz_X}\right)^{k_X}\left(\frac{d}{dz_Y}\right)^{k_Y} P(z_x, z_y|A)\right]_{z_X,z_Y=0} \tag{2.24}$$

The generating function encapsulates the connection between the joint number counts and the cross-correlation functions $\omega^{(k_X,k_Y)}$ between the underlying fields traced by X and Y at all orders, and is given by (Banerjee & Abel, 2021b)[9]

$$P(z_X, z_Y|A) = \exp\left[\sum_{k_X=1}^{\infty}\sum_{k_Y=1}^{\infty} \frac{\bar{n}_X^{k_X}(z_X-1)^{k_X}}{k_X!}\frac{\bar{n}_Y^{k_Y}(z_Y-1)^{k_Y}}{k_Y!}\int_A d\hat{\Omega}_1...d\hat{\Omega}_{k_X}d\hat{\Omega}'_1...d\hat{\Omega}'_{k_Y}\,\omega^{(k_X,k_Y)}\right] \tag{2.25}$$

Note that $\omega^{(k_X,k_Y)}$ represents the correlation function between k_X factors of δ_X and k_Y factors of δ_Y, and is defined as

$$\omega^{(k_X,k_Y)}(\hat{\Omega}_1...\hat{\Omega}_{k_X};\hat{\Omega}'_1...\hat{\Omega}'_{k_Y}) = \left\langle \delta_X(\hat{\Omega}_1)...\delta_X(\hat{\Omega}_{k_X})\delta_Y(\hat{\Omega}'_1)...\delta_Y(\hat{\Omega}'_{k_Y})\right\rangle \tag{2.26}$$

where the average is over all possible configurations of unit vectors $\{\hat{\Omega}_1...\hat{\Omega}_{k_X};\hat{\Omega}'_1...\hat{\Omega}'_{k_Y}\}$ that form the same polyhedron[10]. Note that $\omega^{(N,0)}$ and $\omega^{(0,N)}$ represent the N-point auto-correlation functions of X and Y respectively, and $\omega^{(1,1)}$ represents the two-point cross-correlation function defined in section 2.2.1.

An equivalent quantity to characterise cross-clustering between X and Y is the joint probability $\mathscr{P}_{\geq k_X,\geq k_Y|A}$ of finding more than k_X data points of X and more than k_Y data points of Y in randomly placed spherical caps of area A. One can define a generating function $C(z_X, z_Y|A)$ for $\mathscr{P}_{>k_X,>k_Y|A}$:

$$C(z_X, z_Y|A) \triangleq \sum_{k_X=0}^{\infty} \sum_{k_Y=0}^{\infty} \left(\mathscr{P}_{>k_X,>k_Y|A}\right) z_X^{k_X} z_Y^{k_Y} \tag{2.27}$$

[9] See appendix A of Banerjee & Abel (2021b) for a derivation in 3D cartesian coordinates. The argument is similar for 2D angular coordinates.

[10] While the two-point function is a function of an angle, the higher order correlation functions are functions of polyhedra.

Now, by definition

$$\mathscr{P}_{>k_X,>k_Y|A} = 1 - \sum_{m_X=0}^{k_X} \mathscr{P}_{m_X|A} - \sum_{m_Y=0}^{k_Y} \mathscr{P}_{m_Y|A} + \sum_{m_X=0}^{k_X}\sum_{m_Y=0}^{k_Y} \mathscr{P}_{m_X,m_Y|A} \tag{2.28}$$

Using equations 2.27 and 2.28, $C(z_X, z_Y|A)$ can be written as (Banerjee & Abel, 2021b)

$$C(z_X, z_Y|A) = \frac{1 - P(z_X|A) - P(z_Y|A) + P(z_X, z_Y|A)}{(1-z_X)(1-z_X)} \tag{2.29}$$

where $P(z|A)$ is the generating function for the individual number count distribution given by equation 2.14. Finally, $\mathscr{P}_{>k_X,>k_Y|A}$ can be expressed as

$$\mathscr{P}_{>k_X,>k_Y|A} = \frac{1}{k_X!}\frac{1}{k_Y!}\left[\left(\frac{d}{dz_X}\right)^{k_X}\left(\frac{d}{dz_Y}\right)^{k_Y} C(z_x, z_y|A)\right]_{z_X,z_Y=0} \tag{2.30}$$

From the generating function, it is evident that the joint number count distributions are sensitive to all higher-order correlation functions of the underlying fields traced by X and Y, therefore these statistics are expected to be more potent measures of clustering than the individual correlation functions. Now that we have a formalism in place that connects the cumulative joint number counts of the tracers to the cross-correlation between their underlying fields, we discuss an alternate interpretation that allows for efficient measurement of these probabilities using the positions of the tracers without having to compute the higher-order correlation functions. The argument presented below is very similar to the one presented in section 2.1.2.

Suppose we want to compute the value of $\mathscr{P}_{>k_X-1,>k_Y-1|A}$. Consider a set of N_r area-filling, randomly distributed query points in the sky, such that $N_r \gg N_X, N_Y$. To each query point, we can assign two nearest-neighbour data points, one belonging to the tracer set X and the other to Y. The same can be done for the second-nearest neighbour, third-nearest neighbour and so on. Then, we can compute the query point's distances to the k_X-nearest neighbour in X and the k_Y-nearest neighbour in Y. We argue that the cumulative distribution function (CDF) of the larger of these two distances, evaluated at distance θ, is exactly equal to the probability $\mathscr{P}_{>k_X-1,>k_Y-1|A}$. We call this CDF the joint $\{k_X, k_Y\}$NN-CDF of the tracers X and Y.

To understand this connection, consider the case of $k_X = k_Y = 1$. At a fixed distance scale θ, the value of the joint $\{1,1\}$NN-CDF represents the fraction of spherical caps of area A centred at the query points for which the angular distances to the nearest-neighbour data point in X and Y are both

smaller than θ, since it is defined as the fraction for which the larger of the two nearest-neighbour distances is smaller than θ. In the limit of large and area-filling query points, this fraction is equivalent to the probability of finding at least one data point of both X and Y in randomly chosen spherical caps of area A. This argument easily generalises for any given k_X, k_Y pair. Hence, we conclude,

$$\mathscr{P}_{\geq k_X, \geq k_Y | A} = \mathrm{CDF}_{k_X, k_Y} \tag{2.31}$$

Now, it is clear from the generating function (equation 2.25) that the joint CDFs depend not only on the cross-correlation functions of the two tracers but also on the auto-correlation functions. As discussed in (Banerjee & Abel, 2021b), there is a way to isolate the dependence on the cross-correlation functions to get a purer measure of the spatial cross-correlation of X and Y. We describe this below.

If the fields traced by X and Y are statistically independent, meaning that X and Y are spatially uncorrelated, then all the cross-correlation functions are identical zero, i.e., $\omega^{(k_X, k_Y)}$ is non-zero only when either $k_X = 0$ or $k_Y = 0$. This combined with equations 2.25 and 2.29 implies that the generating functions for the joint number counts and cumulative number counts factorise into products of the individual number count and cumulative number count distributions (Banerjee & Abel, 2021b)

$$P(z_X, z_Y | A) = P(z_X | A) P(z_Y | A) \tag{2.32}$$
$$C(z_X, z_Y | A) = C(z_X | A) C(z_Y | A) \tag{2.33}$$

Therefore, for spatially uncorrelated tracers, we have the following factorisation

$$\mathscr{P}_{\geq k_X, \geq k_Y | A} = \mathscr{P}_{\geq k_X | A} \times \mathscr{P}_{\geq k_Y | A} \tag{2.34}$$

This fact provides a convenient way to define a summary statistic that measures the *excess cross-correlation* between the two sets of tracers

$$\psi_{k_X, k_Y} \triangleq \mathscr{P}_{\geq k_X, \geq k_Y | A} \big/ \left(\mathscr{P}_{\geq k_X | A} \times \mathscr{P}_{\geq k_Y | A} \right) \tag{2.35}$$

or equivalently

$$\psi_{k_X, k_Y} = \mathrm{CDF}_{k_X, k_Y} \big/ \left(\mathrm{CDF}_{k_X \mathrm{NN}} \times \mathrm{CDF}_{k_Y \mathrm{NN}} \right) \tag{2.36}$$

where $\mathrm{CDF}_{k\mathrm{NN}}$ is the auto-CDF of a single set of tracers, as defined in section 2.1.2. This is a very

useful quantity, as a positive (negative) measurement for $\psi_{k_X,k_Y} - 1$ would indicate that the tracer X is correlated (anti-correlated) with the field δ_Y, while $\psi_{k_X,k_Y} - 1 = 0$ would indicate that the there is no spatial cross-correlation between them. In this thesis, we report all our cross-correlation results in the form of the excess cross-correlation ψ_{k_X,k_Y}.

The computational recipe for computing the nearest-neighbour excess cross-correlation is as follows (Banerjee & Abel, 2021b):

1. Create a set of area-filling query points by creating a finely-spaced HEALPix grid in the sky, such that the number of pixels $N_{\rm pix}$ is far greater than the number of data points of both sets of tracers.

2. Build a Ball tree from both sets of tracer positions and estimate the query points' angular distances to the k_X-nearest neighbour data point in X and the k_Y-nearest neighbour data point in Y. For each query point, separately store the larger of the two nearest-neighbour distances.

3. Sort both sets of nearest-neighbour distances to produce the empirical k_X and k_YNN-CDFs of X and Y over a range of angular scales θ.

4. Sort the larger set of distances stored in step (ii) to obtain the empirical joint $\{k_X, k_Y\}$NN-CDF over the range of angular scales considered in step (iii).

5. From the quantities calculated above, compute the excess cross-correlation using equation 2.36.

2.3 Tracer-field Cross-clustering

We now describe the summary statistics for measuring the spatial cross-correlation between a set of discrete tracers, X, and a continuous field δ_Y. We only consider continuous fields in the form of dimensionless fluctuations in a quantity, i.e., fields that are bounded below by -1 and average to 0. The continuous field could be an overdensity field derived from another sample of tracers, such as the galaxies relevant for this study, or a true continuum field like the CMB temperature fluctuations.

2.3.1 Two-point statistics

The cross angular power spectrum and the two-point cross-correlation function between tracer X and field Y are both defined exactly as for two tracers X and Y:

$$\mathscr{C}_\ell^{X,Y} = \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} \left\{\alpha_{\ell m}^X\right\}^* \alpha_{\ell m}^Y$$

$$w^{X,Y}(\theta) = \left\langle \delta_X(\hat{\Omega}_1)\delta_Y(\hat{\Omega}_2) \right\rangle_{\hat{\Omega}_1\cdot\hat{\Omega}_2=\cos\theta}$$

The only difference is that we do not need to create an overdensity field for Y since it already is one. The numerical computation for the cross angular power spectrum remains the same. However, the interpretation of the two-point cross-correlation function and the method to estimate it numerically differs from the methods followed in section 2.1.1 and section 2.2.1. Using $\delta_X = (1+\delta_X) - 1$ and rewriting the average in equation 2.21 as an integral, we get

$$w^{X,Y}(\theta) = \frac{1}{\mathscr{N}} \int_{\text{All sky}} d\hat{\Omega}_1 \left(1+\delta_X(\hat{\Omega}_1)\right) \int_{\hat{\Omega}_1\cdot\hat{\Omega}_2=\cos\theta} d\hat{\Omega}_2 \delta_Y(\hat{\Omega}_2) - \frac{1}{\mathscr{N}} \int_{\text{All sky}} d\hat{\Omega}_2 \delta_Y(\hat{\Omega}_2)$$

where $\mathscr{N}$ is some normalisation constant. Now, by definition, the average of an overdensity field is zero, so the second term in the above expression drops out. Let us look at the first term in more detail. Since the data points in X trace the underlying density field $\bar{n}_X(1+\delta_X)$, $(1+\delta_X)$ can be treated as an (unnormalised) probability density function. Therefore, the integral over $\hat{\Omega}_1$ can be approximated by an average over the positions of the tracers X as follows

$$\hat{w}^{X,Y}(\theta) = \frac{1}{N_X} \sum_{\hat{\Omega}_1 \in X} \int_{\hat{\Omega}_1\cdot\hat{\Omega}_2=\cos\theta} d\hat{\Omega}_2 \delta_Y(\hat{\Omega}_2)$$

The summand in the above expression is nothing but the continuous field δ_Y smoothed over a thin spherical band of inner radius θ and thickness $d\theta$, centred at position $\hat{\Omega}_1$. Therefore, the estimator for the two-point cross-correlation function between the tracers X and continuous field δ_Y is given by

$$\hat{w}^{X,Y}(\theta) = \frac{1}{N_X} \sum_{\hat{\Omega}_1 \in X} \delta_{Y,\text{band}}^{\theta}(\hat{\Omega}_1) \tag{2.37}$$

The field δ_Y smoothed over a thin spherical band is given by

$$\delta^{\theta}_{Y,\text{band}} = \frac{A(\theta+d\theta)\delta^{\theta+d\theta}_Y - A(\theta)\delta^{\theta}_Y}{A(\theta+d\theta)-A(\theta)} \tag{2.38}$$

where δ^{θ}_Y is the field Y smoothed over a spherical cap of angular size θ and area $A(\theta)$. In the limit of $N_X \to \infty$, the estimator $\hat{w}^{X,Y}(\theta)$ approaches the true value for $w^{X,Y}(\theta)$. Therefore, similar to Banerjee & Abel (2023), in practice, we compute the two-point cross-correlation by averaging δ_Y in spherical bands at angular radius θ and thickness $d\theta$, around the positions of all data points in the tracer sample X.

2.3.2 Nearest-neighbour distributions

Banerjee & Abel (2023) generalised the kNN formalism to study tracer-field cross-correlations in 3D using the nearest-neighbour distributions. Here, we summarise the main points in the context of 2D angular clustering. First, we discuss the behaviour of the kNN-CDFs of a set of tracers of the field δ_Y as their average number density $\bar{n}_Y$ tends to infinity, i.e., the continuum limit of the kNN-CDFs.

As discussed in section 2.1.2, the probability of finding at least k tracers in a randomly centred spherical cap of area $A = 2\pi(1-\cos\theta)$ in the sky is connected to the kNN-CDF evaluated at the angular scale θ

$$\mathscr{P}_{\geq k|A} = \text{CDF}_{k\text{NN}}(\theta)$$

Since the tracers represent a local Poisson process on the field δ_Y, the probability of finding exactly k tracers in a spherical cap of area A centred at point $\hat{\Omega}$ is given by

$$\mathscr{P}_{k|A}(\hat{\Omega}) = \frac{\left[\lambda(\hat{\Omega})\right]^k}{k!} e^{-\lambda(\hat{\Omega})} \tag{2.39}$$

where

$$\lambda(\hat{\Omega}) = \bar{n}_Y A\left(1+\delta^{\theta}_Y(\hat{\Omega})\right) \tag{2.40}$$

In the limit $\bar{n}_Y \to \infty$ while keeping δ_Y unchanged, as discussed in Banerjee & Abel (2023), the distribution in equation 2.39 can be well approximated by a Gaussian of vanishingly small width, which implies

$$\mathscr{P}_{k|A}(\hat{\Omega}) \to \delta^D(k-\lambda(\hat{\Omega}))$$

where δ^D is the Dirac delta function. This, combined with equation 2.40, means the following: the probability of finding exactly k tracers of δ_Y in a spherical cap of angular radius θ located at a point is non-zero when δ_Y smoothed on scale θ is vanishingly close to δ^* where $\bar{n}_Y A\,(1+\delta^*) = k$. In other words, for a given spherical cap, the value of k for discrete tracers gets mapped onto a specific value of enclosed overdensity, and we can write

$$\mathscr{P}_{k|A}(\hat{\Omega}) \to \delta^D(\delta_Y^\theta(\hat{\Omega}) - \delta^*)$$

Since $\mathscr{P}_{k|A}$ depends on the centre of the spherical cap only through the smoothed field, any integral over the centres $\hat{\Omega}$ can be rewritten as an integral over possible values of the smoothed field δ_Y^θ. Hence, the area-averaged probability of finding k tracers points in spherical caps of angular radius θ can be written in terms of the probability density function (PDF) $\phi(\delta_Y^\theta)$ of the smoothed field

$$\mathscr{P}_{k|A} = \int_{\text{All sky}} d\hat{\Omega}\,\mathscr{P}_{k|A}(\hat{\Omega}) \to \int \delta^D(\delta_Y^\theta - \delta^*)\phi(\delta_Y^\theta)d\delta_Y^\theta \propto \phi(\delta^*)$$

Finally, the expression for the probability of finding at least k tracers points in spherical caps of angular radius θ will be mapped onto the probability of getting $\delta_Y^\theta > \delta^*$

$$\mathscr{P}_{\geq k|A} \to \mathscr{P}_{>\delta^*}(\theta) = \int_{\delta^*}^{\infty} \phi(\delta_Y^\theta)d\delta_Y^\theta = 1 - \text{CDF}(\delta^*) \tag{2.41}$$

Equation 2.41 implies that the continuum version of the kNN measurements at a spatial scale θ are thresholded evaluations of the CDF of the smoothed continuous field δ_Y^θ. As discussed above, the nearest-neighbour index k for discrete data maps to a threshold δ^* on the smoothed continuous field[11].

Now that we have the analogue of the kNN-CDFs for a continuous field, we discuss the characterisation of spatial cross-correlations using the kNN formalism. Following Banerjee & Abel (2023), the joint probability $\mathscr{P}_{\geq k,>\delta^*}(\theta)$ of finding at least k tracers and the smoothed continuous field δ_Y^θ to cross threshold δ^* in spherical caps of angular radius θ is taken as a measure of the spatial cross-correlations between tracers and a continuous field. Assuming that the tracers X represent a local Poisson process on the overdensity field δ_X, we have, similar to Banerjee & Abel (2023),

$$\mathscr{P}_{k,>\delta^*}(\theta) = \int_{\delta^*}^{\infty} \frac{\left[\lambda(\delta_X^\theta)\right]^k}{k!} e^{-\lambda(\delta_X^\theta)} \phi(\delta_X^\theta, \delta_Y^\theta) d\delta_X^\theta d\delta_Y^\theta \tag{2.42}$$

[11] Note that at fixed k, δ^* is also a function of θ, as evident from its definition

where $\phi(\delta_X^\theta, \delta_Y^\theta)$ is the joint probability distribution of the two fields when smoothed on angular scale θ. The quantity of our interest, $\mathscr{P}_{\geq k, >\delta^*}$, can be written in terms of $\mathscr{P}_{k,>\delta^*}$ as

$$\mathscr{P}_{\geq k, >\delta^*} = \mathscr{P}_{>\delta^*} - \sum_{j<k} \mathscr{P}_{j,>\delta^*} \tag{2.43}$$

Suppose the tracers X are completely uncorrelated and statistically independent of the continuous field δ_Y. In that case, the joint distribution function can be factored into a product of the individual PDFs of the smoothed fields, i.e., $\phi(\delta_X^\theta, \delta_Y^\theta) \propto \phi(\delta_X^\theta)\ \phi(\delta_Y^\theta)$. In this case, we have (Banerjee & Abel, 2023) $\mathscr{P}_{\geq k, >\delta^*} = \mathscr{P}_{\geq k} \times \mathscr{P}_{>\delta^*}$. Similar to the case of tracer-tracer cross-correlations discussed in section 2.2.2, this can be used to define a convenient summary statistic that measures the *excess cross-correlation* between the tracers and the continuous field:

$$\psi_{k,\delta *} \triangleq \mathscr{P}_{\geq k, >\delta^*} / \left(\mathscr{P}_{\geq k} \times \mathscr{P}_{>\delta^*}\right) \tag{2.44}$$

As discussed before, a positive (negative) measurement for $\psi_{k,\delta *} - 1$ would indicate that the tracer X is correlated (anti-correlated) with the field δ_Y, while $\psi_{k,\delta *} - 1 = 0$ would indicate that the there is no spatial cross-correlation between them.

All the physical information about the spatial cross-correlation between fluctuations in the sky distribution of tracer X and the field Y is contained in the joint distribution $\phi(\delta_X^\theta, \delta_Y^\theta)$. Therefore, it is clear from equations 2.42 to 2.44 that the excess cross-correlation, as defined using the nearest-neighbour distributions, will be sensitive not just to the linear or Gaussian correlations in the density fluctuations of the tracers and the continuous field, but to correlations in fluctuations at all orders (See Banerjee & Abel (2023) for a detailed demonstration). Therefore, the kNN formalism provides a powerful way to characterise cosmological cross-correlations.

The joint probability distributions $\mathscr{P}_{\geq k, >\delta^*}$ and excess cross-correlations $\psi_{k,\delta *}$ are simple to compute numerically. We follow the procedure laid out in Banerjee & Abel (2023)

1. Create a set of area-filling query points by creating a finely-spaced HEALPix grid in the sky, such that the number of pixels $N_{\rm pix}$ is far greater than the number of data points.

2. Build a Ball tree from the set of tracer positions and estimate the angular distances to the k-nearest neighbour data points from each query point. For each k, sort the distances to produce the empirical kNN-CDF over a range of angular scales θ. In the limit of large $N_{\rm pix}$, the empirical CDF approaches $\mathscr{P}_{\geq k}$.

3. Smooth the continuous field δ_Y on an angular scale θ using a top-hat filter. The smoothing is done in harmonic space using the $\{\alpha^Y_{\ell m}\}$ of the field, computed via spherical harmonic transforms, to speed up the computation time (see appendix A for details). Interpolate the smoothed field on the query grid defined in step (i).

4. For a given k and threshold δ^*, compute the fraction of query points for which the k^{th} nearest-neighbour lies at an angular distance less than θ and the smoothed field, interpolated to that grid point, exceeds δ^*. In the limit of large N_{pix}, this fraction approaches $\mathscr{P}_{\geq k, >\delta^*}$.

5. Compute the fraction of query points for which the smoothed field, interpolated to the grid point, exceeds δ^*. In the limit of large N_{pix}, this fraction approaches $\mathscr{P}_{>\delta^*}$.

6. From the quantities calculated above, compute the excess cross-correlation using equation 2.44.

7. Repeat steps (iii) to (vi) for different values of the angular scale θ.

While the choice of k is straightforward, it is not clear at first how to decide the threshold value δ^* for the continuous field, especially as it varies with the spatial scale being considered. In this study, we choose the *constant percentile* threshold described in Banerjee & Abel (2023). We define $\delta^* = \delta_{75}$, the value of the 75^{th} percentile of δ^θ_Y. This choice implies that $\mathscr{P}_{>\delta^*} = 0.25$ irrespective of the smoothing scale θ.

Chapter 3

Clustering Measurements on Current Data

3.1 Data

In this section, we discuss the data used in this study. We describe the gravitational wave events selected for this work in section 3.1.1. Typically, a mock catalogue of unclustered data points (known as 'randoms' in the literature) is required to get a reliable measurement of the statistical significance of the clustering signal in the presence of observational selection biases (see., e.g., Wang et al., 2022). In section 3.1.2, we motivate this requirement for the specific case of gravitational wave data, discuss the procedure to create the unclustered catalogue, and present the resulting mock data. Finally, we describe the large-scale structure catalogue used for cross-correlating the BBHs in section 3.1.3.

3.1.1 Gravitational Wave Events

This work uses the compact binary merger events detected in the first three observing runs of LIGO-Virgo-KAGRA, as reported in LIGO Scientific Collaboration et al. (2023b). Following Zheng et al. (2023), from this parent set of $\sim$80 events, we select the events detected with a false alarm rate (FAR) less than 1 per year and crossed a detection threshold of network matched-filtered signal-to-noise ratio (SNR) greater than 10. Since we are interested in binary black holes (BBHs), we further restrict our sample to those events that have a probability of being a BBH merger greater

than 0.5[1].

Zheng et al. (2023) restrict their sample to events detected in all three detectors that were in science mode during the LVK observing period, namely LIGO Livingston, LIGO Hanford (LIGO Scientific Collaboration et al., 2015) and Virgo (Accadia et al., 2012), to get better-localised events. However, this step removes a significant fraction of the BBHs selected above. Since it is not clear a priori whether the resulting gain in sky localisation accuracy would compensate for the reduction in the BBH sample size, we keep two-detector events in our final sample. To ensure better homogeneity in the sky localisations of the BBHs, we remove all events from our sample detected before the Virgo detector joined the observing run. Note that a non-detection in one of the detectors does carry some information about the location of the merger event in the sky. Hence, the two-detector events not detected in Virgo when it is in science mode are still expected to be better localised than those observed in the absence of Virgo.

Finally, we are left with 53 BBHs that constitute our observed catalogue. Using the parameter estimation posterior samples on declination and right ascension made publicly available by the LVK collaboration, we generate skymaps for each event representing the uncertainty in their localisation in the sky. Figure 3.1 shows a combined skymap of all events generated by stacking the individual skymaps. We summarise the properties of the observed BBHs in Figure 3.2.

3.1.2 Mock BBH Catalogue

In this section, we describe the procedure to create the mock BBH catalogue that will serve as the set of 'randoms' used for the clustering analysis. At first, it appears that simply distributing points uniformly in the sky should be sufficient, as the resulting data set would be unclustered and uncorrelated with large-scale structure. This would be a valid approach if we had perfect observations of a sample of BBHs representative of the entire BBH population in the universe. Unfortunately, due to the limited sensitivity of the current gravitational wave detectors, the data are plagued with selection biases and systematic effects; the BBHs selected for this study do not constitute a representative sample of the population[2]. Furthermore, the detectors are not equally sensitive to all regions in the sky; each detector is most sensitive to the merger events that go off

[1]The classification probabilities were calculated using the `GW` package of `PESummary` (Hoy & Raymond, 2021).

[2]It should be noted, however, that the third generation of gravitational wave detectors is expected to detect almost all BBH mergers in the universe up to a very high redshift (Iacovelli et al., 2022; Borhanian & Sathyaprakash, 2022; Hall & Evans, 2019).

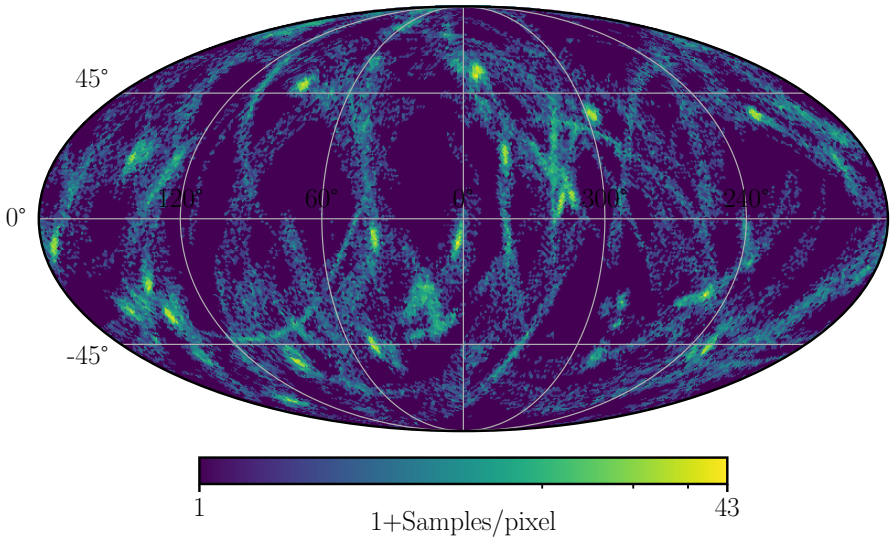


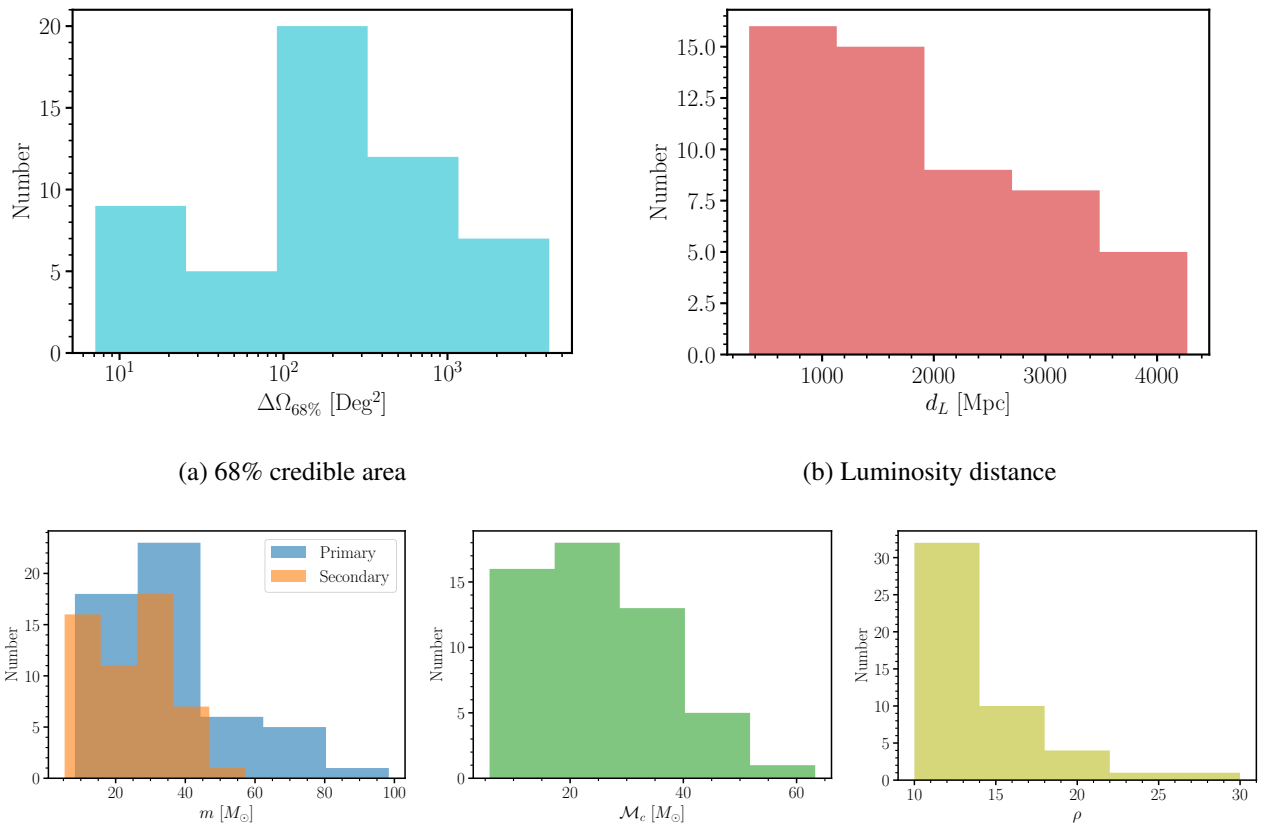
Figure 3.1: Mollweide projection of the combined skymap of the 53 observed events in equatorial (J2000) coordinates. Each banana-shaped cloud represents a single BBH. Skymaps were generated using parameter estimation posterior samples for each event through the `Healpy` package. The colour represents the number of posterior samples per pixel in a logarithmic scale. The HEALPix `NSIDE` for this map is 64.

directly on top of it, i.e., perpendicular to the plane of detector arms. This is a consequence of the transverse nature of gravitational waves. As a result, there is a selection function in the sky for the detector network as a whole (see Chen et al. (2017) for example). These observational systematics have to be carefully folded into the clustering analysis to avoid getting biased or spurious signals.

One way of mitigating the selection biases outlined above is to create realistic mock BBHs that reproduce the properties of the observed BBH sample in a statistical sense but which are inherently unclustered and spatially uncorrelated with the large-scale structure of the universe. Such a mock data set allows us to naturally incorporate the effects of observational biases on the clustering measurements.

We follow a procedure outlined in Zheng et al. (2023) to create our mock catalogue. First, we distribute a population of BBH merger events isotropically in the sky by sampling their locations from a uniform distribution ($\mathscr{U}$), which translates to drawing their right ascension (α) from $\mathscr{U}(0, 2\pi)$ and sine of declination ($\sin\delta$) from $\mathscr{U}(-1, 1)$. We next draw their source parameters from the population distributions inferred by the LVK collaboration (LIGO Scientific Collaboration et al., 2023a) as implemented by the `GWPopulation` package[3]. These are as follows:

[3] https://colmtalbot.github.io/gwpopulation/



(a) 68% credible area

(b) Luminosity distance

(c) Component mass

(d) Chirp mass

(e) SNR

Figure 3.2: The top panel shows the distribution of the BBH properties most relevant for clustering, namely the 1σ sky localisation uncertainty areas (*left*) and luminosity distances (*right*) of the observed BBH catalogue. The credible areas are computed using the rapid Bayesian localisation code `BAYESTAR` (Singer & Price, 2016). The bottom panel shows the distribution of the component masses (*left*), chirp masses (*middle*) and SNRs (*right*) of the observed BBH catalogue. In the left panel, the primary (heavier) BBH mass distribution is shown in blue while the secondary (lighter) BBH mass distribution is shown in orange. The characteristic peak at $\sim 30 M_\odot$ is clearly visible.

1. `Power Law + Peak` model for the mass of the primary (heavier) BBH and a power law distribution for the ratio of component masses (Talbot & Thrane, 2018)

2. power law distribution for redshift evolution of merger rate per unit comoving volume per unit source-frame time (Fishbach et al., 2018)

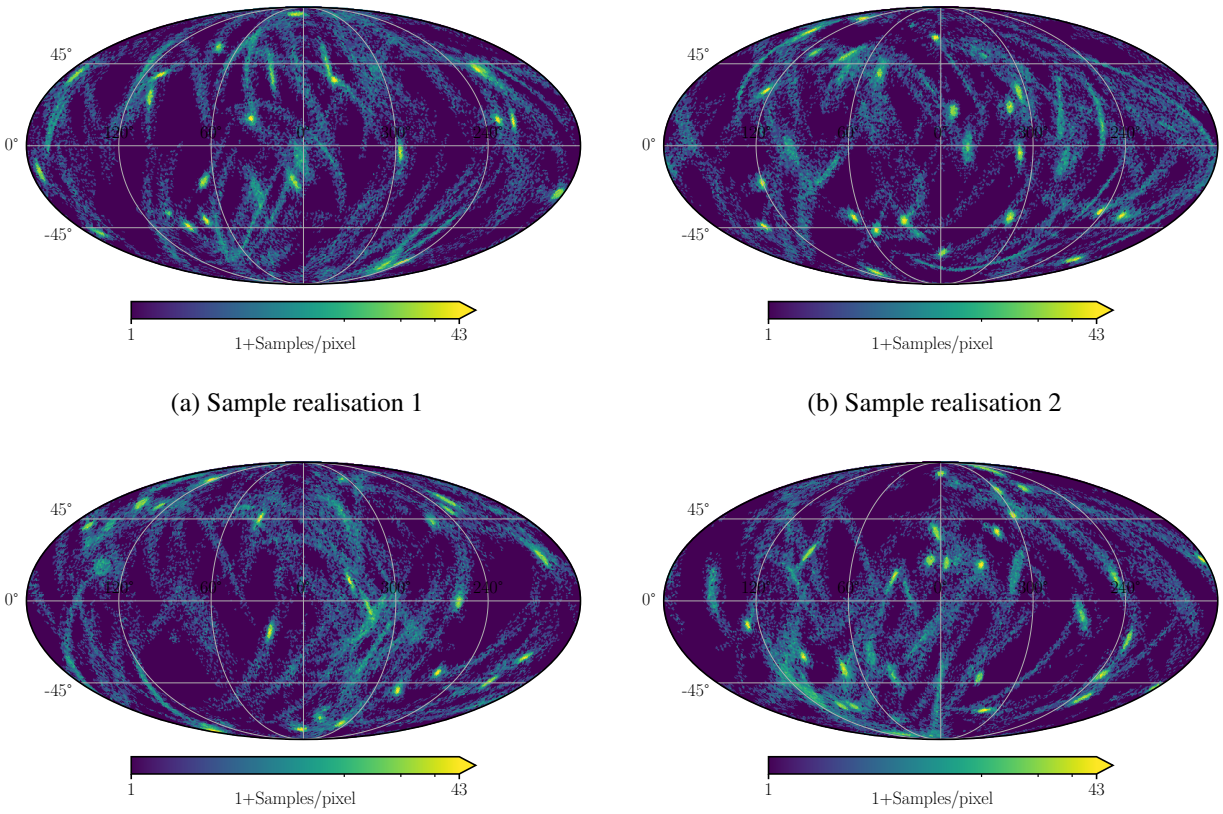
The mathematical details of these models are discussed in appendix B. We assume uniform distributions for inclination angle ι w.r.t. the plane of orbital angular momentum, polarisation angle ψ and phase at coalescence Φ_c over their allowed physical ranges, i.e., $\iota, \psi \in \mathscr{U}(0,\pi)$ and $\Phi_c \in \mathscr{U}(0,2\pi)$, and uniformly sample the BBH merger time during the LIGO observation period after Virgo started taking data. For simplicity, we set the black hole spins identically to zero since we do not expect them to affect the clustering properties or the sky localisation uncertainties, which are most relevant to us. We have further checked that including the spins does not affect our analysis; the final mock BBH catalogues with and without spins turned on are statistically similar in all aspects.

After creating the mock BBH population, we determine which events can be 'detected' by the current gravitational wave detectors to reproduce the selection biases in the data. Here, we consider a detector network consisting of LIGO Livingston, LIGO Hanford and Virgo, the same network that collected the data for our observational sample. We conduct the following gravitational wave data analysis using the rapid Bayesian localisation code for gravitational wave events, `BAYESTAR`[4] (Singer & Price, 2016). First, we simulate the gravitational wave signals for each BBH using the `IMRPhenomXPHM` model (Pratten et al., 2021), which is the same waveform used in the LVK analysis of the data. Next, we inject the simulated signals in stationary Gaussian noise created using analytic estimates for the third observing run power spectral densities for the LIGO Livingston, LIGO Hanford and Virgo detectors, as provided by the `PyCBC` package[5]. Finally, we compute the (phase-maximised) network matched-filtered signal-to-noise (SNR henceforth) for each event and classify the events with an SNR ≥ 10 as 'detections'[6]. Once we have the selected events, we use `BAYESTAR` to localise them. `BAYESTAR` also returns the estimated luminosity distances and credible intervals for the area of sky localisation uncertainty.

[4]We follow a similar procedure to the one outlined in https://lscsoft.docs.ligo.org/ligo.skymap/quickstart/bayestar-injections.html.

[5]http://pycbc.org/pycbc/latest/html/index.html

[6]Technically, the false alarm rate (FAR) is a better measure of whether an event should be considered detectable, but computing FARs is computationally expensive, as it requires doing parameter estimation for the full set of injected events. An SNR cutoff of 10 is a reasonable proxy (see (LIGO Scientific Collaboration et al., 2023b) or (Essick, 2023) for more details).



(a) Sample realisation 1

(b) Sample realisation 2

(c) Sample realisation 3

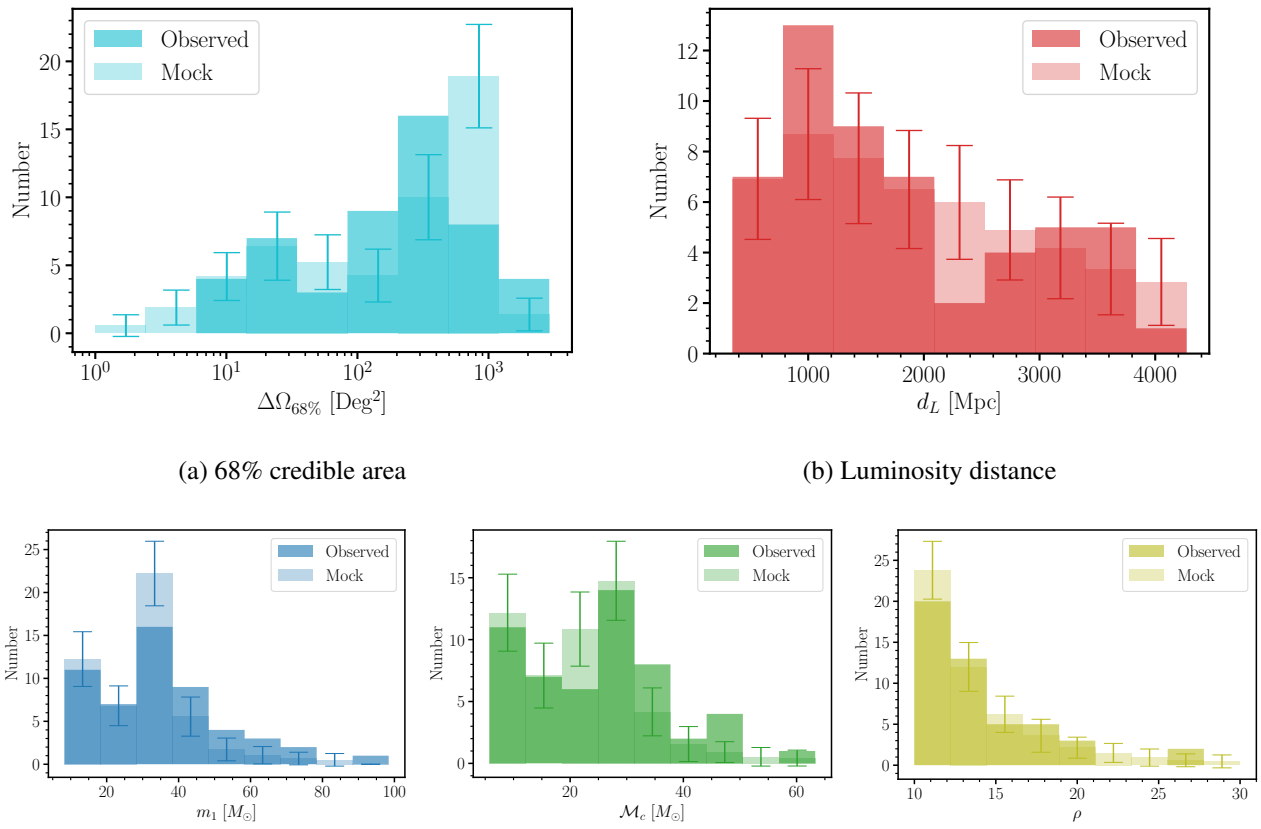
(d) Sample realisation 4

Figure 3.3: Mollweide projection of the combined skymap of the 4 sample realisations of the mock BBH catalogue in equatorial (J2000) coordinates. Each banana-shaped cloud represents a single BBH. These are visually similar to figure 3.1. Skymaps were generated using `BAYESTAR`. As before, the colour represents the number of posterior samples per pixel in a logarithmic scale. The HEALPix `NSIDE` for these maps is 64.

Using the procedure outline above, we generate a mock catalogue of 135 realisations of 53 BBHs each. The combined skymaps of 4 sample realisations are shown in figure 3.3, with the colour palate and resolution identical to figure 3.1 for ease of comparison. The skymaps of the observed and mock BBHs are visually similar.

To investigate if the mock catalogue is statistically similar to the observational data, we compare the distribution of BBH properties, averaged over the 135 mock realisations, with the corresponding distributions measured in the data. The results are shown in figure 3.4.

Within the limit of sample variance across the realisations, the mock catalogue is reasonably statistically similar to the observed BBH sample. However, it includes more events with higher



(a) 68% credible area

(b) Luminosity distance

(c) Primary mass

(d) Chirp mass

(e) SNR

Figure 3.4: The top panel shows the distribution of the 1σ sky localisation uncertainty areas (*left*) and luminosity distances (*right*) of the observed (bold histograms) and mock (light histograms) BBHs, while the bottom panel shows the distribution of the primary masses (*left*), chirp masses (*middle*) and SNRs (*right*) of the observed and mock BBH catalogue. The error bars on the mock histograms show the variance over 135 realisations of the mock catalogue. Except for one or two bins, the mocks and the data agree reasonably within the error bars, signifying that the mock dataset statistically reproduces the observations.

uncertainty in the sky localisation, as seen from the $\sim 3\sigma$ deviation from the observed distribution in the second last histogram bin of figure 3.4a. However, this is not expected to bias the clustering results since it would only lead to slightly larger measurement errors for the clustering statistics of each mock realisation. As we discuss in section 3.3.2, the relevant quantity for measuring the significance of the clustering signal is the variance across the realisations, which is independent of the measurement errors on the individual realisations.

3.1.3 Galaxy Catalogue

We use galaxies and quasars from the publicly available WISE×SuperCOSMOS (hereafter WSC) catalogue (Bilicki et al., 2016), which is a cross-match between two parent full-sky catalogues: the AllWISE release (Cutri et al., 2013) from the Wide-field Infrared Survey Explorer (WISE) (Wright et al., 2010), a mid-infrared ($\lambda \sim \mu m$) space survey; and the SuperCOSMOS Sky Survey (Hambly et al., 2001), consisting of data from digitised optical photographic plates taken by the United Kingdom Schmidt Telescope (UKST) in the southern hemisphere[7] and the Palomar Observatory Sky Survey-II (POSS-II), in the northern hemisphere (Reid et al., 1991). WISE, a NASA space-based mission, surveyed the entire sky in four bands, $W_1 = 3.4\mu m$, $W_2 = 4.6\mu m$, $W_3 = 12\mu m$, and $W_4 = 23\mu m$, while SuperCOSMOS has data in three optical bands, B, R, and I. The interested reader is referred to Bilicki et al. (2016) for more details on the WISE and SuperCOSMOS surveys and the cross-matching procedure. We use photometric redshifts for the WSC catalogue provided by Bilicki et al. (2016), which have been estimated using the artificial neural network code `ANNz` (Collister & Lahav, 2004).

We work with the 'SVM' release of the WSC catalogue (Krakowski et al., 2016), which classifies sources into galaxies, stars and quasars using a support vector machines (SVM) learning algorithm. The reason for choosing the SVM catalogue is as follows: in creating the original WSC catalogue, colour cuts were placed that already removed the quasars. Since we expect quasars to trace the large-scale fluctuations in the universe alongside galaxies, removing quasars is unnecessary for a cross-correlation study like ours. We remove the sources classified as 'stars' in the SVM catalogue and select the remaining objects to form our raw catalogue. Finally, we remove sources lying in problematic regions in the sky with unreliable data using the publicly available WSC mask to create the final catalogue cross-correlated with the BBH catalogue created in section 3.1.1. This process removes regions such as those obscured by the plane of the Milky Way and by the Large

[7] https://www.roe.ac.uk/ifa/wfau/ukstu/telescope.html

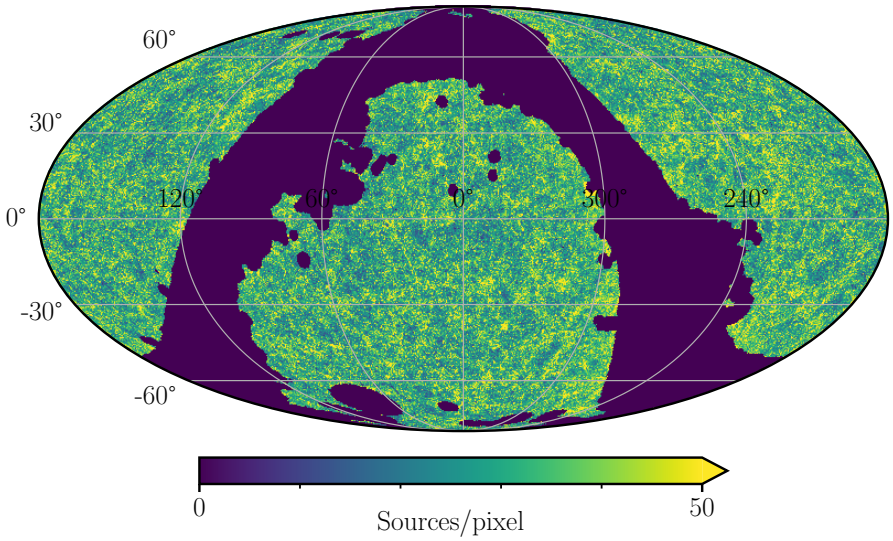


Figure 3.5: Mollweide projection of the skymap of $\sim 1.7 \times 10^7$ galaxies and quasars in the WSC catalogue, in equatorial (J2000) coordinates. Skymaps were generated from the sky locations of the sources using the `Healpy` package. The colour represents the number of sources per pixel on a linear scale. The colour bar has been limited to a maximum of 50 samples per pixel to enhance contrast, which makes it easier to visualise the large-scale structure in the distribution of the galaxies and quasars. The empty navy regions represent regions in the sky with unreliable data and have been masked out. The HEALPix `NSIDE` for this map is 256.

and Small Magellanic Clouds (SMC and LMC) and the areas with high stellar contamination (see Bilicki et al. (2016) for a detailed description of the masking procedure). Some authors (for example, Mukherjee et al., 2022) impose additional colour cuts on $E(B-V)$ and/or on $W_1 - W_2$ to mitigate dust extinction and further reduce stellar contamination. However, this increases the purity of the galaxy sample at the cost of completeness, which is undesirable for cross-correlation studies. Since we do not expect nearby stars to correlate with the extragalactic BBHs, we do not impose any additional colour cuts.

After masking out regions in the sky with unreliable data, we are left with a catalogue that covers $\sim 3\pi$ steradians in the sky, corresponding to a sky coverage of $\sim 68\%$. This makes this catalogue suitable for a cross-correlation study with BBH catalogues which are inherently all sky since gravitational waves are not susceptible to medium propagation effects[8]. The distribution of the WSC galaxies and quasars is shown in figure 3.5.

Furthermore, the redshift distribution of the WSC sources significantly overlaps with that of

[8]Note that gravitation waves, like light, are indeed susceptible to weak gravitational lensing (Meena & Bagla, 2019; Oguri, 2016).

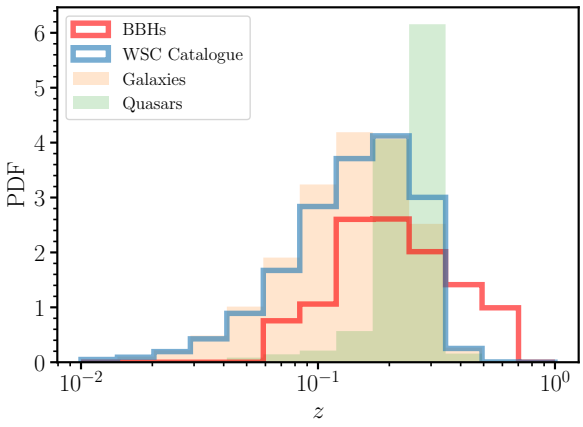


Figure 3.6: A comparison of the redshift distributions of the observed BBHs (red histogram) and the WSC catalogue sources (blue histogram). There is a significant overlap between the redshifts of the two datasets, which is crucial for conducting cross-correlation studies since we do not expect cosmological fluctuations at different redshifts to be correlated. Also plotted are filled histograms showing the distribution of galaxy (orange) and quasar (green) redshifts separately. The quasars are at a higher redshift on average.

the BBHs selected for this study, as shown in figure 3.6. This is important for cross-correlation studies since we do not expect cosmological fluctuations at different redshifts to be correlated. Note that the redshifts of the BBHs are not direct observables but are computed from the luminosity distances, assuming a cosmological model for the expansion history of the universe. For this dataset, the LVK collaboration (LIGO Scientific Collaboration et al., 2023b) assumed a cosmological model consistent with the Planck 2015 results (Ade et al., 2016).

The final catalogue contains $\sim 1.7 \times 10^7$ sources, with ~ 15 million classified as galaxies and ~ 2 million as quasars. This translates to an average number density of more than 600 sources per sq. deg. in the sky.

3.2 Application of clustering formalism to data

The formalism to study clustering developed in chapter 2 is applicable to discrete tracers that can be treated as point objects in the sky, i.e., objects that can be localised to a single sky position (δ, α). However, as shown in section 3.1, due to the limited resolving power of gravitational wave detectors, we can only assign each event with a probability distribution over an extended region in the sky (see figure 3.1 for example). Moreover, as discussed in section 3.1, certain regions in

the sky do not have reliable galaxy data. Hence, we can not define the galaxy overdensity field there. In this section, we discuss our strategy to deal with these challenges, namely the uncertain sky localisation of the BBHs and the complicating effects due to the presence of the WSC mask. In chapter 2, we described two methods of computing cross-correlations, namely the tracer-tracer and tracer-field formalisms. In this section, we also discuss which of the two is more appropriate for the specific data under consideration.

3.2.1 Strategy to deal with the uncertainty in BBH sky localisations

In this section, we discuss our strategy to deal with the extended sky localisation of the BBHs. For each BBH, we have a probability distribution in the sky. Since we need a single location for the BBHs, one possible approach is to assign each BBH the most probable position in its sky localisation area, i.e., the position where the probability distribution is maximised. However, the sky posteriors of many events show signs of bimodality, and assigning a single representative location to them is problematic. Furthermore, in reducing a probability distribution to a single point, we lose a lot of information contained in the shapes of the posteriors. For example, utilising the full sky distribution can allow us to characterise the measurement errors on the clustering strength naturally. Therefore, we adopt a different strategy in this work, which is as follows

1. for each BBH, draw an (RA, Dec) pair from the sky location posterior

2. using the drawn samples as the 'true' locations of the events, compute the auto-clustering statistics as defined in section 2.1

3. repeat steps (i) and (ii) for 1000 draws from the posteriors

The average over the 1000 draws gives the estimated value of the clustering strength, while the variance over the 1000 draws gives the estimated measurement error due to the uncertainty in the sky localisation of the BBHs. Therefore, our strategy naturally preserves the information present in the full sky distribution of each BBH while giving us an estimate of the measurement errors on the clustering strength. Vijaykumar et al. (2023b) took a similar approach in a recent clustering analysis with forecast BBH data for the third generation of gravitational wave detectors.

3.2.2 Cross-clustering: Tracer-Tracer, or Tracer-Field?

There are two possible approaches to quantify the cross-clustering between the BBHs and the WSC catalogue:

1. directly cross-correlate the WSC source positions with the BBH positions
2. cross-correlate fluctuations in the WSC source number density field and the BBH positions.

Although both approaches sound similar and ultimately capture the same physical quantity, there is a subtle conceptual difference: approach (i) treats the galaxies and quasars as discrete point sources, while approach (ii) treats the entire catalogue as a continuous field, discarding the idea of individual sources and their positions. How do we choose between the two approaches?

Given that galaxies and quasars are, in fact, discrete objects, and there is no underlying physical 'galaxy field'[9], taking approach (i) is, in principle, the correct decision, whereas taking approach (ii) would need careful justification. However, the vast discrepancy between the number densities of the WSC and BBHs makes implementing approach (i) difficult using nearest-neighbour measurements.

As discussed in chapter 2, the relevant measure of cross-correlation between two sets of tracers X and Y is the joint probability of finding $\geq k_X$ data points of X and $\geq k_Y$ data points of Y in randomly placed spherical caps of fixed angular radius in the sky. If the number density of Y is much larger than X, it can be shown that this joint probability approaches the auto k_XNN-CDF of X. Thus, if we treat the WSC galaxies as discrete tracers, we would only be able to capture the auto-clustering of the BBHs.

Conceptually, this is easy to understand: the angular scales involved in the clustering analysis are determined by the sparser tracer X and are of the order of the mean inter-particle separation of X, which is much larger than the mean inter-particle separation of Y. The probability of finding $\geq k_y$ tracers of Y in spherical caps of such large angular radii would saturate to 1. As a result, the joint probability of finding $\geq k_X$ data points of X and $\geq k_Y$ data points of Y is practically the same as the unconditional probability of finding $\geq k_X$ data points of X.

A possible solution to this problem is to down-sample the galaxies to match the number densi-

[9] Although there is a physical matter field, galaxies are biased tracers, and the galaxy positions are not results of a Poisson process on the same.

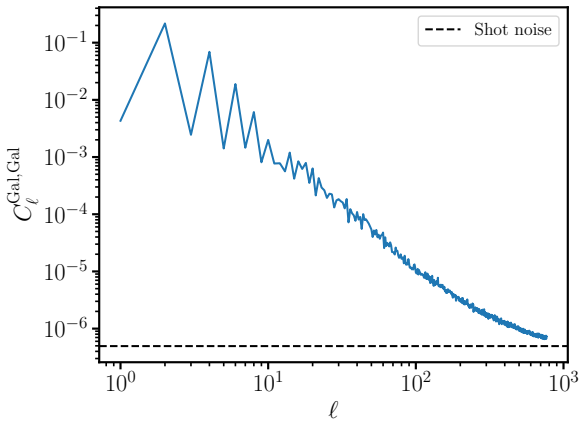


Figure 3.7: The angular power spectrum for the WSC sources computed using `healpy`'s `anafast` routine, with the shot (Poisson sampling) noise plotted for reference. The power spectrum is well above the shot noise up to an $\ell_{\mathrm{max}} \sim 10^3$, corresponding to spatial scales much smaller than those considered in this analysis. Therefore, the treatment of the WSC sources as a continuous field is a reasonable approximation.

ties of the BBHs, compute the joint CDF, and perform a bootstrap average over many realisations of the down-sampling procedure to account for sample variance. However, this step is computationally prohibitive given that we would need to average over $\sim 10^5$ bootstrap samples to get a single measurement.

In this work, we adopt approach (ii) and compute the BBH-Galaxy cross-clustering using a tracer-field cross-correlation formalism. Since the WSC catalogue has an average number density of more than 600 sources per sq. deg. in the sky, and the angular scales involved in our analysis are of the order of 1° or larger, treating the galaxy number density as a continuous field is a reasonable approximation. Furthermore, as shown in figure 3.7, the Poisson sampling noise, or shot noise, is subdominant to the angular power spectrum at all scales of interest. This means that the (fictitious) underlying galaxy density field is well-sampled by WSC source positions. Therefore, taking approach (ii) is justified.

3.2.3 Strategy to deal with the WSC Mask

The computational procedure outlined in chapter 2 gives unbiased measurements for the tracer-field cross-correlations only when the continuous field is defined on the entire sky. However, we do not have reliable data in regions outside the WSC mask. How do we compute cross-correlations

with the BBHs in this scenario? One possible approach is to assign a $\delta_{\rm Gal} = 0$ to all pixels in the masked region since that is the expected average value of an overdensity field. Although this would not bias the results since the BBHs outside the mask would not contribute to the cross-correlation signal, it would unnecessarily add to the noise budget. Instead, we take the approach of removing the BBH events that lie outside the mask. However, since the BBHs are not perfectly localised, we must be careful while handling the events whose sky localisation areas are partly inside the mask. We follow the following strategy:

1. Draw 1000 samples of 53 (RA, Dec) pairs from the sky location posteriors of the BBHs
2. for each sample, remove the sky locations that lie outside the WSC mask

By keeping the posterior samples for events which are partly outside the mask, our method not only preserves the information in the sky distribution of the BBHs but also leads to more number of BBHs contributing to the analysis than simply removing all BBHs whose localisation area intersects the mask would. However, this process leads to different tracers in each sample. Since the *k*NN-CDFs are highly sensitive to the number density of the tracers (see equation 2.14), care needs to be taken to ensure that averaging the CDFs over samples with different number densities does not lead to any issues. An important check is whether the distribution of the number of events inside the mask over the 1000 samples for the mock catalogue generated in section 3.1.2 is statistically similar to that of the data. Figure 3.8 shows that, indeed, that is the case, and hence any systematics that arise due to this would affect the clustering of the observed and mock BBHs equally.

We now have positions for the BBHs, from which we can compute the overdensity fields needed to compute the cross angular power spectrum. Since there is no data outside the mask, the value of the overdensity field computed there would be artificially low (negative). To account for this, we set the BBH overdensity fields outside the mask to zero before calculating their spherical transforms.

To compute the BBH-Galaxy cross-correlation, we need to smooth the galaxy density field on various spatial scales. To minimise the effects due to the mask, we set the field outside the mask to zero before smoothing. Moreover, we take the following precautions to avoid biasing the measurement of the cross-correlation signal:

1. We compute the two-point cross-correlation function for a large number of randomly sampled points inside the mask and subtract this from the two-point function computed for the

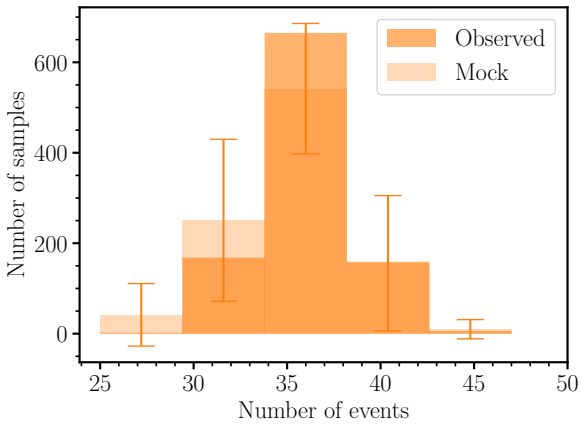


Figure 3.8: Distribution of the number of unmasked events over 1000 samples drawn from the sky distributions of the observed (bold histogram) and mock (light histogram) BBHs. The error bars on the mock histogram show the variance over 135 realisations of the mock catalogue. There is an excellent match between the two distributions within error bars. Therefore, any systematics that arise due to differing number densities between different samples would affect the clustering of the observed and mock BBHs equally.

tracers. This ensures that the two-point function of the tracers is unbiased, as any systematic effects get cancelled out during the subtraction, whereas the true clustering signal is preserved[10].

2. We restrict the query points to inside the mask. This is needed because the query points outside the mask would have artificially large nearest-neighbour distances and would skew the measured distributions. We also remove all query points within a certain angular distance from the mask boundaries to ensure that the smoothed density field interpolated at the query points is not affected by spurious contributions from the regions outside the mask. Wang et al. (2022) followed a similar procedure to analyse the spatial clustering of SDSS clusters using *k*NN-CDFS in a recent study. In practice, we observe that a threshold distance of roughly half the maximum angular scale used in the analysis leads to an unbiased measurement of the excess cross-correlation.

Note that these steps are not necessary for computing auto-clustering of the BBHs as we have BBH data on the entire sky.

[10]Note that this is similar to how random points are used in the Landy-Szalay estimator for the auto-correlation function.

3.3 Hypothesis-testing Framework

Now that we have a pipeline for computing the summary statistics that quantify the clustering of BBHs and their spatial cross-correlations with the WSC sources, how do we interpret the statistical significance of such measurements performed on data? We address this problem using a hypothesis-testing approach: we consider a null hypothesis, which proposes that there is no statistical significance for a clustering signal in the data, and attempt to rule it out by investigating the likelihood of reproducing the observed data, assuming the null hypothesis to be true. To test the null hypothesis, we require a control dataset consistent with its premise. We already described the procedure for creating a catalogue of unclustered mock BBHs in section 3.1.2. This mock catalogue automatically serves as a control set to test the null hypothesis.

We describe our null hypothesis in section 3.3.1 and discuss the method to calculate the statistical significance of the clustering signal in section 3.3.2.

3.3.1 Null Hypothesis

Our null hypothesis is as follows

> *The BBHs currently detected by the LVK collaboration are spatially unclustered, distributed uniformly (isotropically) in the sky and are not spatially correlated with other tracers of the large-scale structure of the universe, such as galaxies and quasars.*

Any dataset consistent with this hypothesis would not contain a statistically significant clustering signal.

3.3.2 Statistical significance

In this section, we describe the way we compute the statistical significance of the clustering measurements once the summary statistics have been computed for the observed and mock BBHs. Consider a summary statistic as a function of angular scale, $S(\theta)$, which could be the angular power spectrum, the two-point function or the nearest-neighbour distribution, either as a measure

of the BBH auto-correlation or BBH-Galaxy cross-correlation. Let the scales considered in the analysis be $\{\theta_1, ..., \theta_p\}$. We define the *data vector* D_a as the summary statistic S evaluated on the observed BBH catalogue at angular scale θ_a. Similarly, we define a *mock vector* M_b^i as $S(\theta_b)$ evaluated on the i^{th} realisation of the mock BBH catalogue for each of the n realisations. Note that $S(\theta)$ represents the summary statistic already averaged over 1000 samples drawn from the sky distribution of the BBHs, as prescribed in section 3.2.1. To characterise the noise properties of the measurement, we compute the *covariance matrix*

$$\Sigma'_{ab} = \left\langle \left(M_a^i - \langle M_a \rangle\right) \left(M_b^i - \langle M_b \rangle\right) \right\rangle \tag{3.1}$$

where the angular brackets denote an average over the n realisations of the mock catalogue. The covariance matrix is, by definition, a $p \times p$ matrix. The object that is relevant for the statistical calculations is the inverse of the covariance matrix, which is multiplied by the Hartlap correction factor (Hartlap, J. et al., 2007) to get an unbiased estimate

$$\Sigma^{-1} = \frac{n-p-2}{n-1} \left(\Sigma'\right)^{-1} \tag{3.2}$$

Once we have the corrected inverse covariance matrix, we characterise the signal-to-noise for clustering using the χ^2 statistic. For the observed BBHs and each realisation of the mock BBHs, we define the χ^2 value as

$$\chi^2_D = \left(D - \langle M \rangle\right)^T \Sigma^{-1} \left(D - \langle M \rangle\right) \tag{3.3}$$

$$\chi^2_{M^i} = \left(M^i - \langle M \rangle\right)^T \Sigma^{-1} \left(M^i - \langle M \rangle\right) \tag{3.4}$$

where $D \triangleq \{D_1, D_2, ..., D_p\}$ and $M^i \triangleq \{M_1^i, M_2^i, ..., M_p^i\}$. The distribution of $\chi^2_{M^i}$ (henceforth the null distribution) represents the signal-to-noise expected from data consistent with the null hypothesis, and χ^2_D represents the signal-to-noise measured from the data. A larger value for χ^2_D relative to the null distribution implies a stronger statistical significance of the clustering signal.

From the null distribution and the measured signal-to-noise, we compute the p-value, or probability of reproducing the observations assuming the null hypothesis is true, by estimating the area enclosed under the (normalised) null distribution curve after it crosses the measured signal-to-noise. In practice, this can be estimated by counting the fraction of mock realisations with $\chi^2_{M^i} > \chi^2_D$. If the signal is strong enough that none of the mock realisations have a larger χ^2 than the data, then one must fit a χ^2 distribution to the null distribution to compute the p-value. The

null hypothesis is ruled out if the p-value is smaller than a chosen detection threshold.

3.4 Angular scales

We conduct the clustering analysis on angular distance scales from $\sim 1^\circ$ to $\sim 35^\circ$, equivalent to $\ell = 6$ to $\ell = 180$. This choice ensures that

1. we have sufficient sampling in the nonlinear regime, where the nearest-neighbour distributions can capture information not accessible through two-point statistics
2. the measurements are not affected by the lack of sampling towards the right tail of the auto kNN-CDF (see Banerjee & Abel (2021a) for a discussion)

These angular distances correspond to projected transverse distance scales of ~ 15 to ~ 400 Mpc for a median redshift of ~ 0.2 for the WSC catalogue. We choose 10 log-spaced angular bins and 10 linearly-spaced ℓ bins in the given range, which leads to a Hartlap factor of 0.92 for 135 realisations of the mock catalogue.

For computing the overdensity fields and the query points for the nearest-neighbour measurements, we use an `NSIDE` = 256 HEALPix grid with $\sim 7.8 \times 10^5$ pixels and an angular resolution of $\sim 0.22^\circ$. As required for the nearest-neighbour analysis, the number of query pixels is much larger than the number of data points, and the query grid has sufficient resolution to sample the smallest spatial scales analysed.

As discussed in section 3.2.3, we remove all query points within 20° of the WSC mask boundaries for computing nearest-neighbour excess cross-correlation to avoid any biases due to the presence of the WSC mask.

3.5 An illustrative example

As discussed in chapter 2, the nearest-neighbour distributions are sensitive to all higher-order cross-correlation functions of the discrete tracers and the continuous field. Consequently, nearest-neighbour measurements are expected to measure a stronger clustering signal significant than the

two-point summary statistics at small angular scales where the underlying fields are non-Gaussian, and these higher-order correlation functions cannot be neglected. In this section, we demonstrate the gains in clustering power obtained using the kNN tracer-field formalism compared to the angular power spectrum using an illustrative example.

We create a set of discrete tracers by picking the centres of 36 pixels in the WSC footprint[11] that have the highest number density of sources in the sky, and compute their spatial cross-correlations with the full WSC overdensity field using both the nearest-neighbour measurements and the angular power spectrum. We further assume that we know the locations of these tracers perfectly. To estimate the cosmic variance associated with the clustering measurements, we also compute these clustering statistics for 100 realisations of 36 randomly chosen points in the sky. We compute the cross-correlations on angular distance scales from $\sim 1^\circ$ to $\sim 35^\circ$, equivalent to $\ell = 6$ to $\ell = 180$.

The results for the first nearest-neighbour distribution are shown in the left panel of figure 3.9, and those for the power spectrum are shown in the right panel. The inset in the right panel presents a zoomed-in version of the full subplot focusing on the smaller scales. In each plot, the solid line shows the excess cross-correlation between the highest-density locations, the shaded band represents the cosmic variance, and the dash-dot line represents the expected value in the absence of cross-correlation.

The excess cross-correlation as measured by the first nearest-neighbour distribution lies well outside the shaded region representing the 3σ cosmic variance in the randoms at all angular scales smaller than $\sim 4^\circ$, whereas the power spectrum fluctuates within the shaded band, barely crossing the detection threshold on one or two bins. Thus, figure 3.9 clearly demonstrates that the nearest-neighbour measurements are able to capture a statistically significant clustering signal on small, nonlinear scales for a well-localised sample of rare, highly biased tracers, whereas the power spectrum can not.

We would like to draw the attention of the reader to the following important implication of the example studied above: *with a set of < 50 well-localised events, the nearest-neighbour measurements on small spatial scales are statistically robust enough to investigate whether BBHs reside in highly biased environments in the universe*. It should be noted however that on large scales, the nearest-neighbour distributions do not perform any better than the power spectrum, and there is no detectable signal in either statistic. This is because, on large scales, the galaxy density field is well approximated by a Gaussian random field, and the higher-order correlation functions are

[11] This is chosen to match the average number of mock BBHs in the WSC footprint, see figure 3.8 for details.

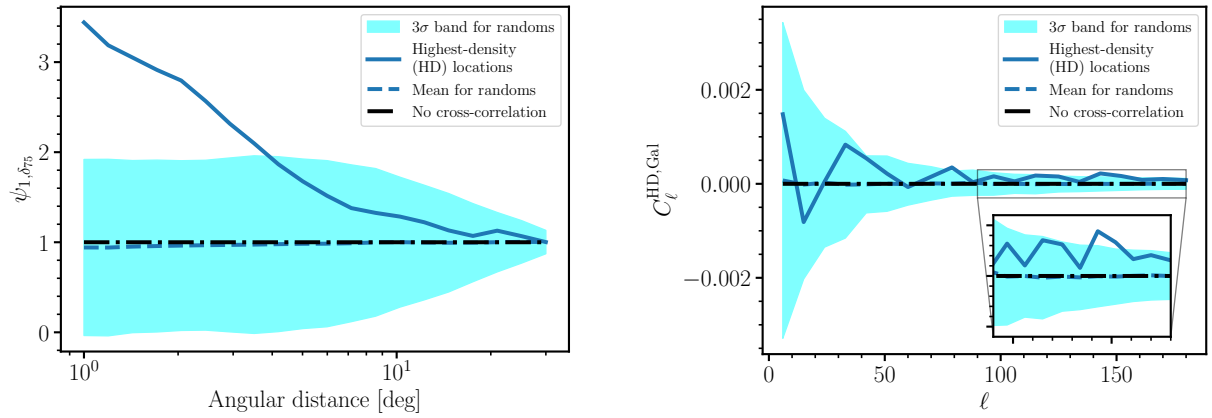


Figure 3.9: The excess cross-correlation between 36 locations with the highest density of WSC sources and the WSC overdensity map as measured by the first nearest-neighbour distribution (left) and the cross angular power spectrum (right). The inset on the right panel shows a zoomed-in view of the power spectrum at smaller scales. The solid lines represent the measurement on the tracers, the shaded band represents the 3σ variance in the same measurement performed on randomly chosen points in the sky, and the dash-dot line represents the expected value in the absence of cross-correlation.

negligible.

3.6 Results

In this section, we present the results of our clustering analysis on the data, with section 3.6.1 devoted to the auto and cross angular power spectrum and section 3.6.2 to the nearest-neighbour measurements. We discuss the implications of our findings in section 3.7

3.6.1 Angular power spectrum

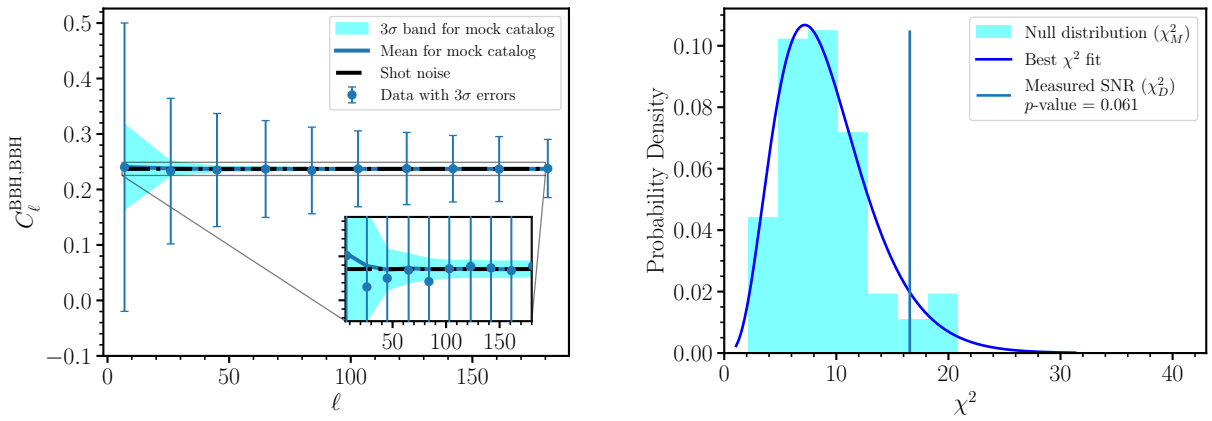
Figure 3.10a shows the angular power spectrum of the BBHs. Filled circles represent the power spectrum of the observed BBHs, and the error bars are measurement errors, defined as 3 times the standard deviation across 1000 samples drawn from the BBH skymaps. The bold line and shaded band show the mean angular power spectrum and 3σ variation around the mean for 135 realisations of the mock BBH catalogue. The dash-dot line represents the Poisson sampling noise (shot noise)

corresponding to the number density of the BBH sample, which is equal to $1/(\bar{n}_{\rm BBH})$. Since the error bars make it difficult to visualise the shaded band, we plot a zoomed-in version of the full figure in the inset.

The mean power spectrum for the mock catalogue is very close to the shot noise, as should be the case for a dataset consistent with the null hypothesis. The variance over realisations of the mock catalogue gives an estimate of the cosmic variance expected in the power spectrum if the null hypothesis holds. It is evident from the figure that with the present number of BBH detections, the angular power spectrum is shot noise-dominated and cannot capture a clustering signal if any is present. As can be seen from the figure, the measurement errors on the data are extremely large and even exceed the cosmic variance. This is a consequence of the significant uncertainties in the sky-localisation of the BBHs.

The χ^2 significance test results for the angular power spectrum are presented in figure 3.10b, with the histogram representing the null distribution and the solid vertical line representing the measured χ^2 for the observed BBHs. The data is consistent with the null hypothesis with a p-value of 0.061, calculated by fitting a χ^2 function to the null distribution. Note that the relatively small p-value is most likely caused by the large (3σ) fluctuation at $\ell \sim 80$. As clear from figure 3.10a, this fluctuation does not indicate a clustering signal. Hence, we conclude that *the angular power spectrum does not capture a statistically significant clustering signal in the presently available BBH data*.

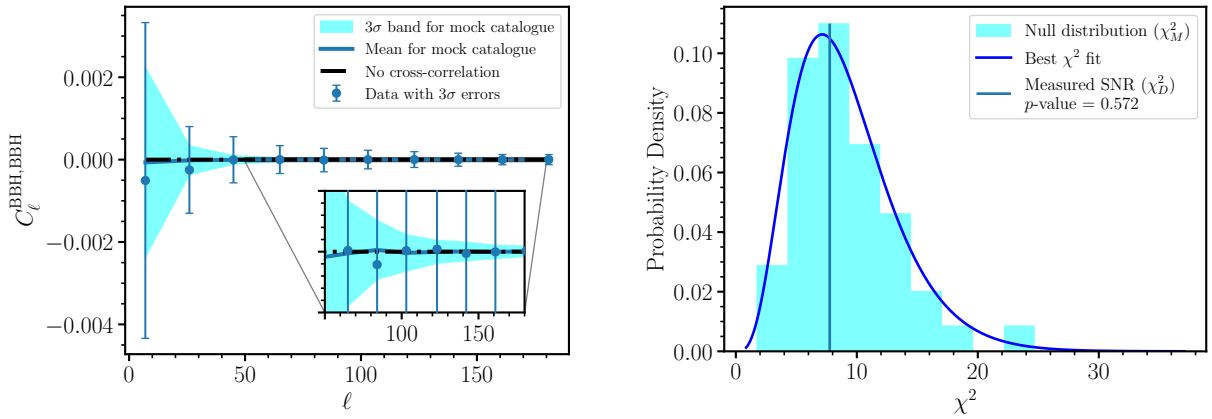
Figure 3.11a shows the cross angular power spectrum measurements between the BBHs and the WSC sources, with the plotting scheme identical to figure 3.10a. Again, we plot a zoomed-in version in the inset for better visibility of the shaded band. The mean power spectrum for the mock catalogue is very close to 0 at all scales, as expected from the null hypothesis, which stipulates that the BBHs and WSC sources are spatially uncorrelated (the cross power spectrum is unaffected by shot noise). Even for cross-correlation, the measurement errors on the data are extremely large and often exceed the cosmic variance. The χ^2 significance test for the cross angular power spectrum, summarised in figure 3.11b, confirms the visual impression given by figure 3.11: the data is consistent with the null hypothesis at a p-value of 0.572, implying that *the cross angular power spectrum does not capture statistically significant evidence for spatial cross-correlation between the presently observed BBHs and the WSC sources.*



(a) Data vectors

(b) χ^2 distributions

Figure 3.10: *Left*: Results of auto-clustering analysis conducted using the angular power spectrum as the summary statistic, with a zoomed-in view provided in the inset for better visibility. Filled circles represent the measured angular power spectrum of the observed BBHs, with error bars representing variance across 1000 samples drawn from the BBH skymaps. The bold line represents the angular power spectrum averaged over 135 realisations of the mock BBH catalogue. The shaded band, which shows the variance of the power spectrum across realisations, represents the cosmic variance. The dash-dot line shows the shot noise. All errors are displayed at the 3σ level. It is evident from the figure that with the present number of BBH detections, the angular power spectrum is shot noise-dominated; visually, there are no signs of a clustering signal. *Right:* Results of the statistical significance test for the auto angular power spectrum of the BBHs. The histogram represents the distribution of χ^2 values over 135 realisations of the mock catalogue, and the curve enveloping it represents the best-fit χ^2 distribution. The vertical line represents the measured χ^2 value for the data. The data is consistent with the null hypothesis at a p-value of 0.061, calculated using the CDF of the best χ^2 fit. There is no evidence for a statistically significant clustering signal in the present data.



(a) Data vectors

(b) χ^2 distributions

Figure 3.11: *Left:* Results of BBH-Galaxy cross-clustering analysis conducted using the cross angular power spectrum as the summary statistic, with a zoomed-in view provided in the inset for better visibility. The plotting scheme is identical to figure 3.10a, except the dash-dot line, $y = 0$, represents the cross power spectrum of two uncorrelated data sets. With the present number of BBH detections, no visual evidence exists for a clustering signal in the cross angular power spectrum. *Right:* Results of the statistical significance test for the BBH-Galaxy cross angular power spectrum. The plotting scheme is the same as figure 3.10b. The data is consistent with the null hypothesis at a p-value of 0.572, calculated using the CDF of the best χ^2 fit. There is no statistical evidence for spatial cross-correlation between the presently detected BBHs and the galaxies and quasars from the WSC catalogue.

3.6.2 Nearest-neighbour measurements

Figure 3.12 shows the first two kNN-CDFs of the BBHs in blue and orange standing for the first and second neighbours, respectively. The plotting scheme is similar to figure 3.10a, except the dash-dot line represents the expectation for the kNN-CDFs of an unclustered, Poisson-distributed dataset in the sky. The analytic expression for the kNN-CDFs of Poisson distributed points is only a function of $\bar{n}_{\mathrm{BBH}}A$, as can be seen from equation 2.14 (see also Banerjee & Abel (2021a)).

The left panel shows that $\mathrm{CDF}_{1\mathrm{NN}}$ is smaller than $\mathrm{CDF}_{2\mathrm{NN}}$ at all scales. This is intuitive since, for each query point, the distance to the k^{th} nearest-neighbour is always smaller than the distance to the $(k+1)^{\mathrm{th}}$ nearest-neighbour. As a result, for a given scale θ, the fraction of query points with the first nearest neighbour at a distance less than θ would be larger than those with the second nearest neighbour. This generally holds, i.e., the kNN-CDFs always shift towards larger angular scales with increasing k.

Since the CDFs span a large range relative to the error bars, it is difficult to visualise the results from the left panel. We plot the CDFs normalised by the mean of the mock catalogue in the right panel of figure 3.12. It is clear from these plots that the mean of the mock catalogue is consistent with the expected value for Poisson distributed data. As a consequence of the uncertainties in the sky localisation of the BBHs, the measurement errors on the CDFs of the observed BBHs are large, even exceeding the variance across the mock catalogue, similar to what was observed for the angular power spectrum in figure 3.10a.

Note how the variance across the mock realisations for the $\mathrm{CDF}_{1\mathrm{NN}}$ becomes vanishingly small as we approach the smallest angular scales. This happens because at spatial scales much smaller than the mean tracer-tracer separation, the $\mathrm{CDF}_{1\mathrm{NN}}$ approaches the expected value for a Poisson distribution regardless of the positions of the tracers. Mathematically, as the area A goes to 0, the leading order term in the summation inside the exponential in equation 2.14 dominates. Since the leading term does not contain any correlation functions, it corresponds to the CDF for a Poisson distribution. Hence, any deviations from the Poisson expression, either due to a clustering signal or due to sampling noise, are exponentially suppressed (see also Banerjee & Abel, 2021a).

The data seems to indicate no presence of a clustering signal in any of the kNN-CDFs. To confirm the visual intuition, we perform a χ^2 significance test using measurements at 5 spatial bins each from the first and second nearest-neighbour distributions and find that the data is consistent with the null hypothesis at a p-value of 0.081. Note that in this case, since the null distribution is

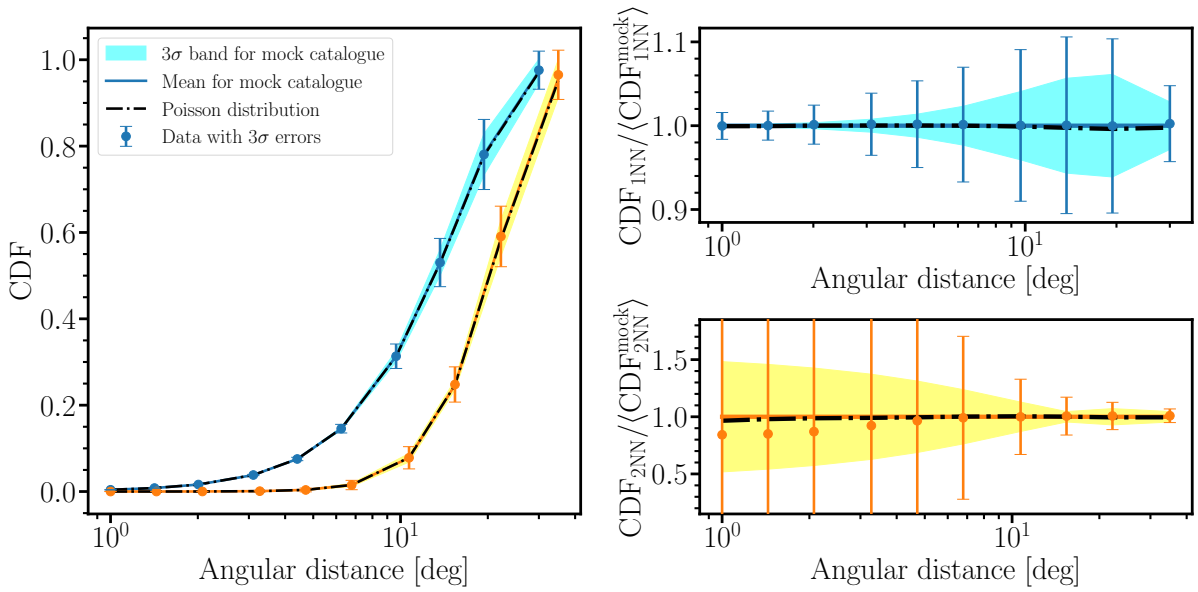


Figure 3.12: *Left panel*: Results of auto-clustering analysis conducted using the first and second nearest-neighbour cumulative distribution functions as the summary statistic. The plotting scheme is similar to figure 3.10a, except the dash-dot line represents the analytic expectation for the kNN-CDFs of an unclustered, Poisson-distributed dataset in the sky. Different colours represent different values of the nearest-neighbour index k, with blue and orange standing for $k = 1$ and 2, respectively. *Right panel*: CDFs divided by the mean over the mock catalogue to reduce the dynamic range of the plot. The error bars for the bottom plot have not been shown to the full extent to make the rest of the plot clearer. With the present number of BBH detections, there is no visual evidence for a clustering signal in the kNN-CDFs.

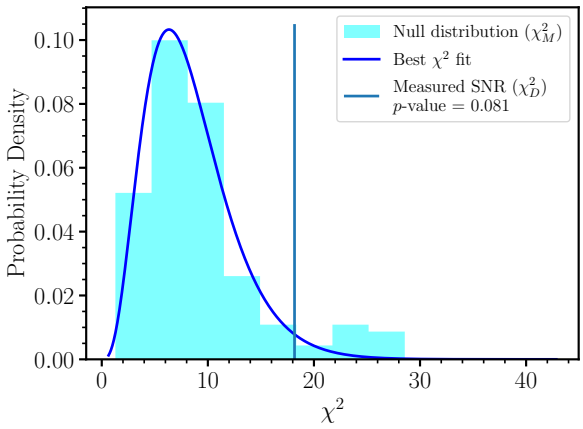


Figure 3.13: Results of the statistical significance test for the combined first and second nearest-neighbour cumulative distribution functions. The plotting scheme is the same as figure 3.10b. The data is consistent with the null hypothesis at a p-value of 0.081, calculated by counting the number of mock realisations with $\chi^2_{M^i} > \chi^2_D$ since the best chi-square fit does not characterise the right tail of the distribution well. There is no evidence for a statistically significant clustering signal in the present data.

heavy-tailed and poorly characterised by a χ^2 function, we compute the p-value by counting the number of mock realisations with $\chi^2_{M^i} > \chi^2_D$. These results are summarised in figure 3.13. We conclude that *the nearest-neighbour distributions do not capture a statistically significant clustering signal in the presently available BBH data*.

Figure 3.14 shows the excess cross-correlation between the BBHs and the quasars and galaxies from the WSC catalogue, as measured by the first and second nearest-neighbour measurements. The plotting and colour schemes are identical to figure 3.12, except the dash-dot line here represents the expected excess cross-correlation between two spatially uncorrelated datasets and is identically equal to 1 at all scales. The inset on the right shows a zoomed-in version of the full subplot for better visibility. The mean of the mock catalogue is consistent with 0, as expected from the null hypothesis. As was the case for the other summary statistics, the measurement errors on the data are extremely large even for the excess cross-correlation[12].

Interestingly, the nearest-neighbour distributions indicate a mild anti-correlation between the BBHs and the WSC catalogue at all angular scales considered, which is not picked up by the cross angular power spectrum. However, even though the anti-correlation seems to manifest in both

[12]Note that the excess cross-correlation, by definition, cannot take negative values; the error bars extending to negative values is an effect of the choice to plot the mean $\pm$ 3 $\times$ standard deviation, but the actual measurements are always positive.

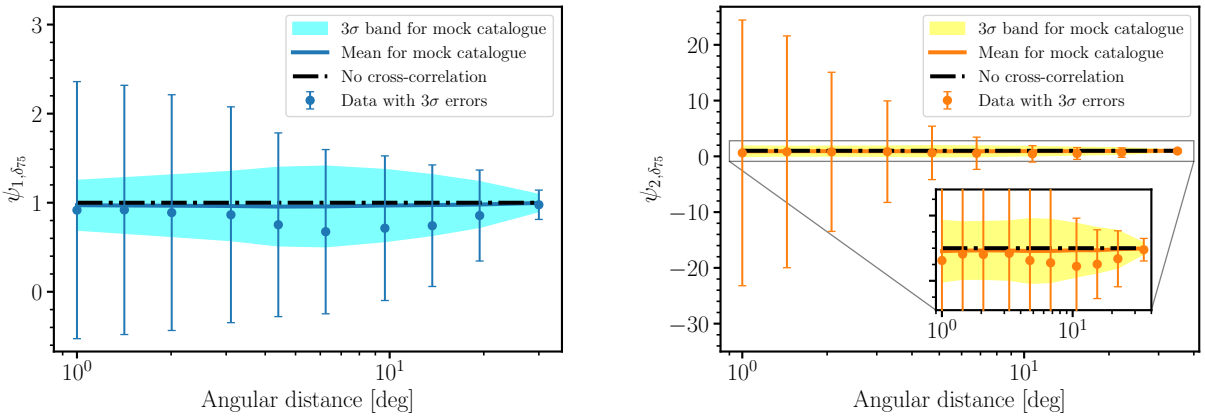


Figure 3.14: The excess BBH-Galaxy cross-correlation measured by the first (left) and second (right) nearest-neighbour measurements. The plotting scheme is identical to the right panel of figure 3.12, except the dash-dot line here represents the expected excess cross-correlation between two spatially uncorrelated datasets and is identically equal to 1. Since the error bars for the right panel make visualising the rest of the figure difficult, a zoomed-in view is provided in the inset. The plots visually indicate a mild anti-correlation between the BBHs and the WSC catalogue at all angular scales.

the nearest-neighbours systematically, extreme care needs to be taken while analysing such plots of the nearest-neighbour distributions; since nearest-neighbour measurements are cumulative, a noise-driven fluctuation on one spatial scale can affect the measurement at nearby scales, and our visual intuition can not be trusted. We need to take this into account by calculating the full covariance matrix before reaching any conclusions. Moreover, the deviation from $\psi = 1$ in each case is well within the limits of cosmic variance as characterised by the mock catalogue. This is likely an effect of sample variance due to the small number of observed BBHs, and we will return to this point in section 3.7.

We perform a χ^2 significance test using 5 spatial bins each from the first and second nearest-neighbour distributions and find that the data is consistent with the null hypothesis at a p-value of 0.444. Similar to the auto-CDFs, the null distribution here also is heavy-tailed. Hence, we compute the p-value by counting the number of mock realisations with $\chi^2_{M^i} > \chi^2_D$. Figure 3.15 shows the summary plot of this analysis. We conclude that *the nearest-neighbour measurements do not capture statistically significant evidence for spatial cross-correlation between the presently observed BBHs and the WSC sources.*

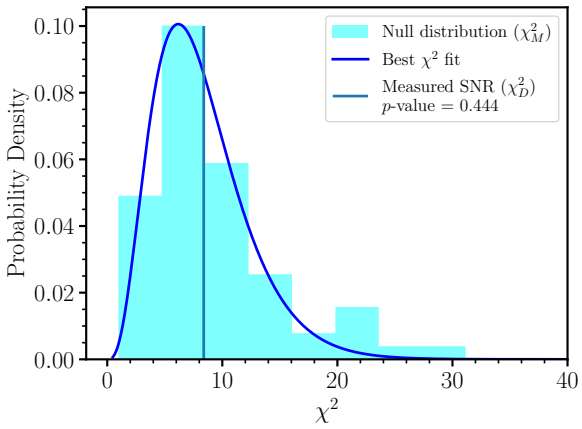


Figure 3.15: Results of the statistical significance test for the excess BBH-Galaxy cross-correlation as measured using a combination of the first and second nearest-neighbour measurements. The plotting scheme is the same as figure 3.10b. The data is consistent with the null hypothesis at a p-value of 0.444, calculated by counting the number of mock realisations with $\chi^2_{M^i} > \chi^2_D$ since the best chi-square fit does not characterise the right tail of the distribution well. This implies that despite a visual indication for an anti-correlation, there is no statistical evidence for any spatial cross-correlation between the presently detected BBHs and the galaxies and quasars from the WSC catalogue.

3.7 Discussion

As we saw in section 3.6, none of the summary statistics considered in this study were able to capture a statistically significant signal, either for the auto-clustering of BBHs or for spatial cross-correlations of BBHs with the large-scale structure of the universe, in the presently available data. Our results are consistent with previous attempts in the literature (see, for example, Zheng et al. (2023), Cavaglià & Modi (2020) and Mukherjee et al. (2022)). What explains these results? If BBHs reside primarily in galaxies, where most stars in the universe live and die, their locations are expected to be clustered and spatially cross-correlated with the observed fluctuations in large-scale structure surveys.

We believe two aspects of the data conspire to obscure the clustering signal: first, the observed BBHs constitute a statistically small sample that is susceptible to sample variance; with only $\sim$50 data points, it is very difficult to tell apart a clustered sample from a Poisson-distributed one. Second, with the current sensitivities of the gravitational wave detectors, there is considerable uncertainty in the sky localisation of the BBHs, which tends to smear out any clustering signal at small scales that the nearest-neighbour distributions are the most sensitive to. Of course, there is always the possibility that the null hypothesis is true, in which case, we would not see a clustering

signal even if we had perfect observations and a statistically large sample of BBHs.

The detection of a clustering signal in a small sample such as the one selected for this study would imply that binary black holes reside in extremely biased environments, such as highly dense nodes of the cosmic web or huge cosmic voids. As discussed in section 3.5, the nearest-neighbour distributions are statistically powerful enough to detect the clustering of rare and highly biased tracers at nonlinear scales. Unfortunately, the non-detection of a clustering signal in the current BBH data does not rule out the possibility of BBHs being highly biased tracers because the uncertainty in the sky localisations would completely wash it out even if a signal existed. Repeating this analysis with better-localised events from future gravitational wave observations would be a worthwhile exercise.

We briefly discussed in section 3.6 that the cross-correlation measurements indicate a mild anti-correlation between the observed BBHs and the WSC sources, albeit statistically insignificant. To investigate this further, we plot, in figure 3.16, the most probable sky locations of the observed BBHs on top of the galaxy overdensity field smoothed on 10° scale using a top-hat filter. Many of the observed BBHs appear to lie near large-scale underdense regions. Due to the small sample size, this is picked up as a mild anti-correlation in the nearest-neighbour distributions. However, this is not evidence for anti-correlation, as established earlier using the χ^2 test. We have further checked that a significant number of the mock realisations show similar behaviour, which is most likely a result of sample variance.

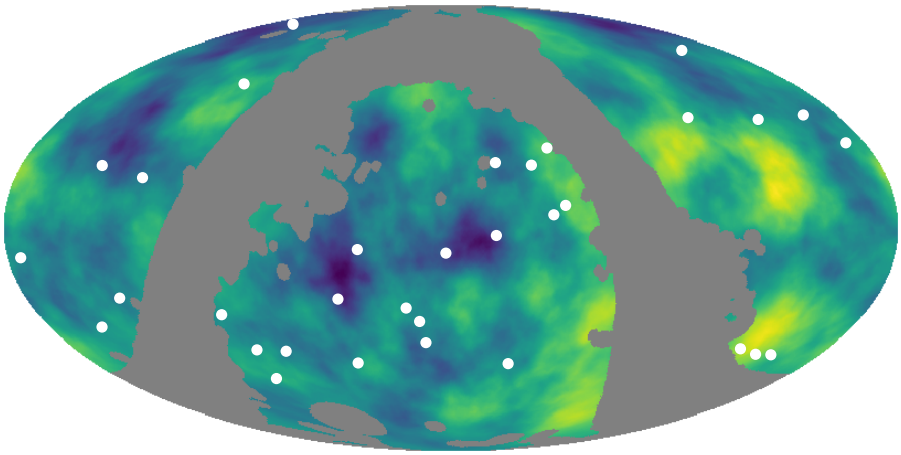


Figure 3.16: Mollweide projection of the overdensity field for the WSC catalogue smoothed on a 10° scale using a top-hat filter, with the superimposed white dots representing the most probable positions of the observed BBHs. Warmer colours represent a higher density of galaxies and quasars, while cooler colours represent underdensities. Many of the observed points happen to lie near large-scale underdensities, leading to a slight anti-correlation in the cross-clustering measurements. We believe that this may be due to sample variance since the anti-correlation is not statistically significant, and a significant number of the mock realisations show similar behaviour.

Chapter 4

Forecasts

In chapter 3, we found no evidence for spatial cross-correlations between the currently observed catalogue of binary black holes and large-scale structure tracers from the WISE×SuperCOSMOS galaxy catalogue. In this chapter, we investigate the feasibility of detecting the spatial clustering of BBHs in the coming decades using the nearest neighbour distribution. To achieve this goal, we attempt to measure the BBH-galaxy cross-clustering signal in forecast data for future LIGO observing runs and stage-IV large-scale structure surveys. Specifically, we cross-correlate the overdensity field of galaxies expected from the first year of operations of the Vera C. Rubin Observatory's Legacy Survey of Space and Time[1] (LSST Y1, Željko Ivezić et al., 2019) with the BBH catalogue expected from 10 years of gravitational wave observations by a network of 5 ground-based detectors, namely LIGO Hanford, LIGO Livingston (LIGO Scientific Collaboration et al., 2015), LIGO India (Saleem et al., 2022), Virgo (Accadia et al., 2012) and KAGRA (Akutsu et al., 2020).

4.1 Simulated Data

In this section, we describe the simulated data used to conduct the forecast study. We discuss the forecast LSST Y1 galaxy density field in section 4.1.1 and the mock BBH catalogues in section 4.1.2.

[1] https://www.lsst.org/

4.1.1 Galaxy Overdensity Field

For conducting forecast studies of angular clustering in the sky, we need simulated data that mimics observational data from large-scale surveys. Such observations typically consist of the angular positions and redshifts of tracers of structure formation, such as galaxies. The more distant these tracers are, the higher their observed redshifts are. Furthermore, as we survey deeper into the sky, we cover larger and larger cosmological volumes. For example, a survey that observes out to a redshift of 1 covers a (comoving) cosmological volume of ~ 157 Gpc^3. However, it is computationally prohibitive to simulate such large volumes; cosmological simulations typically focus on rectangular regions in space (known as a boxes) with comoving volumes of the order of 1 Gpc^3, and produce 'snapshots' of the contents of these boxes at different redshifts[2]. Therefore, to recreate the observations expected from large-scale surveys, one must arrange many such simulated boxes in spherical shells around a fictitious observer, and project the 3D positions of the simulated tracers in the boxes on the 'sky' of the observer. Each shell contains snapshots at the same redshift, and shells that are further from the observer have a higher redshift than shells that are closer. The resulting simulation product is often referred to as a *lightcone* in the literature. In this work, we use the simulated LSST Y1 galaxy overdensity fields implemented in the Agora lightcone[3] (Omori, 2022).

The Agora lightcone is constructed using data products from the MultiDark Planck 2 (MDPL2) simulation[4] (Klypin et al., 2016), which is a dark matter-only N-body simulation that contains 3840^3 dark matter particles in a 1 h^{-1} Gpc box[5]. Here, we present a brief summary of the process followed by Omori (2022) to create the lightcone from the individual boxes MDPL2. The interested reader is requested to refer to Klypin et al. (2016) for further details on MDPL2 and to Omori (2022) for a detailed description of the lightcone construction.

To construct the lightcone, Omori (2022) create a tesselation of the simulation snapshots and their associated halo catalogues using periodic boundary conditions and extract concentric spherical shells of thickness 25 h^{-1} Mpc from the tiled volume. Then, they randomly rotate the shells every 1 h^{-1} Gpc (See figure 3 of Omori, 2022) to avoid repeating structures along the line of sight. Finally, they project the dark matter particles in the boxes onto HEALPix shells of `NSIDE` = 8192.

[2] All elements of a snapshot have the same redshift.
[3] https://yomori.github.io/Agora-docs/#/
[4] https://www.cosmosim.org/metadata/mdpl2/
[5] h refers to the value of the Hubble constant assumed in the simulation, in units of 100 Km s^{-1} Mpc^{-1}

Omori (2022) compute galaxy overdensity fields corresponding to the LSST Y1 survey[6] in 5 redshift bins. For each bin, they multiply the projected dark matter overdensity shells lying in the given redshift range by the expected linear galaxy bias values for LSST Y1 galaxies, weigh the shells by the LSST Y1 redshift distribution $n(z)$ in that bin, and sum the weighted shells to obtain the desired projected galaxy overdensity field (see section 3.7 of Omori, 2022, for more details). The values for the bias and the redshift distributions of the LSST Y1 galaxies are derived using analytic relations given in the Dark Energy Science Collaboration Scientific Requirement Document (The LSST Dark Energy Science Collaboration et al., 2018)[7].

We calculated the combined redshift distribution and the combined overdensity field by performing a number-density weighted average over the redshift distribution and overdensity fields of the individual bins. The individual $n(z)$ for the 5 bins and the overall $n(z)$ for the total sample are displayed in figure 4.1.

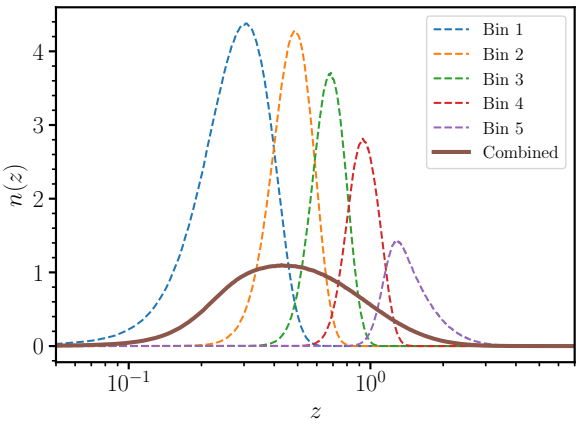


Figure 4.1: The modelled $n(z)$ for galaxies from the LSST Y1 survey. The lighter dashed lines show the distribution for data in 5 redshift bins, while the solid line shows the combined redshift distribution obtained by performing a number-density weighted average of the individual $n(z)$. The distribution peaks at $z\sim0.4$ and has support till a redshift of about 4.

The Agora galaxy density fields are computed on an 8192 `NSIDE` grid. Such fine resolution is unnecessary for our purpose, and we downgrade the density field to a `NSIDE`=1024 to speed up calculations, using `healpy`'s `pixelfunc.ud_grade` method. We have checked that the angular

[6]Note that here, galaxies always refer to the so-called *clustering* galaxies in the LSST survey, and NOT the *background* galaxies.

[7]see Omori (2022) for the values of the linear bias parameters and the analytic expressions for the redshift distributions.

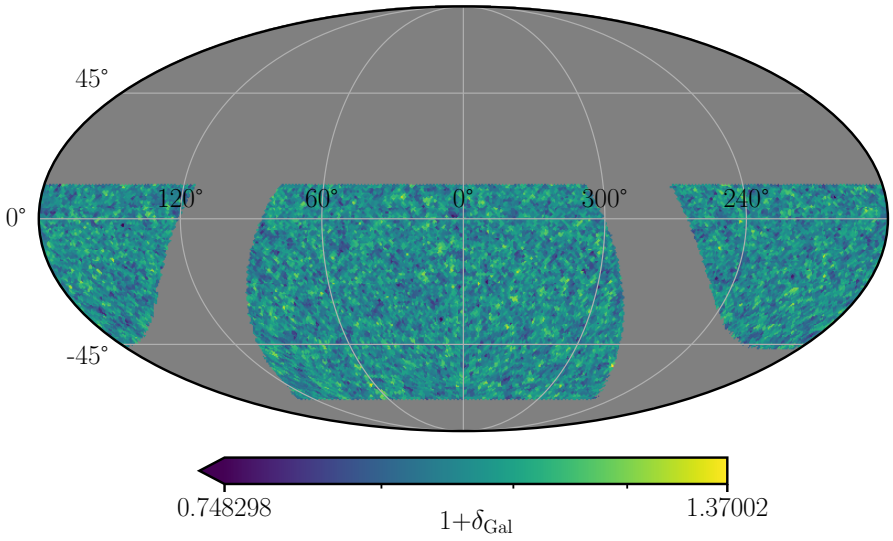


Figure 4.2: Mollweide projection of the combined LSST Y1 overdensity field simulated using the Agora lightcone. Warmer colours represent a higher density of galaxies, while cooler colours represent underdensities, and the colour bar is in logarithmic scale to enhance contrast. The empty grey regions represent the portion of the sky expected to be outside the LSST survey footprint (see text and references for the procedure to create the mask). Note that the simulations produce an all-sky density map; we artificially remove the data outside the mask to mimic the expected observations. The HEALPix `NSIDE` for this map is 64.

power spectrum of the galaxy density field is preserved in the downgrading process.

The Agora simulations create an all-sky galaxy density field. However, the survey footprint of LSST does not cover the entire sky. To make the study more realistic, we restrict the simulated data to the expected survey footprint of $-70^\circ < \text{Dec} < 12.5^\circ$, following the DESC Recommendations for optimizing the LSST Observing Strategy (Lochner et al., 2018). Since data close to the plane of the Milky Way is also expected to be contaminated, we further remove data with absolute galactic latitude less than 15°. This process results in a mask with $\sim$43.3% sky coverage corresponding to an area of $\sim 1.78 \times 10^4$ sq. deg. Figure 4.2 displays the simulated LSST Y1 overdensity fields, further downgraded to `NSIDE` = 64 to enhance the density contrast.

4.1.2 Mock BBH Catalogues

In this section, we describe the BBH catalogues used in this forecast study. We create a set of BBHs that are clustered and spatially correlated with the simulated LSST Y1 galaxy density field,

and 100 realisations of spatially unclustered and randomly distributed BBHs that are otherwise statistically identical to the clustered BBH catalogue. The clustered set serves as a proxy for future 'data' while the unclustered realisations serve as the 'control' data used to account for selection effects and quantify the cosmic variance in the cross-correlation measurements.

We begin with the description of the unclustered catalogue. The procedure to create the injected population of the BBHs is very similar to the one followed in section 3.1.2 to create the mock BBH used in the analysis of the currently available data.

First, we distribute a population of BBH merger events isotropically in the sky by sampling their locations from a uniform distribution. Next, we draw their component masses assuming a `Power Law + Peak` model for the primary mass and a power law distribution for the mass ratio. The mass model is described in detail in appendix B.

As discussed in section 4.1.1, the redshift distribution of the LSST Y1 galaxy sample extends up to $z\sim4$. Therefore, we must incorporate redshifts beyond $z\sim1$ in the injected BBH population. In section 3.1.2, we sampled the redshifts of the BBHs from a power law redshift evolution model with the power law index set at the value inferred in the LVK population analyses using the presently observed BBHs. However, since the LVK collaboration has only observed BBHs till a redshift of ~1, this model is currently constrained only at small redshifts, and we need to be extremely careful while extrapolating it to higher redshifts. In most formation channels, black holes have a stellar origin. Consequently, the formation rate for BBHs is expected to be related to the rate at which stars are formed in the universe (see, e.g., Fishbach & Kalogera, 2021, and references therein). A recent study by Vijaykumar et al. (2023a) indicates that current observations also favour a BBH merger rate that follows the star formation rate. Since the star formation rate peaks at $z\sim2$ and falls off subsequently (Madau & Dickinson, 2014), there is no physical motivation to assume that the BBH merger rate should keep increasing with redshift as a power law at higher redshifts.

For the purposes of this forecast study, following Iacovelli et al. (2022), assume that the redshift evolution of the BBH merger rate per unit comoving volume per unit source-frame time follows the Madau-Dickinson profile (Madau & Dickinson, 2014) for the star formation rate density. As a caveat, we emphasize that this assumption is only a first approximation which serves as a starting point in the absence of observation constraints; there are physical effects, such as time delays between the formation and merger of the binaries (Fishbach & Kalogera, 2021), and metallicity-dependent effects (Santoliquido et al., 2021; Chruślińska, 2024), that could result in the redshift

evolution of the BBH merger rate deviating from the evolution of the star formation rate. As our understanding of the redshift evolution of compact object merger rates improves with upcoming gravitational wave observing runs, future studies will have to be updated to account for these effects.

The mathematical details of our redshift model are discussed in appendix B. We sample from this distribution out to a redshift of 4. This high-redshift cutoff for the injected population is chosen to match the redshift distribution of the galaxy sample; any BBHs detected at redshifts that have very few or no galaxies would not contribute to the cross-correlation signal between the BBHs and the galaxy density field. Finally, following Iacovelli et al. (2022), we assume that the BBH merger rate per unit comoving volume per unit source-frame time at $z = 0$ is equal to 17 $\mathrm{Gpc}^{-3}\mathrm{Yr}^{-1}$, as inferred in LIGO population analyses using the GWTC-3 catalogue.

Once we have assigned the masses and redshifts to the injected BBHs, we draw their inclination angles, polarisation angles and phases from uniform distributions over their allowed physical ranges, similar to section 3.1.2. We sample the BBH coalescence times uniformly over the observing period of 10 years. The black hole spins are again set identically to zero. As discussed in section 3.1.2, we do not expect the spins to affect the clustering properties or the sky localisation uncertainties, which are most relevant for our spatial cross-correlation study.

After creating the mock BBH population, we determine which events can be 'detected' by a gravitational wave detector network consisting of 5 ground-based detectors, LIGO Hanford, LIGO Livingston, Virgo, LIGO India and KAGRA (henceforth, HLVIK). Following Calore et al. (2020), we assume that each detector has a duty cycle of 80%, i.e., is in science mode for 80% of the observing period. Following the procedure outlined in section 3.1.2, we first simulate the gravitational wave signals for each BBH using the `IMRPhenomXPHM` model. Next, we inject the simulated signals in stationary Gaussian noise created using analytic estimates for the power spectral densities for detectors provided by the `PyCBC` package, assuming the advanced LIGO A+ sensitivities for the 3 LIGO detectors and design sensitivities for Virgo and KAGRA. Finally, we compute the network matched-filtered signal-to-noise for each event and classify the events with an SNR ≥ 10 as 'detections'. This results in $\sim 2.8 \times 10^4$ detections on the entire sky, of which $\sim 1.1 \times 10^4$ lie inside the expected LSST Y1 survey footprint. Hence, *an HLVIK detector network observing at an A+ sensitivity and an 80% duty cycle for a period of 10 years is expected to detect* $\sim 2.8 \times 10^4$ *merging binary black holes per year*. Our findings are consistent with other forecasts in the literature. For example, with the same detector network, duty cycle and observing period, Calore et al.

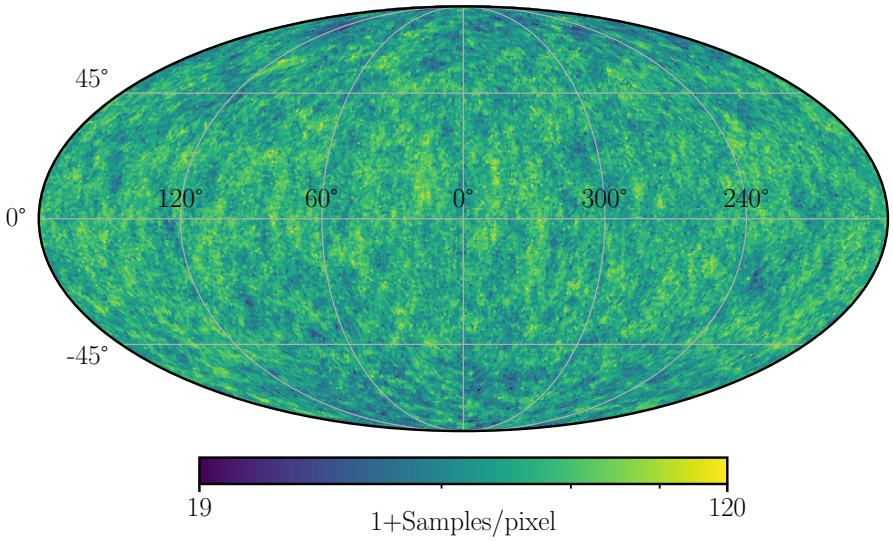


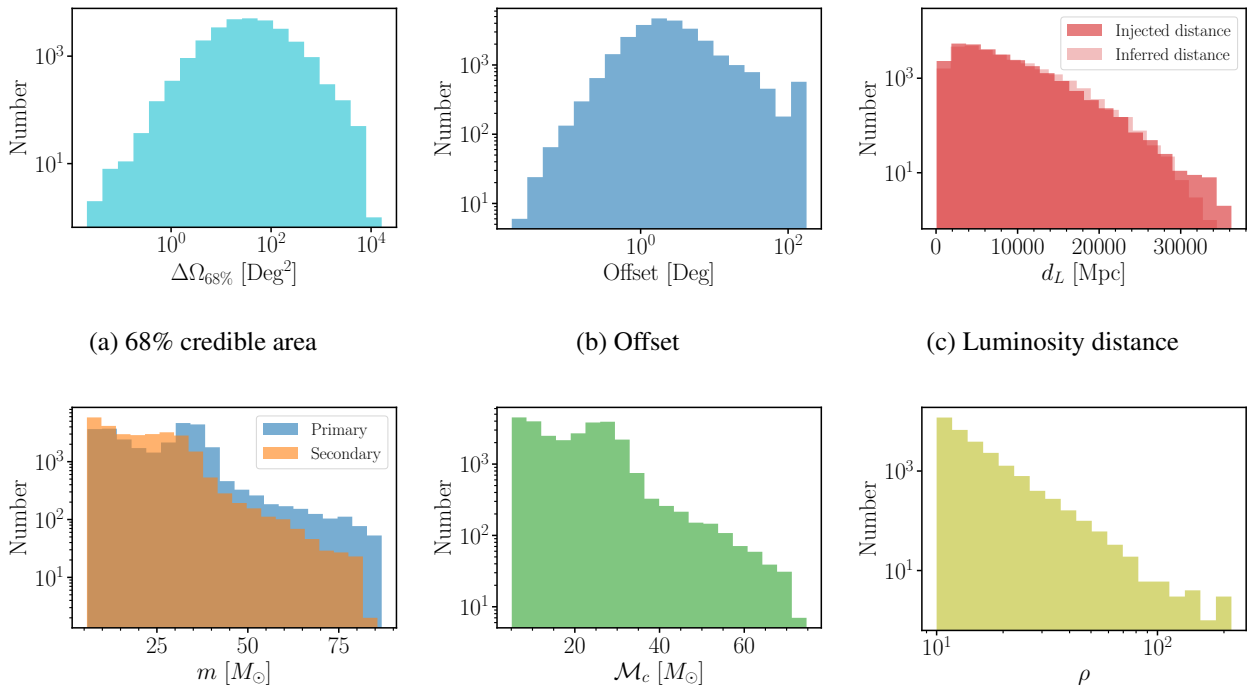
Figure 4.3: Mollweide projection of the combined skymap of the $\sim 2.8 \times 10^4$ BBHs constituting one realisation of the unclustered mock BBH catalogue. The colour represents the number of posterior samples per pixel in a logarithmic scale. The HEALPix `NSIDE` for this map is 64. This plot represents what the BBH skymap is expected to look like after 10 years of gravitational wave observations with an HLVIK detector network operating at A+ sensitivity.

(2020) find that $\sim 2 \times 10^4$ BBHs are detected. Their numbers are slightly different because they use design sensitivities for the LIGO detectors instead of A+ sensitivities, and an SNR cutoff of 8 instead of 10. We have explicitly checked that with their analysis choices, we reproduce their numbers.

Once we have the selected events, we use `BAYESTAR` to localise them. `BAYESTAR` also returns the estimated luminosity distances, the offsets between the injected and highest probability sky locations, and credible intervals for the area of sky localisation uncertainty for each detected event. Figure 4.3 plots the combined skymap of the detected BBHs and represents the expected BBH skymap from 10 years of data taken with an HLVIK detector network operating at A+ sensitivity.

Figure 4.4 shows the distribution of various BBH properties of the unclustered BBH catalogue. It is evident from a comparison of figures 3.4a that the addition of LIGO India and KAGRA, combined with better sensitivities of each detector, results in significant improvements in sky localisation of the detected BBHs.

So far, we have created one realisation of the unclustered mock BBH catalogue. In principle, one can generate the full catalogue by repeating the procedure outlined above 100 times, with



(a) 68% credible area (b) Offset (c) Luminosity distance

(d) Component mass (e) Chirp mass (f) SNR

Figure 4.4: The top panel shows the distribution of the BBH properties most relevant for clustering, namely the 1σ sky localisation uncertainty areas (*left*), angular offset between injected and inferred sky locations (middle), and luminosity distances (*right*) of the unclustered mock BBH catalogue. The light (bold) histogram in the right subplot represents the distribution of the inferred (injected) distances. The bottom panel shows the distribution of the component masses (*left*), chirp masses (*middle*) and SNRs (*right*) of the mock catalogue. In the left panel, the primary (heavier) BBH mass distribution is shown in blue while the secondary (lighter) BBH mass distribution is shown in orange.

a different random seed chosen each time for creating the injected BBH population. However, generating the skymaps for $\sim 2.8 \times 10^4$ events is computationally expensive. Furthermore, it demands a significant storage requirement ($\sim$30 GB for one realisation). Therefore, generating 100 realisations from scratch is practically infeasible[8]. Fortunately, there exists a symmetry in the sky distribution of the BBHs that allows us to generate multiple realisations of the unclustered catalogue from just one realisation.

As discussed before, gravitational wave detectors are most sensitive to sky locations directly above the plane of their arms, which results in the sensitivity of the detector network to a gravitational wave signal being a function of the declination and right ascension of its source. Due to the earth's rotation about its axis, this selection function peaks at different right ascensions in the sky at different times during the day, but at the same declination. This is because the detectors move along lines of constant latitude with the rotation of the earth. As a result, the dependence of the selection function on right ascension gets averaged out over the large number of complete rotations in the observing run, and the final selection function is expected to be only a function of the declination of the sources[9]. To verify this, we plot the number counts, sky localisation uncertainty, and offsets of the detected BBHs against their right ascensions, and show the results in figure 4.5. As expected, we find no correlation between these properties and the right ascension. Hence, it is safe to assume that the BBH properties most important for clustering are independent of the right ascension of the BBHs.

We utilise this symmetry to generate 100 realisations of the unclustered catalogue from the BBH sample created originally. The BBHs in each new realisation are created by randomly displacing the posterior distributions of the original BBHs in the sky along constant declination latitudes. In practice, this is achieved by adding a random number between 0 and 2π to the right ascensions of the `BAYESTAR` posterior samples[10].

Now that we have the control set of unclustered mock BBHs, we describe the methodology to create the clustered catalogue that will serve as the proxy for future BBH events. There is only

[8] If we are to finish the Master's Thesis in time.

[9] This is not strictly true for observations, since the overall sensitivity of the detectors is also a function of the time of the day. For example, detectors are more sensitive at times when human activity is minimal. This can lead to residual selection effects on the right ascension. However, this is a higher-order effect that can be ignored for the purposes of a study with simulated data.

[10] Note that the same random number is added to all samples from a single BBH to preserve the shape of the distribution, but different random numbers are added to different BBHs. Moreover, each BBH is shifted by different random numbers to create different realisations. For example, to create 100 realisations of 28,000 BBHs, we need 2,800,000 random numbers.

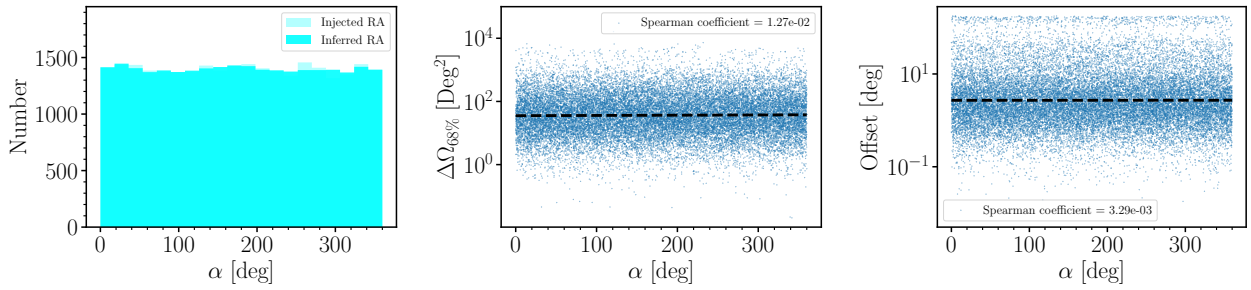


Figure 4.5: *Left:* the distribution of the right ascension of the unclustered mock BBHs, with the light histogram representing the inferred values and the bold histogram representing the injected values. *Middle:* scatter plot showing the variation of the 1σ sky localisation uncertainty area with right ascension. The dashed curve shows the best-fit straight line through the data. *Right:* same as the middle panel, but for the offset between injected and inferred sky locations. These plots demonstrate that the BBH properties most important for clustering, namely number counts, sky localisation uncertainty, and offsets, are uncorrelated with the right ascension of the BBHs. Therefore, the procedure for creating the other realisations of the unclustered catalogue by randomly shifting the posteriors of each event along constant declination latitudes is justified.

step we need to do differently from the procedure for creating the control data: since we need the injected population of BBHs to be inherently correlated with the simulated LSST Y1 map, we cannot distribute their positions isotropically in the sky. In this study, we implement the following steps to assign positions to the injected BBH population that are correlated with the galaxy density field:

1. First, we create a set of points that represent a local Poisson process on the simulated LSST Y1 galaxy density field $\delta_{\rm Gal}$. In practice, this is done by assuming that the probability for a BBH to merge at a location (δ, α) is proportional to $1 + \delta_{\rm Gal}(\delta, \alpha)$. We use the `healpytools.rand_pix_from_map` method from the `python` library `astrotools`[11] for this purpose, which implements inverse CDF sampling to draw random pixels from a HEALPix map[12]. For each randomly sampled pixel, we assign the centre of the pixel as a temporary BBH location.

2. Step (i) creates a set of tracers that sample the galaxy density field in an unbiased manner,

[11] https://astro.pages.rwth-aachen.de/astrotools/index.html

[12] The choice of `NSIDE` = 1024 for the LSST field created in section 4.1.1 results in a small dynamic range for $1 + \delta_{\rm Gal}$, leading to the CDF being saturated quickly, which causes systematic effects at large angular scales in the distribution of the samples created using an inverse CDF sampling procedure. To mitigate this, we downsample the field to `NSIDE` = 512 before drawing the samples.

meaning that the bias for the BBH population and the galaxy sample with respect to the underlying matter field are identical. This is an undesirable situation since there is no physical reason to assume that the bias of these two samples should be the same. To correct this, we randomly scatter the temporary BBH positions on scales of $\sim 0.15^{\circ}$[13], which correspond to transverse spatial scales of ~ 4 Mpc for a median redshift of ~ 0.4 for the LSST galaxies. This scale represents an estimate for the typical displacement of a BBH system from the centre of its host galaxy, which is expected to be smaller than the size of typical dark matter halos[14]. This step serves another purpose: it ensures that the injected BBH positions no longer lie on a grid, unlike the temporary postitions that were taken to be centres of HEALPixels.

3. Figure 4.6 indicates that there is a $\sim 80\%$ overlap between the redshift distributions of the unclustered BBHs and the LSST Y1 galaxies. This overlap is expected to be similar for the clustered mock BBHs. Therefore, if we draw all injections from the LSST Y1 density field, $\sim 20\%$ of the clustered mock BBHs that will have inferred redshifts outside the overlap region will be spuriously correlated with the galaxy field, leading to artificial enhancement in the clustering measurement. To mitigate this potential bias, we replace 20% of the injections by random draws from a Poisson distribution in the sky.

Once we have the positions, the rest of the procedure for creating the spatially clustered catalogue is identical to the one followed for creating the unclustered one. We have checked that the two catalogues are identical in all respects other than their clustering properties, but omit the analogue of figure 4.4 for the clustered catalogue for brevity.

4.2 Cross-correlation Analysis

In this section, we present a preliminary analysis of the spatial cross-correlation between the mock HLVIK BBH catalogues and the simulated LSST Y1 galaxy density field created in the previous section.

[13] In practice, this corresponds to shifting each BBH by a random angular distance drawn from a Gaussian of standard deviation 0.15°, along a randomly chosen great circle intersecting its temporary location.

[14] Note that dark matter halos are typically 1-2 Mpc in size, and we take a more conservative value to account for low redshift events, for which the projected angular size would be larger.

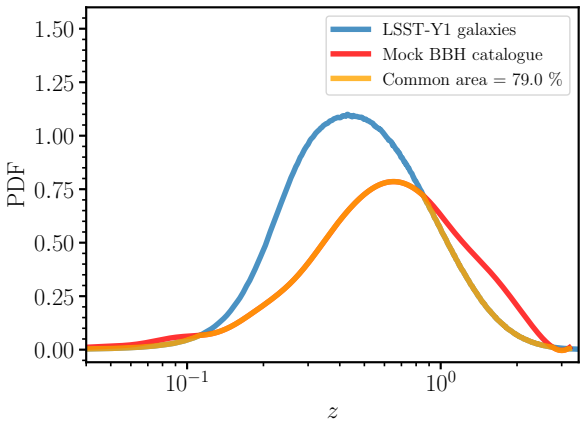


Figure 4.6: A comparison of the distribution of inferred redshifts for the unclustered mock BBHs (red curve) and the modelled $n(z)$ for LSST Y1 galaxies (blue curve). There is a $\sim 80\%$ overlap between the two redshift distributions, as shown by the orange curve.

In chapter 3, we used the angular power spectrum as the two-point statistic of choice. Here, we focus on the two-point cross-correlation function instead[15]. As before, we also compute the excess cross-correlation as measured by the first and second nearest neighbour distributions.

4.2.1 Robustness of cross-clustering statistics

Since we are conducting a forecast study with simulated data, we have the 'true' (injected) sky locations of the 'detected' BBHs in addition to their 'observed' (inferred) posterior probability distributions in the sky. Therefore, we can perform two measurements:

1. cross-correlation between the observed (inferred) sky distributions of the mock BBHs and the galaxy density field

2. cross-correlation between the true (injected) sky locations of the mock BBHs and the galaxy density field

The first measurement provides a prediction of the cross-correlation signal expected from realistic

[15]There is no objective reasoning for this decision since the two-point function and the power spectrum are completely equivalent measures of clustering. We choose to work with the correlation function simply because we have already explored the power spectrum in the previous chapter.

future data and the second measurement estimates the signal expected if all detected BBHs were perfectly localised.

The additional information present in the second measurement is extremely valuable for the following reasons. As discussed in the previous chapter, the uncertainty in the sky localisation of the BBHs tends to reduce the strength of the clustering signal. By comparing the cross-correlation signal detected in the inferred sky distribution of the BBHs against that in the injected distribution (which is the true underlying signal), we can judge the robustness of the summary statistic used to perform the measurement in the presence of uncertainties in sky localisation. For example, a summary statistic that can measure the clustering signal using the injected locations but not using the inferred distributions is less robust than another summary statistic that can detect the signal using both the injected and inferred distributions.

In order to compare the measurements with each other and determine the robustness of the summary statistics to the presence of sky localisation uncertainty, we need to ensure that both are computed using exactly the same procedure. As discussed in chapter 3, performing the first measurement requires computing the summary statistics using random samples from the posterior distributions of the BBHs and averaging over many such samples (see sections 3.2.1 and 3.2.3). However, each BBH has a unique injected sky location and this sampling-averaging procedure is not applicable to the second measurement. How do we ensure consistency in the treatment of the two measurements?

For the purposes of this preliminary study, we assign each BBH the most probable position in its sky localisation area to perform the first measurement. Although not the most realistic approach for computing the clustering of objects with uncertain sky localisations[16], it is a reasonable first approximation for the BBH catalogue considered in this forecast study, which has much fewer events with bimodal sky distributions and significantly improved sky localisations than the data considered in chapter 3[17]. However, we caution the reader that the measurements presented here are not realistic predictions, since we neglect the information in the full sky distributions of the individual BBHs. We will compute the realistic measurements using the sampling-averaging procedure in future work.

[16] See the discussion in section 3.2.1.

[17] This is partly due to the addition of two more detectors (LIGO India and KAGRA) to the network and partly because of the increased sensitivities of each detector.

4.2.2 Angular scales

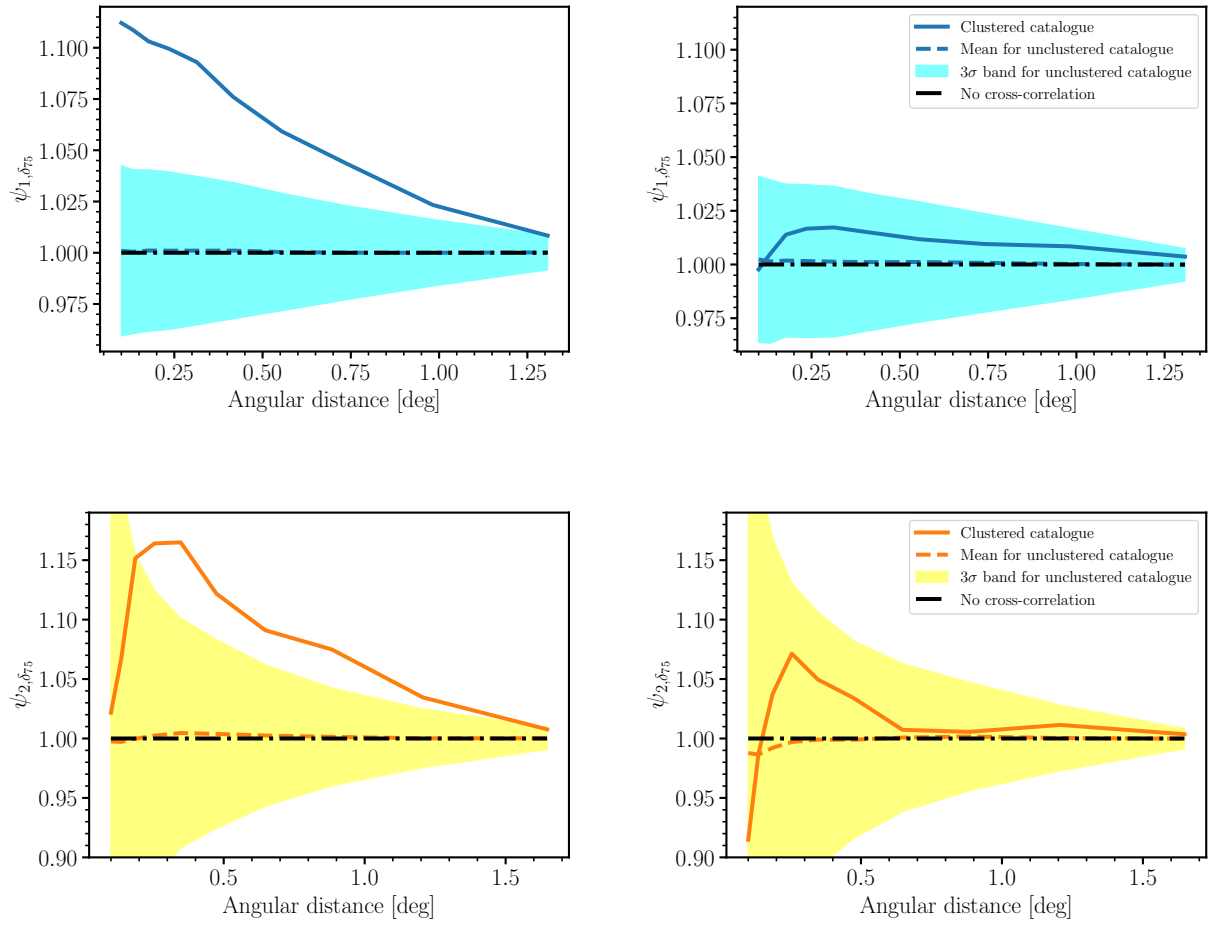
We conduct the clustering analysis on angular distance scales from $\sim 0.1^\circ$ to $\sim 3^\circ$. These angular distances correspond to projected transverse distance scales of ~ 3 to ~ 90 Mpc for a median redshift of ~ 0.4 for the LSST Y1 galaxies. The majority of these scales are in the nonlinear or quasi-nonlinear regime. For computing the overdensity fields and the query points for the nearest-neighbour measurements, we use an `NSIDE` = 1024 HEALPix grid with $\sim 1.2 \times 10^7$ pixels and an angular resolution of $\sim 0.06^\circ$. As required for the nearest-neighbour analysis, the number of query pixels is much larger than the number of data points, and the query grid has sufficient resolution to sample the smallest spatial scales analysed. We remove all query points within 2° of the WSC mask boundaries for computing nearest-neighbour excess cross-correlation to avoid any biases due to the presence of the mask.

4.2.3 Results

In this section, we present the results of our preliminary analysis[18]. We focus on evaluating the robustness of the summary statistics to the presence of offsets between the injected and inferred sky locations of the BBHs, and leave realistic clustering measurements for future work.

Figure 4.7 plots the excess cross correlation between the $\sim 1.1 \times 10^4$ BBHs of the clustered mock catalogue and the simulated LSST Y1 galaxy density field as measured by the first (top panel) and second (bottom panel) nearest neighbour distributions. In each panel, the left plot shows cross-correlation measurements using the 'true' (injected) sky location for each BBH, while the right plot shows the measurements using the highest probability (inferred) sky location for each BBH. The solid line in each subplot represents the measurement for the clustered BBH catalogue, while the dashed line and shaded band represent the mean and 3σ deviations over the 100 realisations of the unclustered BBH catalogue. The dash-dot line indicates the expected value in the absence of cross-correlations.

[18] We emphasize that these results do not represent the realistic cross-correlation signals expected in the forecast data since we do not account for the full sky distributions of the individual BBHs.



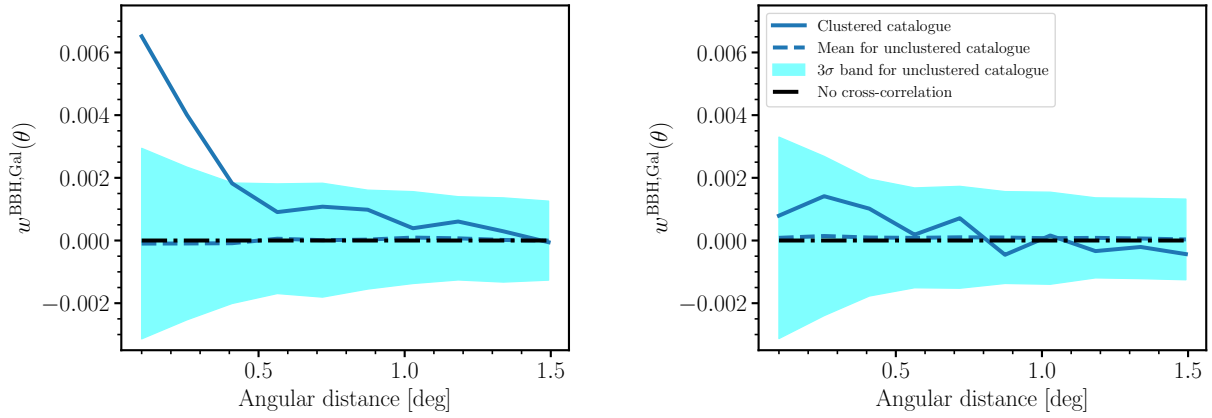
(a) Injected sky locations

(b) Inferred sky locations

Figure 4.7: Excess cross correlation between the $\sim 1.1 \times 10^4$ clustered mock BBHs and the simulated LSST Y1 galaxy density field as measured by the first (top panel) and second (bottom panel) nearest neighbour distributions. The plotting scheme for each subfigure is as follows: the solid line represents the measurement for the clustered BBH catalogue, while the dashed line and shaded band represent the mean and 3σ deviations over 100 realisations of the unclustered BBH catalogue. The dash-dot line indicates the expected value in the absence of cross-correlations. In each panel, the left plot shows cross-correlation measurements using the 'true' (injected) sky location for each BBH, while the right plot shows the measurements using the highest probability (inferred) sky location for each BBH. The y-axis scales for the left and right plots have been made identical for ease of comparison. Both nearest neighbour distributions capture a statistically significant cross-correlation signal between the distribution of the true locations of the detected BBHs and the galaxy density field but are not able to detect any cross-correlation signal using the inferred locations of the BBHs.

As demonstrated in figure 4.7a, both nearest neighbour distributions capture a large, statistically significant cross-correlation signal between the true locations of the detected BBHs and the

galaxy density field, but are not able to detect any cross-correlation signal using the inferred locations of the BBHs. Similar behaviour is observed for the two-point cross-correlation function, plotted in figure 4.8, which captures a much weaker signal than the nearest neighbour distributions using the true locations, and no signal using the inferred locations.



(a) Injected sky locations

(b) Inferred sky locations

Figure 4.8: The two-point cross-correlation function of the $\sim 1.1 \times 10^4$ clustered mock BBHs and the simulated LSST Y1 galaxy density field. The plotting scheme is similar to figure 4.7. The y-axis scales for the left and right plots have been made identical for ease of comparison. The correlation function is able to detect a cross-correlation between the distribution of the true locations of the detected BBHs and the galaxy density field, but the signal is much weaker as compared to the nearest neighbour distributions (cf. figure 4.7a). Similar to the nearest neighbour distributions, the two-point function can not detect any cross-correlation signal using the inferred locations of the BBHs.

These results imply the following: *the expected sample size of BBHs detected in 10 observing years of an HLVIK-like detector network is large enough to allow for the detection of spatial cross-correlations with an LSST-like galaxy survey, but the uncertainty in the sky localisation of the BBHs leads to a significant reduction in the measured signal.* As shown in figure 4.4b, there are even $10^\circ - 100^\circ$ offsets between the injected and inferred locations for a significant fraction of the BBHs. Given that we are performing the measurement on $\sim 1^\circ$ scales, such large offsets are likely to wash out any clustering signal.

Figure 4.9 shows that the offset between the injected and inferred sky locations of the BBHs is highly correlated with the 1σ sky localisation area. As expected, on average, better-localised events are also localised closer to their 'true' locations in the sky.

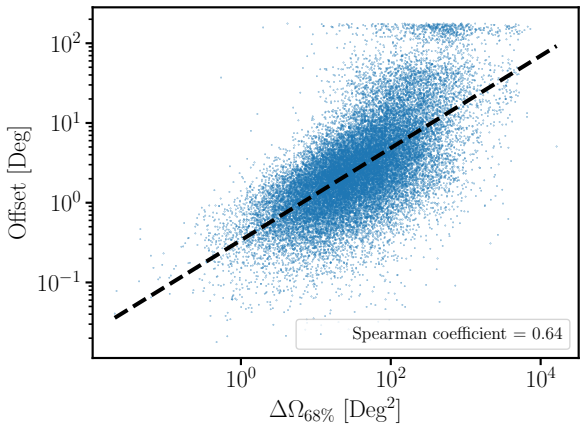
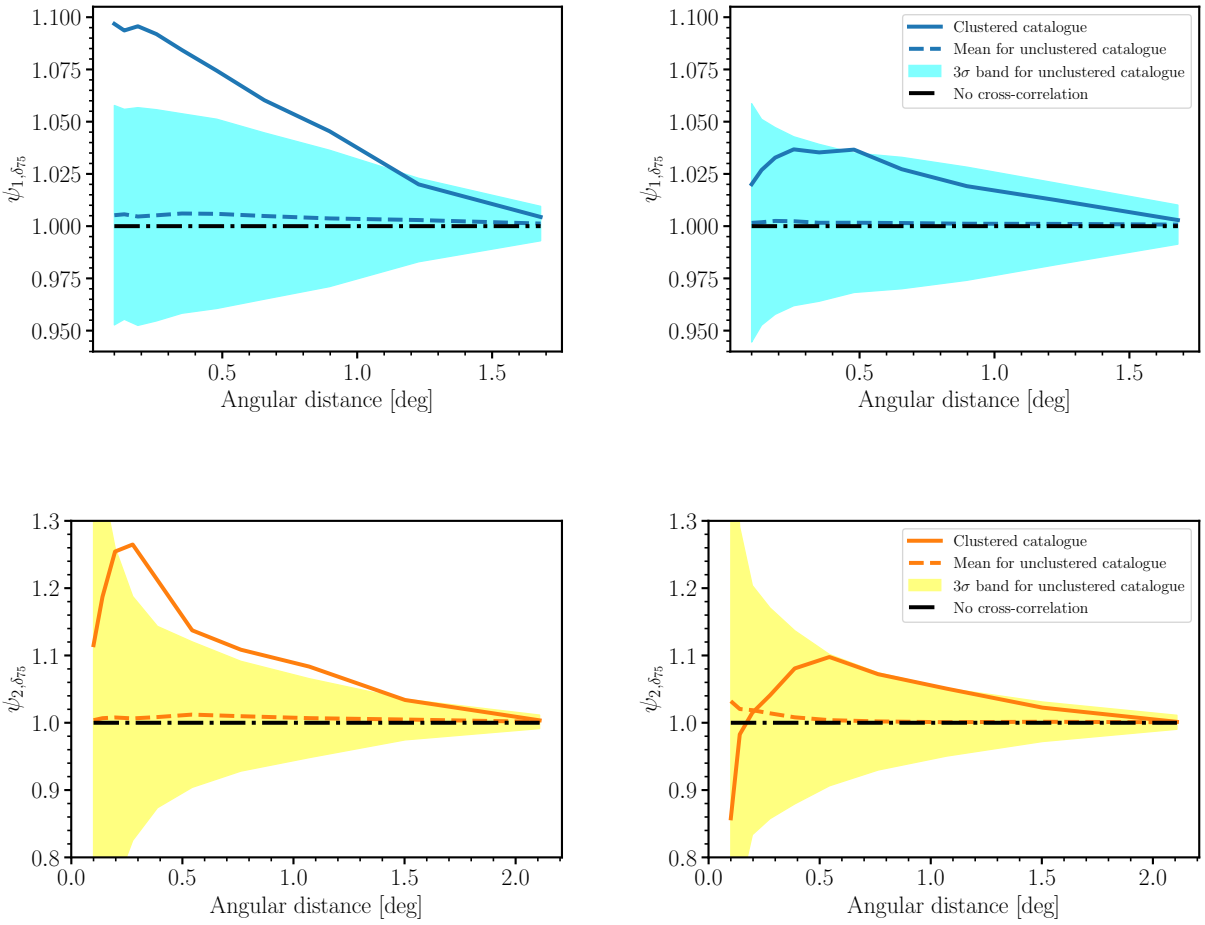


Figure 4.9: Scatter plot showing the variation of the offset between injected and inferred sky locations of the BBHs with their 1σ sky localisation uncertainty area. The dashed curve shows the best-fit straight line through the data. As expected, there is a clear correlation between the sky localisation uncertainty and the offsets, and on average, better-localised events are localised closer to their 'true' locations in the sky.

Since the number of BBHs in the mock catalogue is more than enough to detect a clustering signal using the true locations, one way of reducing the impact of the sky localisation uncertainty might be to restrict the sample to well-localised BBHs that have smaller offsets. To test this hypothesis, we remove all events with 1σ sky localisation area greater than 50 sq. deg. from the clustered and unclustered mock BBH catalogues[19]. After performing this cut, we are left with $\sim 1.6 \times 10^4$ BBHs in total and $\sim 6.7 \times 10^3$ BBHs inside the simulated LSST survey footprint, for which we compute the clustering signal. The results for the nearest neighbour measurements and the two-point correlation function are shown in figures 4.10 and figure 4.11 respectively.

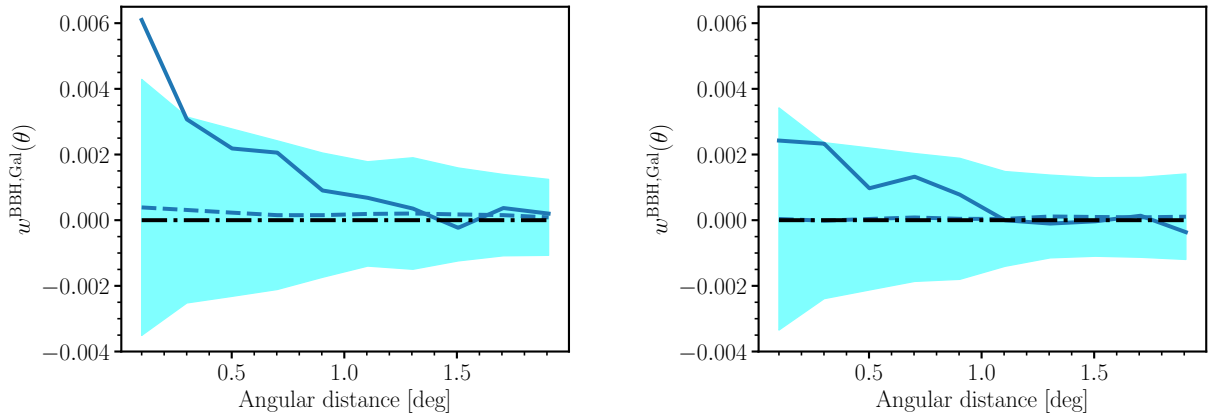
[19]Note that we cannot place a cut on offsets since they are not observables, unlike the sky localisation areas, which can be measured in data.



(a) Injected sky locations

(b) Inferred sky locations

Figure 4.10: Same as figure 4.7, but only using the $\sim 6.7 \times 10^3$ BBHs with 1σ sky localisation areas less than 50 sq. deg. The y-axis scales for the left and right plots have been made identical for ease of comparison. Both nearest neighbour distributions continue to capture a statistically significant cross-correlation signal between the distribution of the true locations of the restricted BBH sample and the galaxy density field as before. The loss in signal in going from injected to inferred sky locations is smaller for the restricted BBH sample than for the entire sample, resulting in an almost statistically significant detection in the second nearest neighbour distribution.



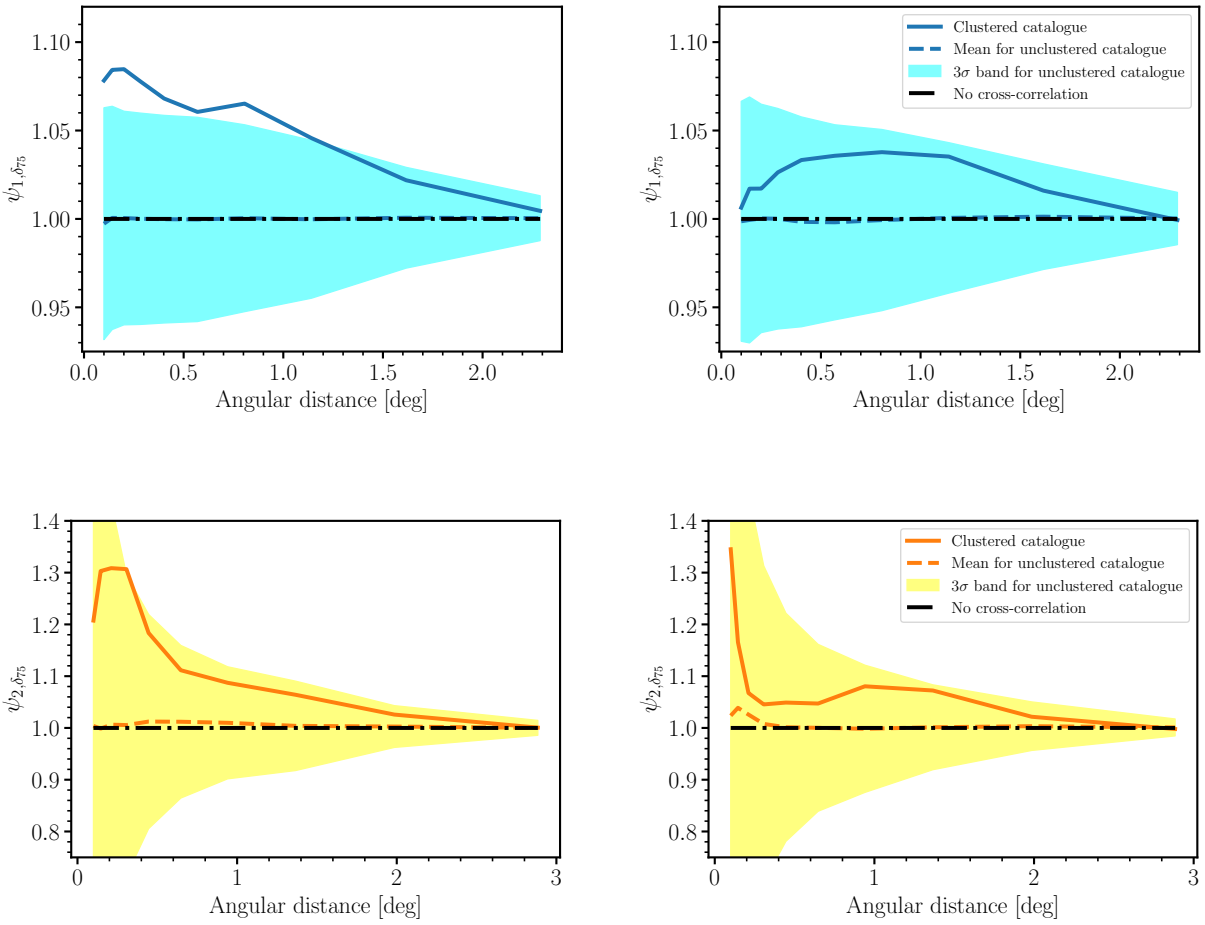
(a) Injected sky locations

(b) Inferred sky locations

Figure 4.11: Same as figure 4.8, but only using the $\sim 6.7 \times 10^3$ BBHs with 1σ sky localisation areas less than 50 sq. deg. The y-axis scales for the left and right plots have been made identical for ease of comparison. For this restricted BBHs sample, the correlation function detects a much weaker cross-correlation signal between the distribution of the true locations of the detected BBHs and the galaxy density field than the nearest neighbour distributions (cf. figure 4.10), and the cross-correlation signal is completely lost when the inferred locations of the BBHs are used instead.

Even after removing $\sim 40\%$ events with sky localisation uncertainty above the chosen threshold, both nearest neighbour distributions continue to capture a statistically significant cross-correlation signal between the distribution of the true locations of the detected BBHs and the galaxy density field. Due to the offsets for the set of better-localised BBHs being smaller, the loss in the cross-correlation signal is significantly less pronounced in both distributions as compared to the loss in signal for the entire BBH sample (cf. figure 4.7), and the second nearest neighbour distribution almost captures a statistically significant signal even using the inferred sky locations at scales larger than $\sim 0.5^\circ$. Note how the small scales are strongly affected by the offsets, and the loss in signal reduces as we go to larger scales. The two-point correlation function, on the other hand, detects a much weaker signal even for the true sky locations and is not able to detect any signal for the inferred BBH locations.

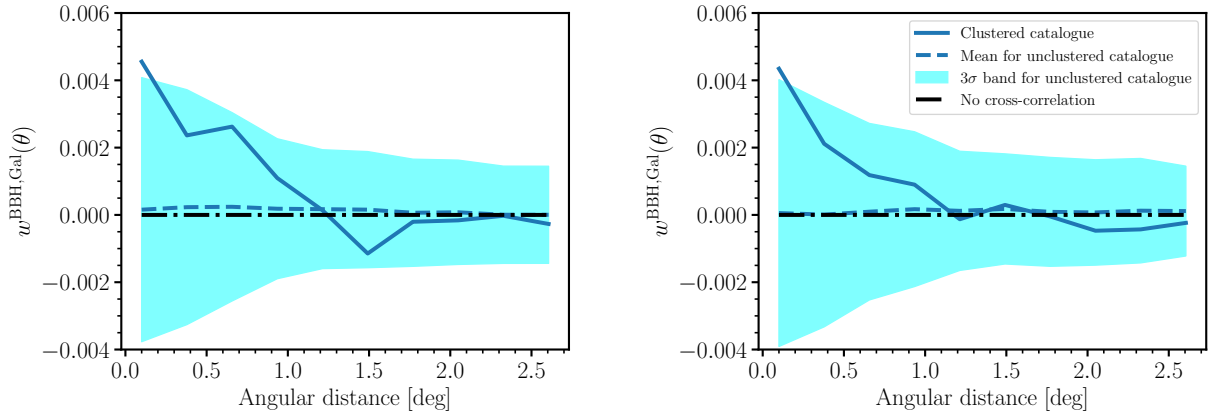
As a further test, we also repeat the analysis with all events localised to 1σ areas more than 20 sq. deg. removed from the BBH catalogues. This leaves us with $\sim 9 \times 10^3$ events in total and $\sim 3.6 \times 10^3$ events in the LSST footprint. The results are presented in figure 4.12 and 4.13.



(a) Injected sky locations

(b) Inferred sky locations

Figure 4.12: Same as figure 4.7, but only using the $\sim 3.6 \times 10^3$ BBHs with 1σ sky localisation areas less than 20 sq. deg. The y-axis scales for the left and right plots have been made identical for ease of comparison. For the BBHs that satisfy this (more stringent) selection criteria, the cross-correlation signal between the distribution of the true locations of the detected BBHs and the galaxy density field is detected only in the first nearest neighbour distribution, and no signal is detected using the inferred sky locations in either of the two distributions.



(a) Injected sky locations

(b) Inferred sky locations

Figure 4.13: Same as figure 4.8, but only using the $\sim 3.6 \times 10^3$ BBHs with 1σ sky localisation areas less than 20 sq. deg. The y-axis scales for the left and right plots have been made identical for ease of comparison. For the BBHs that satisfy this (more stringent) selection criteria, the two-point function does not capture a correlation signal using either the injected or the inferred sky locations.

We observe that with this more stringent cut on sky localisation, the number density of BBHs becomes so small that even for the injected locations, the signal is detected only in the first nearest neighbour distribution, and none of the summary statistics are able to measure any cross-correlations using the inferred locations.

4.2.4 Discussion

The main takeaways of our analysis are as follows:

- Two opposing factors determine whether a BBH-galaxy cross-correlation signal can be detected: the sample size of the BBHs and the offsets between true and inferred sky locations.
- If no restrictions are placed in the sky localisation uncertainty of the events, the sample size of the BBHs is large enough to detect a significant clustering signal using the true sky locations, but the resulting large offsets lead to a non-detection using the inferred sky locations.
- On the other hand, the gains from smaller offsets achieved by placing too aggressive sky

localisation cuts become redundant since the remaining sample of BBHs becomes too small for a clustering signal to be detected even in the true locations.

- A 50 sq. deg. selection criteria leads to an optimal combination of sample size and offsets: there are enough BBHs left in the sample to get a significant clustering measurement for the true locations, and the offsets are small enough to not smear out the observed signal at the angular scales of interest.

- The number density of BBHs with $\Delta\Omega_{68\%} < 50$ sq. deg. expected from 10 years of observations by an HLVIK-like detector network is sufficient to allow for a marginal detection of their spatial cross-correlations with an LSST-like galaxy survey in the second nearest neighbour distribution, but not large enough to allow for a detection in the two-point correlation function or the first nearest neighbour distribution.

- The first nearest neighbour distribution captures a larger cross-correlation signal for perfectly localised BBHs than the second nearest neighbour distribution, but the second nearest neighbour distribution is more robust to uncertainty in sky localisation than the first nearest neighbour.

- The nearest neighbour distributions are more robust to sky localisation uncertainty than the two-point correlation function.

Currently, we do not understand why the second nearest neighbour distribution seems to be more robust to offsets than the first nearest neighbour distribution. We leave a systematic investigation of this behaviour to future work.

Chapter 5

Conclusion and Outlook

In this thesis, we developed a framework for quantifying the spatial clustering of sources of gravitational waves and their cross-correlation with the large-scale structure of the universe, using two-point summary statistics and nearest-neighbour distributions as summary statistics. We extended the k-nearest-neighbour formalism, originally developed in Banerjee & Abel (2021a) and Banerjee & Abel (2023) for 3D clustering, to angular clustering in the sky. Our framework implements robust strategies to deal with the extended sky localisation of sources and selection biases associated with gravitational wave detections. It can handle observational systematics due to the presence of masked regions in the sky with unreliable electromagnetic observations.

We illustrated the statistical power of the nearest-neighbour distributions as measures of spatial clustering of sparsely sampled and highly biased tracers by cross-correlating the overdensity field of the WISE$\times$SuperCOSMOS (WSC) all-sky catalogue with 36 tracers residing in the highest density regions in the sky. Even with such a small sample size, the first nearest-neighbour distribution captured a statically significant signal at small scales where the angular power spectrum did not. Through this example, we demonstrated that the nearest-neighbour distributions are able to access information in the higher-order correlation functions at small scales where cosmological fluctuations are non-Gaussian.

As a first application to data, we measured the angular power spectrum and nearest-neighbour distributions of the Binary Black Hole (BBH) mergers detected in the first three observation runs of LIGO-Virgo-KAGRA and cross-correlated these sources with galaxies and quasars from the WSC catalogue. We adopted a hypothesis-testing approach to determine the significance of the

clustering signal, with the null hypothesis stipulating that BBHs are distributed uniformly in the sky. To mitigate observational biases in the BBH data, we created a catalogue of mock BBHs that statistically reproduce the observed properties of the detected BBHs but are spatially unclustered and uncorrelated with the large-scale structure of the universe. This sample served as a natural control set to compare with the data while testing the null hypothesis.

Using chi-squared distributions to measure statistical deviations from the null hypothesis, we found no evidence for spatial clustering of BBHs or their cross-correlation with large-scale structure in the presently available data. These results are consistent with similar studies in the literature (Zheng et al., 2023; Mukherjee et al., 2022; Cavaglià & Modi, 2020). We discussed that an absence of a clustering signal is not unexpected, given the small sample size and large uncertainty in the sky localisation of the BBHs.

A detection of clustering with so few events would indicate that BBHs reside in extremely biased environments in the universe, such as cosmic web nodes and massive voids. However, a non-detection of this cross-correlation in currently available data does not rule out this scenario, since the sky localisation uncertainty smears out the clustering signal at small scales where the measurements are most sensitive. We demonstrated that with well-localised BBHs, the kNN tracer-field formalism has the exciting potential to test the possibility of BBHs being highly biased tracers of large-scale structures.

Although we were not able to measure the clustering signal in the presently available data on binary black hole mergers, our framework provides a powerful means to study spatial cross-correlations between continuous fields and transient events with uncertain sky localisation in the presence of selection effects and observational systematics, as demonstrated by the results of our forecast study for future observing runs of LIGO and stage-IV galaxy surveys.

We forecast 10 years of GW observations with a network of 5 ground-based detectors, resulting in a mock BBH catalogue of $\sim 2.8 \times 10^4$ BBHs, of which $\sim 1.6 \times 10^4$ were localised to better than 50 sq. deg. and $\sim 9 \times 10^3$ to better than 20 sq. deg. in the sky. We cross-correlated 3 sets of BBHs (the full sample and the two sets having maximum allowed sky localisation uncertainties of 50 and 20 sq. deg. respectively) with the simulated galaxy overdensity field of an LSST Y1-like survey and found that the second nearest neighbour distribution captures a nearly statistically significant cross-correlation signal for the modestly-localised $\sim 1.6 \times 10^4$ BBHs with sky localisation area smaller than 50 sq. deg., while the two-point cross-correlation function does not capture a clear signal. Our analysis demonstrates that the nearest neighbour distributions can extract higher-order,

non-Gaussian clustering from the small spatial scales.

None of the clustering statistics detected a signal using the inferred locations of the other two samples. This can be understood as follows: if no restrictions are placed in the sky localisation uncertainty of the events, the sample size of the BBHs is large enough to detect a significant clustering sample using the true sky locations, but the resulting large offsets lead to a non-detection using the inferred sky locations. On the other hand, placing an aggressive sky localisation cut (e.g., 20 sq. deg.) is not effective either since the remaining sample of BBHs becomes too small for a clustering signal to be detected even in the true locations. Our findings suggest that a moderate selection criterion (such as sky localisation area less than 50 sq. deg.) leads to an optimal combination of sample size and offsets that ensures a large enough sample size to get a significant clustering measurement for the true locations while keeping the offsets are small enough to not smear out the observed signal at the angular scales of interest.

With a statistically significant population of even better-localised merger events expected to be detected in future observing runs of the third generation of gravitational wave detectors (Iacovelli et al., 2022; Borhanian & Sathyaprakash, 2022; Hall & Evans, 2019), we would have access to even smaller scales where the nearest-neighbour distributions are expected to offer significant gains over a two-point analysis (Banerjee & Abel, 2023, 2021b,a). Hence, the techniques developed in this thesis would be crucial for measuring the clustering of gravitational wave sources that will be detected in the coming decades. In future work, we will conduct forecast studies analysing the angular clustering of BBHs expected to be detected by the third-generation detectors, as well as their cross-correlations with forecast galaxy data for stage-IV large-scale surveys (Gupta & Banerjee 2024 in prep.).

We focused on binary black holes in this work, but our framework can also be applied to study binary neutron star mergers or neutron star black hole mergers with minor modifications. In addition to gravitational wave sources, other astrophysical transients such as gamma-ray bursts are also often poorly localised (see Michael Burgess, J. et al. (2021) and references therein). The methods presented in this thesis will be useful for conducting multi-messenger studies with these objects. Similarly, the tracer-field correlation formalism discussed here can be applied to conduct cross-correlation studies between large-scale structure and cosmological fields, such as the cosmic microwave background and the cosmological 21 cm neutral hydrogen signal.

Programming Software and Data Availability

The `iPython` notebook environment `JupyterLab` [1] and the `python` libraries `healpy`[2], `NumPy`[3], `pandas`[4], `scikit-learn`[5] and `SciPy`[6] were used extensively in this work. All plots in this thesis were made using `Matplotlib`[7].

The following publicly available data were used in this thesis: the gravitational wave parameter estimation data and skymaps for the parent BBH catalogue, which can be found at https://zenodo.org/records/5546663, and the WSC SVM catalogue and mask, which can be found at http://ssa.roe.ac.uk/WISExSCOS.html. The simulation products of the Agora lightcone were obtained on request from Dr. Yuuki Omori, and are gratefully acknowledged. The data generated in this study, including the mock BBH catalogues, are available upon reasonable request.

[1] https://jupyterlab.readthedocs.io/en/latest/index.html
[2] https://healpy.readthedocs.io/en/latest/
[3] https://numpy.org/
[4] https://pandas.pydata.org/
[5] https://scikit-learn.org/
[6] https://scipy.org/
[7] https://matplotlib.org

Bibliography

Abbott B. P., et al., 2017, Nature, 551, 85

Abbott B. P., et al., 2021, The Astrophysical Journal, 909, 218

Abbott R., et al., 2023, The Astrophysical Journal, 949, 76

Accadia T., et al., 2012, Journal of Instrumentation, 7, P03012

Ade P. a. R., et al., 2016, Astronomy & Astrophysics, 594, A13

Adelman-McCarthy J. K., et al., 2006, The Astrophysical Journal Supplement Series, 162, 38

Adhikari S., Fishbach M., Holz D. E., Wechsler R. H., Fang Z., 2020, The Astrophysical Journal, 905, 21

Aghanim N., et al., 2020, Astronomy & Astrophysics, 641, A6

Akutsu T., et al., 2020, Progress of Theoretical and Experimental Physics, 2021, 05A103

Alfradique V., et al., 2024, Monthly Notices of the Royal Astronomical Society, 528, 3249

Amon A., et al., 2023, Monthly Notices of the Royal Astronomical Society, 518, 477

Balaudo A., Garoffolo A., Martinelli M., Mukherjee S., Silvestri A., 2023, J. Cosmology Astropart. Phys., 2023, 050

Banerjee A., Abel T., 2021a, Monthly Notices of the Royal Astronomical Society, 500, 5479

Banerjee A., Abel T., 2021b, Monthly Notices of the Royal Astronomical Society, 504, 2911

Banerjee A., Abel T., 2023, Monthly Notices of the Royal Astronomical Society, 519, 4856

Banerjee A., Kokron N., Abel T., 2022, Monthly Notices of the Royal Astronomical Society, 511, 2765

Bera S., Rana D., More S., Bose S., 2020, The Astrophysical Journal, 902, 79

Bilicki M., et al., 2016, The Astrophysical Journal Supplement Series, 225, 5

Borhanian S., Sathyaprakash B. S., 2022, Listening to the Universe with Next Generation Ground-Based Gravitational-Wave Detectors, doi:10.48550/arXiv.2202.11048, http://arxiv.org/abs/2202.11048

Calore F., Cuoco A., Regimbau T., Sachdev S., Serpico P. D., 2020, Physical Review Research, 2, 023314

Cavaglià M., Modi A., 2020, Universe, 6, 93

Chen H.-Y., Essick R., Vitale S., Holz D. E., Katsavounidis E., 2017, The Astrophysical Journal, 835, 31

Chruślińska M., 2024, Annalen Phys., 536, 2200170

Collister A. A., Lahav O., 2004, Publications of the Astronomical Society of the Pacific, 116, 345

Cutri R. M., et al., 2013, Technical report, Explanatory Supplement to the AllWISE Data Release Products, https://ui.adsabs.harvard.edu/abs/2013wise.rept....1C. https://ui.adsabs.harvard.edu/abs/2013wise.rept....1C

DES Collaboration et al., 2022a, Physical Review D, 105, 023520

DES Collaboration et al., 2022b, Physical Review D, 106, 043520

DES Collaboration et al., 2022c, Physical Review D, 106, 103530

DES Collaboration et al., 2023, Physical Review D, 107, 083504

Davis M., Huchra J., Latham D. W., Tonry J., 1982, The Astrophysical Journal, 253, 423

Devaraju B., 2015, doctoralThesis, doi:10.18419/opus-3985, http://elib.uni-stuttgart.de/handle/11682/4002

Dvornik A., et al., 2023, Astronomy & Astrophysics, 675, A189

Essick R., 2023, Physical Review D, 108, 043011

Ezquiaga J. M., Holz D. E., 2022, Phys. Rev. Lett., 129, 061102

Falco E. E., et al., 1999, Publications of the Astronomical Society of the Pacific, 111, 438

Fang K., Banerjee A., Charles E., Omori Y., 2020, The Astrophysical Journal, 894, 112

Farr W. M., Fishbach M., Ye J., Holz D. E., 2019, The Astrophysical Journal Letters, 883, L42

Fishbach M., Kalogera V., 2021, The Astrophysical Journal Letters, 914, L30

Fishbach M., Holz D. E., Farr W. M., 2018, The Astrophysical Journal Letters, 863, L41

Fumagalli, A. Costanzi, M. Saro, A. Castro, T. Borgani, S. 2024, A&A, 682, A148

Gagnon E. L., Anbajagane D., Prat J., Chang C., Frieman J., 2023, arXiv e-prints, p. arXiv:2312.16289

Gair J. R., et al., 2023, The Astronomical Journal, 166, 22

Geller M. J., Huchra J. P., 1989, Science, 246, 897

Gupta K. R., Banerjee A., 2024, arXiv e-prints, p. arXiv:2404.01428

Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, The Astrophysical Journal, 622, 759

Hall E. D., Evans M., 2019, Classical and Quantum Gravity, 36, 225002

Hambly N., et al., 2001, Monthly Notices of the Royal Astronomical Society, 326, 1279

Hartlap, J. Simon, P. Schneider, P. 2007, A&A, 464, 399

Holz D. E., Hughes S. A., 2005, The Astrophysical Journal, 629, 15

Holz D. E., Hughes S. A., Schutz B. F., 2018, Physics Today, 71, 34

Hoy C., Raymond V., 2021, SoftwareX, 15

Hu J.-P., Wang F.-Y., 2023, Universe, 9, 94

Huchra J. P., Vogeley M. S., Geller M. J., 1999, The Astrophysical Journal Supplement Series, 121, 287

Iacovelli F., Mancarella M., Foffa S., Maggiore M., 2022, The Astrophysical Journal, 941, 208

Klypin A., Yepes G., Gottlöber S., Prada F., Heß S., 2016, Monthly Notices of the Royal Astronomical Society, 457, 4340

Krakowski T., Małek K., Bilicki M., Pollo A., Kurcz A., Krupa M., 2016, Astronomy & Astrophysics, 596, A39

LIGO Scientific Collaboration and Virgo Collaboration et al., 2016, Physical Review Letters, 116, 061102

LIGO Scientific Collaboration a. J. A., et al., 2015, Classical and Quantum Gravity, 32, 074001

LIGO Scientific Collaboration and KAGRA Collaboration V. C., et al., 2023a, Physical Review X, 13, 011048

LIGO Scientific Collaboration and KAGRA Collaboration V. C., et al., 2023b, Physical Review X, 13, 041039

Landy S. D., Szalay A. S., 1993, The Astrophysical Journal, 412, 64

Libanore S., et al., 2021, Journal of Cosmology and Astroparticle Physics, 2021, 035

Libanore S., Artale M., Karagiannis D., Liguori M., Bartolo N., Bouffanais Y., Mapelli M., Matarrese S., 2022, Journal of Cosmology and Astroparticle Physics, 2022, 003

Lochner M., et al., 2018, arXiv e-prints, p. arXiv:1812.00515

MacLeod C. L., Hogan C. J., 2008, Phys. Rev. D, 77, 043512

Madau P., Dickinson M., 2014, Annual Review of Astronomy and Astrophysics, 52, 415

Mancarella M., Genoud-Prachex E., Maggiore M., 2022, Physical Review D, 105, 064030

Mastrogiovanni S., et al., 2021, Phys. Rev. D, 104, 062009

Mastrogiovanni S., Karathanasis C., Gair J., Ashton G., Rinaldi S., Huang H.-Y., Dálya G., 2024, Annalen der Physik, 536, 2200180

Meena A. K., Bagla J. S., 2019, Monthly Notices of the Royal Astronomical Society, 492, 1127

Michael Burgess, J. Cameron, Ewan Svinkin, Dmitry Greiner, Jochen 2021, A&A, 654, A26

Miyatake H., et al., 2023, Physical Review D, 108, 123517

Mukherjee S., Wandelt B. D., Nissanke S. M., Silvestri A., 2021, Phys. Rev. D, 103, 043520

Mukherjee S., Krolewski A., Wandelt B. D., Silk J., 2022, Cross-correlating dark sirens and galaxies: measurement of H_0 from GWTC-3 of LIGO-Virgo-KAGRA, doi:10.48550/arXiv.2203.03643, http://arxiv.org/abs/2203.03643

Namikawa T., Nishizawa A., Taruya A., 2016, Phys. Rev. Lett., 116, 121302

Oguri M., 2016, Phys. Rev. D, 93, 083511

Omohundro S. M., 2009. https://api.semanticscholar.org/CorpusID:61067117

Omori Y., 2022, arXiv e-prints, p. arXiv:2212.07420

Palmese A., et al., 2020, The Astrophysical Journal Letters, 900, L33

Pedregosa F., et al., 2012, Scikit-learn: Machine Learning in Python, https://arxiv.org/abs/1201.0490v4

Peron M., Libanore S., Ravenni A., Liguori M., Artale M. C., 2023, arXiv e-prints, p. arXiv:2305.18003

Pratten G., et al., 2021, Physical Review D, 103, 104056

RIOS J. d. M. Y., 1795, Memoria sobre algunos Métodos nuevos de calcular la Longitud por las distancias lunares: y aplicacion de su teórica á la solucion de otras problemas de navegacion. (Tabla, etc.).. Imp. Real

Raccanelli A., Kovetz E. D., Bird S., Cholis I., Muñoz J. B., 2016, Phys. Rev. D, 94, 023516

Reid I. N., et al., 1991, Publications of the Astronomical Society of the Pacific, 103, 661

Riess A. G., 2020, Nature Reviews Physics, 2, 10

Riess A. G., et al., 2018, The Astrophysical Journal, 855, 136

Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, The Astrophysical Journal, 876, 85

Riess A. G., et al., 2022, The Astrophysical journal letters, 934, L7

Saleem M., et al., 2022, Classical and Quantum Gravity, 39, 025004

Santoliquido F., Mapelli M., Giacobbo N., Bouffanais Y., Artale M. C., 2021, Monthly Notices of the Royal Astronomical Society, 502, 4877

Scelfo G., Bellomo N., Raccanelli A., Matarrese S., Verde L., 2018, Journal of Cosmology and Astroparticle Physics, 2018, 039

Schutz B. F., 1986, Nature, 323, 310

Singer L. P., Price L. R., 2016, Physical Review D, 93, 024013

Soares-Santos M., et al., 2019, The Astrophysical Journal Letters, 876, L7

Stoughton C., et al., 2002, The Astronomical Journal, 123, 485

Strauss M. A., et al., 2002, The Astronomical Journal, 124, 1810

Talbot C., Thrane E., 2018, The Astrophysical Journal, 856, 173

The LSST Dark Energy Science Collaboration et al., 2018, arXiv e-prints, p. arXiv:1809.01669

Trott E., Huterer D., 2022, Challenges for the statistical gravitational-wave method to measure the Hubble constant (arXiv:2112.00241)

Valentino E. D., et al., 2021, Classical and Quantum Gravity, 38, 153001

Vijaykumar A., Fishbach M., Adhikari S., Holz D. E., 2023a, arXiv e-prints, p. arXiv:2312.03316

Vijaykumar A., Saketh M., Kumar S., Ajith P., Choudhury T. R., 2023b, Physical Review D, 108, 103017

Wang Y., Banerjee A., Abel T., 2022, Monthly Notices of the Royal Astronomical Society, 514, 3828

Wong K. C., et al., 2020, Monthly Notices of the Royal Astronomical Society, 498, 1420

Wright E. L., et al., 2010, The Astronomical Journal, 140, 1868

Zheng Y., Kouvatsos N., Golomb J., Cavaglià M., Renzini A. I., Sakellariadou M., 2023, Physical Review Letters, 131, 171403

Zonca A., Singer L. P., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K. M., 2019, Journal of Open Source Software, 4, 1298

Željko Ivezić et al., 2019, The Astrophysical Journal, 873, 111

Appendices

Appendix A

Smoothing in harmonic space

Consider a field $\delta(\Omega)$ defined in the sky. The expression for the field smoothed on an angular scale θ is given by

$$\delta^{\theta}(\Omega) = \frac{1}{2\pi(1-\cos\theta)} \int_{\arccos(\hat{\Omega}\cdot\hat{\Omega}')\leq\theta} d\hat{\Omega}'\,\delta(\hat{\Omega}') \tag{A.1}$$

The smoothed field is equivalent to the field averaged over spherical caps of angular radius θ. Equation A.1 can be re-written as

$$\delta^{\theta}(\Omega) = \int_{\text{All sky}} d\hat{\Omega}'\,\delta(\hat{\Omega}')W^{\theta}(\Omega',\Omega) \tag{A.2}$$

where $W^{\theta}(\Omega',\Omega)$ is the top-hat filter in configuration space, given by

$$W^{\theta}(\Omega',\Omega) = \begin{cases} \frac{1}{2\pi(1-\cos\theta)} & \arccos(\hat{\Omega}\cdot\hat{\Omega}') \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{A.3}$$

The integral in equation A.3 is computationally expensive to perform in configuration space, but one can use properties of the spherical harmonics and massively speed up the calculation by going to harmonic space. Let the expansions of the smoothed and unsmoothed fields in spherical

harmonics be given by

$$\delta(\hat{\Omega}) = \sum_{\ell m} \alpha_{\ell m} Y_{\ell m}(\hat{\Omega}) \tag{A.4}$$

$$\delta^{\theta}(\hat{\Omega}) = \sum_{\ell m} \alpha^{\theta}_{\ell m} Y_{\ell m}(\hat{\Omega}) \tag{A.5}$$

Since the top-hat filter represents a homogeneous and isotropic smoothing kernel that is only a function of the central angle θ between $\hat{\Omega}$ and $\hat{\Omega}'$, it can be expanded in terms of the Legendre polynomials $P_\ell(\cos\theta)$ as:

$$W^{\theta} = \sum_{\ell} b_\ell P_\ell(\cos\theta) \tag{A.6}$$

It can be shown, by substituting equations A.4 to A.6 in equation A.2, that the spherical harmonic expansion coefficients $\alpha^{\theta}_{\ell m}$ of the smoothed field are given by the product of the $\alpha_{\ell m}$ of the unsmoothed field and the Legedre expansion coefficients b_ℓ of the top-hat filter (Devaraju, 2015)[1]:

$$\alpha^{\theta}_{\ell m} = 4\pi \frac{b_\ell}{2\ell+1} \alpha_{\ell m} \tag{A.7}$$

This expression makes sense since computing the integral in equation A.3 is equivalent to performing a *spatial convolution* on the surface of a sphere. A convolution in configuration space corresponds to a product in harmonic space. It is to be noted that for a general inhomogeneous smoothing kernel, however, spatial smoothing is not equivalent to a convolution since each point on the sphere has a different kernel. For such cases, equation A.7 would be different. See chapter 2 of Devaraju (2015) for a nice discussion on the topic.

In practice, we compute the $\{\alpha_{\ell m}\}$ for the unsmoothed field using the `healpy` package. The $\{b_\ell\}$ for the top-hat function are computed as follows. By definition,

$$\begin{aligned} b_\ell &= \frac{2\ell+1}{2} \int_{-1}^{1} W^{\theta} P_\ell(\cos\theta) d(\cos\theta) \\ &= \frac{2\ell+1}{2} \frac{1}{2\pi(1-\cos\theta)} \int_{\cos\theta}^{1} P_\ell(\cos\theta) d(\cos\theta) \end{aligned} \tag{A.8}$$

[1]Note that we have an additional factor of 4π not present in the expression derived by Devaraju (2015), since we use a normalised definition for the top-hat filter that already incorporates an extra factor of $1/4\pi$ that appears in their equivalent of equation A.2.

Using the recursion relation

$$P_\ell(x) = \frac{1}{2\ell+1}\frac{d}{dx}\left[P_{\ell+1}(x) - P_{\ell-1}(x)\right] \tag{A.9}$$

we get

$$b_\ell = \frac{1}{4\pi(1-\cos\theta)}\left[P_{\ell-1}(\cos\theta) - P_{\ell+1}(\cos\theta)\right] \tag{A.10}$$

Appendix B

Population Models for Binary Black Holes

For creating the mock BBH catalogues, we assume the `Power Law + Peak` model for the mass of the primary (heavier) BBH and a power law distribution for the ratio of component masses. For the analysis using the currently available data, we assume a power law distribution for redshift evolution of merger rate per unit comoving volume per unit source-frame time, as specified in LIGO Scientific Collaboration et al. (2023a), while for the the forecast study, we assume that the BBH merger rate follows the Madau-Dickinson star formation rate density (Madau & Dickinson, 2014).

B.1 Mass Model

Let m_1 denote the mass of the primary black hole, and q denote the mass ratio, such that the mass of the secondary black hole is qm_1. The population distribution model for m_1 is given by

$$\pi(m_1|\lambda_{\rm peak},\alpha,m_{\rm min},\delta_m,m_{\rm max},\mu_m,\sigma_m) = \left[(1-\lambda_{\rm peak})\mathscr{P}(m_1|-\alpha,m_{\rm max}+\lambda_{\rm peak}G(m_1|\mu_m,\sigma_m)\right]S(m_1|m_{\rm min},\delta_m) \tag{B.1}$$

where $\mathscr{P}(m_1|-\alpha,m_{\rm max})$ is a power law distribution with spectral index $-\alpha$ and high-mass cut-off $m_{\rm max}$, $G(m_1|\mu_m,\sigma_m)$ is a Gaussian distribution with mean μ_m and standard deviation σ_m, and $\lambda_{\rm peak}$ is the mixing fraction that determines the relative importance of the power law and Gaus-

Parameter	Value
α	3.4
β_q	1.08
$m_{\rm min}$	5.08
$m_{\rm max}$	86.85
$\lambda_{\rm peak}$	0.04
μ_m	33.73
σ_m	3.56
δ_m	4.83

Table B.1: `Power Law + Peak` model parameters

sian components. The lower mass end of the distribution is tapered using a smoothing function $S(m_1|m_{\rm min},\delta_m)$ which rises from 0 to 1 over the interval $(m_{\rm min}, m_{\rm min}+\delta_m)$, given by

$$S(m_1|m_{\rm min},\delta_m) = \begin{cases} 0 & m < m_{\rm min} \\ [f(m-m_{\rm min},\delta_m)+1]^{-1} & m_{\rm min} \leq m < m_{\rm min}+\delta_m \\ 1 & m \geq m_{\rm min}+\delta_m \end{cases} \tag{B.2}$$

with

$$f(m,\delta_m) = \exp\left(\frac{\delta_m}{m} + \frac{\delta_m}{m-\delta_m}\right) \tag{B.3}$$

The conditional mass ratio distribution is given by a power law, also smoothed at the lower mass end

$$\pi(q|m_1,\beta_q,m_{\rm min},\delta_m) \propto q^{\beta_q} S(qm_1|m_{\rm min},\delta_m) \tag{B.4}$$

The values for the model parameters assumed in this work are taken from the publicly available LVK population analysis results and are summarised in table B.1.

B.2 Redshift Evolution Models

The power law redshift evolution model parameterises the merger rate density per comoving volume and source-frame time as

$$\mathscr{R}(z) = \frac{dN}{dV_c dt_s} = \mathscr{R}_0(1+z)^{\kappa} \tag{B.5}$$

where $\mathscr{R}_0$ is the merger rate density at $z = 0$, t_s is the source-frame time, related to observer-frame time as $t_s = t_o/(1+z)$ due to cosmological redshift. This implies that the observed redshift distribution is

$$\begin{aligned} \frac{dN}{dz} &= \int dt_o \frac{dV_c}{dz} \mathscr{R}_0 (1+z)^{\kappa-1} \\ &= t_{obs} \mathscr{R}_0 \frac{dV_c}{dz} (1+z)^{\kappa-1} \end{aligned} \tag{B.6}$$

where t_{obs} is the total observation time and $\frac{dV_c}{dz}$ is the differential comoving volume. The probability distribution function for redshifts is given by normalising equation B.6

$$\pi(z|\kappa, z_{max}) = \frac{\frac{dV_c}{dz}(1+z)^{\kappa-1}}{\int_0^{z_{max}} dz \frac{dV_c}{dz}(1+z)^{\kappa-1}} \tag{B.7}$$

where z_{max} is the maximum redshift out to which the population has been created. Note that the constants $\mathscr{R}_0$ and t_{obs} drop out of the expression since they are simply normalisation constants and do not affect the shape of the distribution. However, they do indeed control the total number of mock events to be drawn from the normalised distribution. Although the z_{max} in LIGO analyses is typically taken to be 2.3, we assume a z_{max} of 0.7 since assuming a lower value of z_{max} is computationally less demanding. We have checked that inputting a higher z_{max} value does not affect our mock catalogue. This is due to the fact that a negligible fraction of injections outside this redshift are detected by our assumed network. We assume $\kappa = 3$ in our analysis.

The Madau Dickinson star formation rate profile gives the star formation rate density per comoving volume and source-frame time as

$$\psi(z) \propto \frac{(1+z)^{2.7}}{1 + \left(\frac{1+z}{1+1.9}\right)^{5.6}} \tag{B.8}$$

We assume that the BBH merger rate follows the star formation rate exactly (see section 4.1.2 for details), therefore, we must have

$$\mathscr{R}(z) = \mathscr{R}_0 \frac{(1+z)^{2.7}}{1+\left(\frac{1+z}{1+1.9}\right)^{5.6}} \tag{B.9}$$

As before, $\mathscr{R}_0$ is approximately[1] the merger rate density at $z = 0$ and t_s is the source-frame time, related to observer-frame time as $t_s = t_o/(1+z)$ due to cosmological redshift. This implies that the observed redshift distribution is

$$\begin{aligned} \frac{dN}{dz} &= \int dt_o \frac{dV_c}{dz} \mathscr{R}_0 \frac{(1+z)^{1.7}}{1+\left(\frac{1+z}{1+1.9}\right)^{5.6}} \\ &= t_{obs} \mathscr{R}_0 \frac{dV_c}{dz} \frac{(1+z)^{1.7}}{1+\left(\frac{1+z}{1+.9}\right)^{5.6}} \end{aligned} \tag{B.10}$$

where t_{obs} is the total observation time and $\frac{dV_c}{dz}$ is the differential comoving volume. The probability distribution function for redshifts is given by normalising equation B.10

$$\pi(z|z_{max}) = \frac{\frac{dV_c}{dz} \frac{(1+z)^{1.7}}{1+\left(\frac{1+z}{1+1.9}\right)^{5.6}}}{\int_0^{z_{max}} dz \frac{dV_c}{dz} \frac{(1+z)^{1.7}}{1+\left(\frac{1+z}{1+1.9}\right)^{5.6}}} \tag{B.11}$$

where z_{max} is the maximum redshift out to which the population has been created. For the forecast study, we set $z_{max} = 4$, $t_{obs} = 10$ and $\mathscr{R}_0 = 17$ Gpc^{-3} Yr^{-1} in our analysis.

[1] This is a very good approximation, since $\mathscr{R}(0) \approx \mathscr{R}_0$