EPJ Web of Conferences **206**, 09017 (2019) https://doi.org/10.1051/epjconf/201920609017
ISMD 2018

Clustering of hotspots in the cosmic microwave background

Low *En Zuo Joel*[1] and *Abel* Yang[1,*]

[1]Department of Physics, National University of Singapore

Abstract. The physics behind the origin and composition of the Cosmic Microwave Background (CMB) is a well-established topic in the field of Cosmology. Literature on CMB anisotropies reveal consistency with Gaussianity [1], but these were conducted on full multi-frequency temperature maps. In this thesis, we utilise clustering algorithms to specifically conduct statistical analyses on the distribution of hotspots in the CMB. We describe a series of data processing and clustering methodologies conducted, with results that conclusively show that the counts-in-cells distribution of hotspots in the CMB does not follow a Poisson distribution. Rather, the distribution exhibits a much closer fit to both the Negative Binomial Distribution (NBD) and the Gravitational Quasi-Equilibrium Distribution (GQED). From this result, we conclude that structure likely existed in the early universe, from the period of the recombination Epoch, possibly opening new insights in the field of galaxy formation.

1 Introduction

1.1 Cosmic Microwave Background

In the current theory of Big Bang Cosmology, the cosmic microwave background (CMB) is leftover electromagnetic radiation from the primordial stages of the formation of our universe. While the CMB may be consistent with Gaussianity on a whole [1], detailed observations reveal pockets of anisotropy scattered across the distribution, forming a pattern similar to that of a hot gas that has expanded over time. Till date, no studies have been conducted specifically on the statistical nature of the distribution of hotspots in the CMB. As such, we utilise clustering algorithms to specifically conduct statistical analyses on the distribution of these hotspots.

1.2 Hypothesis

Through clustering analysis of CMB data, we seek to compare the distribution of hotspots in the CMB with the negative binomial distribution (NBD) and gravitational quasi-equilibrium distribution (GQED). These distributions were chosen as they provide a good description of the counts-in-cells distributions of galaxies [2, 3].

If the GQED or NBD in fact proves to be a close match to the distributions of hot spots in the CMB, we will be able to conclude that matter in the recombination epoch already possessed some form of structure.

*e-mail: phyyja@nus.edu.sg

2 Materials & Methods

2.1 Planck CMB Temperature Map

The Planck Space Telescope was sent into orbit in May 2009 by the European Space Agency (ESA) to survey the CMB. Out of the published CMB maps, the SMICA product is labelled as preferred by the ESA and as such, was selected for further analysis [4]. A Heaviside filter (95th percentile) was applied on the CMB Temperature map. The end result is a list of boolean HEALpix map, from which we can obtain the galactic longitude and latitude of each hot pixel.

To exclude possible sources of foreground contamination from the galaxy, such as galactic dust that lie in the plane of the milky way, we apply a mask on the dataset by excluding all points 10° above and below the galactic equator.

2.2 Clustering Algorithm: HDBSCAN

Conventional clustering algorithms (e.g. *k*-means) make assumptions that do not necessarily hold true for the SMICA dataset. We expect the real data to be noisy which can serve create apparent bridges between two separate clusters, leading to inaccurate clustering results. We also expect clusters of varying densities, and non-spherical clusters to be a concern. The total number of hotspots in the dataset was also an unknown.

HDBSCAN is an acronym which stands for "Hierarchical Density-Based Spatial Clustering Application with Noise" and was eventually selected as the clustering algorithm of choice as it was able to succinctly deal with the above stated conditions [5].

2.3 Probability Distributions

The galaxy counts-in-cells (CIC) distribution describes the spatial location of galaxies. It can be generalised to a form of $f(N, V)$, giving the probability of finding N hot spots in a region of volume V, or solid angle Ω. In the $f_V(N)$ form, the volume is taken to be constant.

The probability mass function and variance of the GQED is as follows:

$$f_V(N) = \frac{\bar{N}(1-b)}{N!}\left(\bar{N}(1-b)+Nb\right)^{N-1} e^{-\bar{N}(1-b))-Nb} \tag{1}$$

$$\langle(\Delta N)^2\rangle = \frac{\bar{N}}{(1-b)^2} \tag{2}$$

where $\bar{N}$ is the mean of the distribution and b is a clustering parameter between 0 and 1.

The probability mass function and variance of the negative binomial distribution (NBD) is as follows:

$$f_V(N) = \frac{\Gamma\left(N+\frac{1}{g}\right)}{\Gamma\left(\frac{1}{g}\right)N!}\frac{\bar{N}^N\left(\frac{1}{g}\right)^{\frac{1}{g}}}{\left(\bar{N}+\frac{1}{g}\right)^{N+\frac{1}{g}}} \tag{3}$$

$$\langle(\Delta N)^2\rangle = \bar{N}^2 g + \bar{N} \tag{4}$$

where $\bar{N}$ is the mean of the distribution and g is a parameter greater than 0 related to the variance.

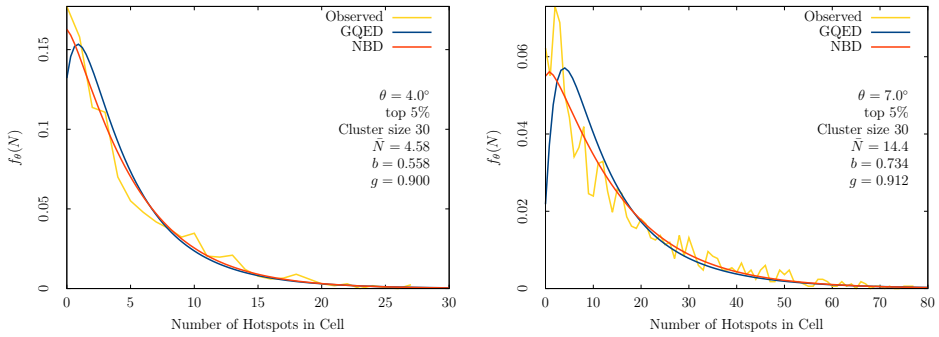

Figure 1. Comparison between NBD and GQED for cell sizes of 4.0° and 7.0°

2.4 Steps taken to obtain CIC distribution of hotspots in CMB

From the masked and filtered CMB data, we use HDBSCAN to obtain the positions of cluster centres, using a haversine metric on the galactic longitude and latitude of each hot pixel.

To lay down the cells, we spread cell centres evenly across the unmasked area. The physical extent of each cell is the cap with radius θ on the surface of a sphere goving a solid angle of $4\pi \sin^2(\theta/2)$. We then count the number of hotspots that lie within each cell to obtain the counts-in-cells distribution.

3 Results

3.1 Least squares goodness of fit measure

The least squares goodness of fit (LSGF) measure presents a quantitative measure of the goodness of fit between an observed, and expected distribution, given by

$$X = \sum_{i=0}^{N} (O_i - E_i)^2 \,. \tag{5}$$

This is the sum of the squares of the difference between the expected and observed $f_V(N)$. Smaller values of X indicate a closer fit.

From figure 1, we see the NBD consistently outperforms the GQED, with a smaller LSGF value across most of the tested parameter sets.

3.2 Accounting for variance by resampling

Variations in the distribution of hotspots in the CMB can result in sub-volumes which are not statistically similar. To obtain an indication of the possible variations across the sky, we conduct a resampling procedure where for each instance, we exclude 25% of the cells based on the galactic longitude of the cell centre. We obtain separate instances by shifting the window by $\pi/3$ radians each time, with an overlap of $\pi/6$ radians.

The minimum and maximum histogram values are then combined to obtain a window of confidence. An optimum result would be a tight window with only one of the fitted probability distribution functions falling into it, indicating an accurate fit with high confidence.

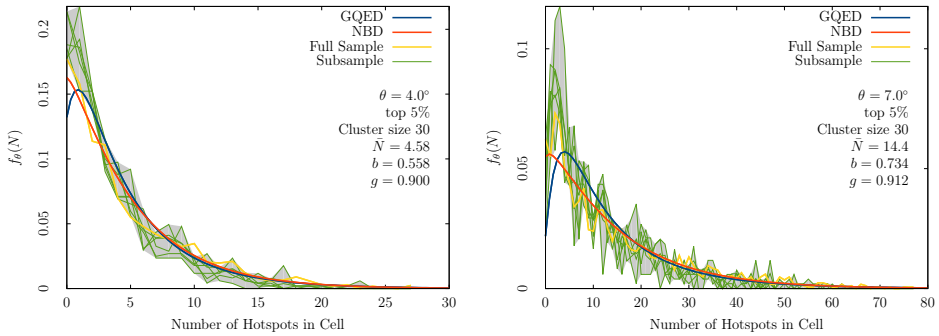

Figure 2. Counts-in-cells for different resampling windows for cell sizes of 4.0° and 7.0°

From figure 2, we observe that the NBD and GQED trace out almost identical paths, diverging only at low values of N. There are however large variations close to small values of N, suggesting the strong presence of noise in the data. This noise is likely to have originated from cosmic variance in the data collected by ESA, and it explains the greater impact in the small N region.

4 Discussion & Conclusion

After obtaining all LSGF values, we compared the results between the NBD and GQED. Out of all 396 parameter choices, the counts-in-cells distribution follows the NBD and GQED somewhat closely, indicating that the hotspots in the CMB are in fact not randomly distributed.

Comparing the closeness of fit between the NBD and GQED, we observe that the GQED generally under predicts the number of voids where $N = 0$. This leads to the NBD presenting a closer fit at small values of N. However, we are unable to conclusively state that the NBD is a closer fit as the resampling window is noisy near low N regions. Also, the NBD is unphysical description of galaxy clustering, violating the 2nd law of thermodynamics [6].

For future work, we intend to use a more comprehensive galactic foreground contamination mask and higher resolution temperature maps. We also intend to consider different percentile cut-offs for the initial Heaviside filter (e.g. 1-σ instead of just 95th percentile).

References

[1] Planck Collaboration, P.A.R. Ade, N. Aghanim, Y. Akrami, P.K. Aluri, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A.J. Banday et al., Astron. & Astrophys. **594**, A16 (2016)

[2] L. Hurtado-Gil, V.J. Martínez, P. Arnalte- Mur, M.J. Pons-Bordería, C. Pareja-Flores, S. Paredes, Astron. & Astrophys. **601**, A40 (2017)

[3] A. Yang, W.C. Saslaw, Astrophys. J. **729**, 123 (2011)

[4] Planck Collaboration, R. Adam, P.A.R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A.J. Banday, R.B. Barreiro et al., Astron. & Astrophys. **594**, A9 (2016)

[5] L. McInnes, J. Healy, S. Astels, Journal of Open Source Software **2(11)**, 205 (2017)

[6] W.C. Saslaw, F. Fang, Astrophys. J. **460**, 16 (1996)