Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921314010886

Towards a better understanding of galaxy clusters

Pedro T. P. Viana[1,2]

[1]Centro de Astrofísica da Universidade do Porto,
Rua das Estrelas, 4150-762 Porto, Portugal

[2]Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto,
Rua do Campo Alegre, 687, 4169-007 Porto, Portugal
email: viana@astro.up.pt

Abstract. Observational data on clusters of galaxies holds relevant information that can be used to determine the relative plausibility of different models for the large-scale evolution of the Universe, or estimate the joint posterior probability distribution function of the parameters that pertain to each model. Within the next few years, several surveys of the sky will yield large galaxy cluster catalogues. In order to make use of the vast amount of information they will contain, their selection functions will have to be properly understood. We argue this, as well as the estimation of the full joint posterior probability distribution function of the most relevant cluster properties, can be best achieved in the framework of bayesian statistics.

Keywords. galaxies: clusters: general, surveys, methods: statistical

1. Introduction

Clusters of galaxies are the largest gravitationally collapsed structures in the Universe. The hierarchical growth of large-scale structure also ensures they are the rarest. Further, although not all important physical processes involved in the assembly of galaxy clusters are well known, they are relatively simple astrophysical objects. All of these characteristics make clusters of galaxies excellent probes of the growth of structure in the Universe and of its large-scale evolution.

Within the next few years, several surveys of the sky will detect many thousands of galaxy clusters. The largest catalogue is expected to result from the Euclid mission (*www.euclid-ec.org*). However, the accuracy with which we will be able to distinguish, for example, competing hypothesis for the cause of the present accelerating expansion of the Universe, will be limited by our understanding of the galaxy cluster catalogue selection function, systematic errors and the uncertainty in the galaxy cluster mass determinations. The combination of Euclid data with that obtained through other galaxy cluster surveys, based on the detection of the galaxy cluster signal on the X-rays (XCS, *www.xcs-home.org*, and eROSITA, *www.mpe.mpg.de/eROSITA*, surveys) and due to the Sunyeav-Zel'dovich effect in the mm/sub-mm (SPT, *www.pole.uchicago.edu/spt*, and Planck, *www.sci.esa.int/planck*, surveys), can help in this respect, by enabling the cross-calibration of the selection functions of those surveys and the mass-observable relations, unearthing possible systematic errors in the process.

2. Overview

Let $\xi = \{\xi_{\alpha,i}\}$ be the set of all galaxy cluster quantities of interest, each associated with a different α, for all clusters in some catalogue, each identified by a given i. We would like to characterise the probability distribution for any $\xi_{\alpha,i}$ given the information contained in some dataset, D, and any relevant prior information, I_0,

$$P(\xi_{\alpha,i}|D,I_0) = \int P(\xi|D,I_0) d\xi_{\beta,j} \tag{2.1}$$

where the integral runs over all possible combinations of β and j that are different from the combination of α and i. Using Bayes Theorem, we can now write

$$P(\xi_{\alpha,i}|D,I_0) \propto \int P(\xi|I_0) P(D|\xi,I_0) d\xi_{\beta,j} \tag{2.2}$$

where the proportionality or normalisation constant is equal to the inverse of the model evidence, $P(D|I_0)$. The estimation of the likelihood, $P(D|\xi,I_0)$, is usually difficult given that the data available can be heterogenous, always consisting of a finite number of measurements affected by heteroscedastic noise and possibly correlated.

The problem simplifies considerably if the data pertaining to each galaxy cluster, D_i, is acquired independently. We can thus write

$$P(\xi_{\alpha,i}|D_i,I_0) \propto \int P(\xi|I_1) P(D_i|\xi,I_1) d\xi_{\beta,i} \tag{2.3}$$

Assuming that all galaxy clusters in the catalogue belong to the same statistical population, the prior probability inside the integral becomes independent of i. And it can be inferred from the rest of the data, that we will continue to call D,

$$P(\xi|I_1) = P(\xi|D,I_2) \propto P(\xi|I_2) P(D|\xi,I_2) \tag{2.4}$$

The most general way to evaluate the likelihood in the previous expression is by using a non-parametric or semi-parametric density estimation procedure (e.g. Bovy, Hogg & Roweis 2011, Sarkar *et al.* 2014). The prior probability can be minimally informative, or include any assumptions that may follow from what we believe were the physical conditions under which galaxy clusters assembled. Presently, the most credible prior assumptions follow from the so-called standard cosmological model (e.g. Lahav & Liddle 2014). Among them, it is particularly relevant the halo mass function, $n(M,z)$, describing the number density of gravitationally collapsed structures as a function of their mass, M, and redshift, z. If the data being considered contains information about the masses and redshifts of at least some of the galaxy clusters in the catalogue, then we can write

$$P(\xi|I_2) = P(\xi_\gamma|M,z,I_3) P(M,z|I_3) = P(\xi_\gamma|M,z,I_4) \frac{n(M,z)}{\int n(M,z) dM dz} \tag{2.5}$$

where γ identifies all cluster quantities of interest except for mass and redshift, while $P(\xi_\gamma|M,z,I_4)$ is a minimally informative prior. Otherwise, assumptions about relations that connect cluster mass and redshift to some other cluster property, information about which is contained in the dataset D, will have to be included in the prior assumptions.

The halo mass function depends on the values taken by some parameters in the standard cosmological model, whose set we will denote by θ, like the mean matter density in the Universe. Thus, in fact

$$P(\xi|I_2) = P(\xi_\gamma|M,z,I_4) \int P(\theta|I_4) \frac{n(M,z,\theta)}{\int n(M,z,\theta) dM dz} d\theta \tag{2.6}$$

where $P(\theta|I_4)$ represents the prior distribution of the cosmological parameters.

https://doi.org/10.1017/S1743921314010886 Published online by Cambridge University Press

In the previous expression, the full joint posterior distribution was marginalised over the cosmological parameters, θ. If we had marginalised instead over ξ, we would have obtained the posterior distribution of those parameters given the dataset, D,

$$P(\theta|D, I_0) \propto \int P(\xi|\theta, I_0)P(\theta|I_0)P(D|\theta, \xi, I_0)d\xi \tag{2.7}$$

where $P(\theta|I_0)$ is a minimally informative prior. In an analogous manner to expression (2.5), we could then write

$$P(\xi|\theta, I_0) = P(\xi_\gamma|M, z, I_4)\frac{n(M, z, \theta)}{\int n(M, z, \theta)dMdz} \tag{2.8}$$

The density estimation procedure needed to evaluate $P(D|\xi, I_2)$ or $P(D|\theta, \xi, I_0)$ can be computationally very intensive. This problem can be alleviated if so-called galaxy cluster scaling relations are used to describe how the quantities ξ_α relate to each other. These are often assumed to be simple power-laws (e.g. Kelly 2007, Maughan 2014), that can be linearised by changing to $\log(\xi_\alpha)$, with some associated (so-called intrinsic) scatter. This is usually taken to be normally distributed with respect to $\log(\xi_\alpha)$, and independent from the values that $\log(\xi_\alpha)$ may take. If we label as S the set of parameters that define such relations, then expression (2.4) becomes

$$P(\xi|I_1) = \int P(\xi, S|D, I_5)dS \propto \int P(\xi|S, I_6)P(S|I_6)P(D|\xi, S, I_6)dS \tag{2.9}$$

where $P(\xi|S, I_6)$ and $P(S|I_6)$ can be minimally informative priors. In this case, had we marginalised instead over ξ, we would have obtained the posterior distribution of the linear regression parameters, S, given the dataset, D,

$$P(S|D, I_0) \propto \int P(\xi|S, I_7)P(S|I_7)P(D|S, \xi, I_7)d\xi \tag{2.10}$$

Taking into account prior information about the cosmological model would imply using expressions (2.5) or (2.6) in both expressions (2.9) and (2.10). Further, expressions (2.7) and (2.10) can be combined to infer the joint posterior distribution function of θ and S.

Finally, it should be remembered that all prior probabilities have to include the effects of the selection procedures followed in the assembly of the galaxy cluster catalogues considered. These depend on at least one galaxy cluster property, and most often also on the assumed cosmological model (e.g. Lloyd-Davies *et al.* 2011).

Acknowledgements

This work is being carried out with financial support from Fundação para a Ciência e a Tecnologia (FCT), through project PTDC/FIS/111725/2009.

References

Bovy, J., Hogg, D. W., & Roweis, S. T. 2011, *Annals of Applied Statistics*, 5, 2B, 1657
Sarkar, A., Pati, D., Mallick, B. K., & Carroll, R. J. 2014, *arXiv:1404.6462*
Lahav, O. & Liddle, A. R. 2014, *The Review of Particle Physics 2014*, ch. 23
Kelly, B. C. 2007, *ApJ*, 665, 1489
Maughan, B. J. 2014, *MNRAS*, 437, 1171
Lloyd-Davies, E. J., *et al.* (XCS collaboration), 2011, *MNRAS*, 418, 14