Spatially Weighted and Measurement Error Corrected Gaussian Mixture Model for Galaxy Clustering Analysis

Jiangang Hao

Center for Particle Astrophysics, Fermi National Accelerator Laboratory, Batavia, IL

jghao@fnal.gov

Introduction

Galaxy cluster plays an important role in modern Cosmology. Their abundance and spatial clustering provide an unique way to differentiate between the modified gravity models and dark energy plus General Relativity models.

Owing to the different physical interactions, galaxy clusters show different signatures at different wavelengths of the associated electromagnetic emission. Therefore, the detections of galaxy clusters have been investigated in multiple wavelength.

In optical band, the major challenge for cluster detection is the projection effects, i.e., the foreground and background galaxies projected into the cluster along the line of sight. We cannot resolve the position of galaxies along the line of sight mainly because most galaxies do not have their spectroscopic redshift available. Fortunately, the red galaxies in galaxy clusters are tightly clustered in colors, and we can detect galaxy clustering by detecting the red sequence clustering in color space.

Gaussian Mixture Model (GMM) [1] can be used for density estimation in large astronomical datasets [2]. When applying it to red sequence galaxy clustering, measurement errors of the colors need to be included [3] and a cluster finding algorithm, GMBCG, has been developed [4].

In this work, we future extend the error corrected Gaussian Mixture Model by including the spatial weights from cluster galaxies and field galaxies. This helps to boost the detection of galaxies clusters and reduce the false detection.

Color Distribution

Around a galaxy cluster, the color distribution of galaxies can be well approximated by a mixture of Gaussian Distributions [3]

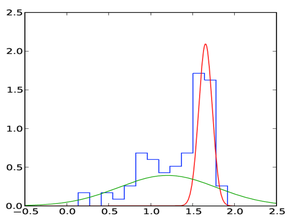
$$p(c_i, r_i|\theta) = \sum_{j=1}^{M} w_j \left[\frac{1}{\sqrt{2\pi\sigma_{cj}^2}} \exp\left[-\frac{(c-\mu_{cj})^2}{2\sigma_{cj}^2}\right] \right]$$

The measurement errors on the colors are not negligible and can be modeled as

$$p(c_i|c_i^{(m)}) = \frac{1}{\sqrt{2\pi\delta_i^2}} \exp\left[-\frac{(c_i - c_i^{(m)})^2}{2\delta_i^2}\right]$$

Putting them together, we have the Error Corrected Gaussian Mixture Model [3]

$$p(c^{(m)}, r|\theta) = \prod_{i=1}^{N}\sum_{j=1}^{M} w_j \left[\frac{1}{\sqrt{2\pi(\sigma_{cj}^2+\delta_i^2)}} \exp\left[-\frac{(c_i^{(m)}-\mu_{cj})^2}{2(\sigma_{cj}^2+\delta_i^2)}\right] \right]$$

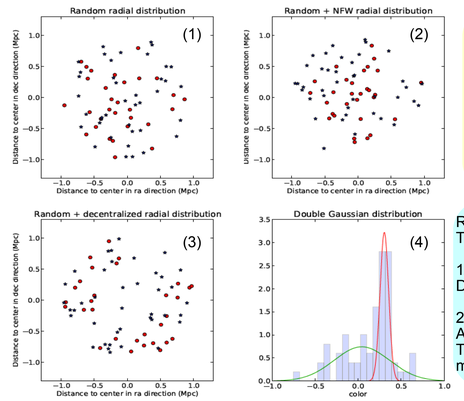


Based on detecting the bimodal feature of in color space around galaxy cluster, a large optical cluster catalog, GMBCG, for SDSS DR7 has been constructed [4]

Mock & Real data

What we are most interested in is how much the spatial distribution will help to reduce the false detection even if there are bimodal features in color space. This can be quantified as the difference between the BICs. The larger difference means a better detection.

Monte Carlo Simulation



Mock data set (1): spatial distribution is random

Mock data set (2): spatial distribution is from a composite of random and NFW distribution.

Mock data set (3): spatial distribution is random + a decentralized distribution

Colors in the above 3 data sets are all from a bi-Gaussian distribution as shown in (4)

Running our fitting to the three data sets, we have The following results:

1. Spatial weight can significantly increase the Difference between the BICs.

2. When the spatial distribution is very far from our Assumed distribution for cluster members, it will make The one- component fit outperforms the two component model

Simulation	BIC - one component	BIC - two component	Clustering Amplitude
Random Distribution	17.665	11.519	17.32
Random + NFW	50.093	19.077	24.68
Random + Decentralized	14.675	17.522	0.98

Spatial Distribution

Around a galaxy cluster, the spatial distribution of galaxies is a composite distribution of the cluster members and random field galaxies. The galaxy cluster members' distribution can be well approximated by the NFW distribution while the random field galaxies follow a spatial Poisson distribution.

PDF of cluster members' distribution

$$p(r|\rho_s(R_c), r_s) = \frac{2\rho_s r_s r}{(r/r_s)^2-1} \begin{cases} 1 - \frac{2}{\sqrt{(r/r_s)^2-1}} \tan^{-1}\sqrt{\frac{(r/r_s)-1}{(r/r_s)+1}} & (r/r_s) > 1 \\ 1 - \frac{2}{\sqrt{1-(r/r_s)^2}} \tanh^{-1}\sqrt{\frac{1-(r/r_s)}{(r/r_s)+1}} & (r/r_s) < 1 \\ 0 & (r/r_s) = 1 \\ 0 & (r/r_s) > 20. \end{cases}$$

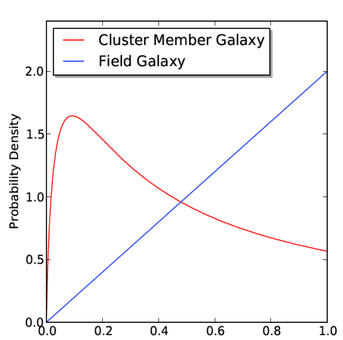
PDF of random field galaxies' distribution

$$p(r|R_c) = \frac{2r}{R_c^2}$$

Both PDFs are normalized at R_c

$$\int_0^{R_c} p(r|\rho_s, r_s)dr = 1 \qquad \int_0^{R_c} p(r|R_c)dr = 1$$

The Figure on the right shows the PDFs of the cluster member galaxies and field galaxies.



Likelihood & EM Recursion Relation

Combining the spatial and color distribution, we have the total likelihood

For random field galaxies only

$$p_1(c^{(m)}, r|\theta) = \prod_{i=1}^{N} \left[\frac{p(r_i|R_c)}{\sqrt{2\pi(\sigma_c^2+\delta_i^2)}} \exp\left[-\frac{(c_i^{(m)}-\mu_c)^2}{2(\sigma_c^2+\delta_i^2)}\right] \right]$$

For random field galaxies + cluster member galaxies

$$p_2(c^{(m)}, r|\theta) = \prod_{i=1}^{N} \left[\frac{w_1 p(r_i|\rho_s(R_c), r_s)}{\sqrt{2\pi(\sigma_{c1}^2+\delta_i^2)}} \exp\left[-\frac{(c_i^{(m)}-\mu_{c1})^2}{2(\sigma_{c1}^2+\delta_i^2)}\right] + \frac{w_2 p(r_i|R_c)}{\sqrt{2\pi(\sigma_{c2}^2+\delta_i^2)}} \exp\left[-\frac{(c_i^{(m)}-\mu_{c2})^2}{2(\sigma_{c2}^2+\delta_i^2)}\right] \right]$$

Expectation – Maximization recursive relation for Maximizing the likelihood

$$\begin{aligned} \mu_{cj}^{(t+1)} &= \frac{\sum_{i=1}^{N} c_i^{(m)} p(z_i = j|c_i^{(m)}, \theta_j^{(t)})/(1+\delta_i^2/\sigma_{cj}^{(t)2})}{\sum_{i=1}^{N} p(z_i = j|c_i^{(m)}, \theta_j^{(t)})/(1+\delta_i^2/\sigma_{cj}^{(t)2})} \\ \sigma_{cj}^{(t+1)} &= \left[\frac{\sum_{i=1}^{N} (c_i^{(m)}-\mu_{cj})^2 p(z_i = j|c_i^{(m)}, \theta_j^{(t)})/(1+\delta_i^2/\sigma_{cj}^{(t)2})}{\sum_{i=1}^{N} p(z_i = j|c_i^{(m)}, \theta_j^{(t)})/(1+\delta_i^2/\sigma_{cj}^{(t)2})} \right]^{1/2} \\ w_j^{(t+1)} &= \frac{1}{N}\sum_{i=1}^{N} p(z_i = j|c_i^{(m)}, \theta_j^{(t)}) \end{aligned}$$

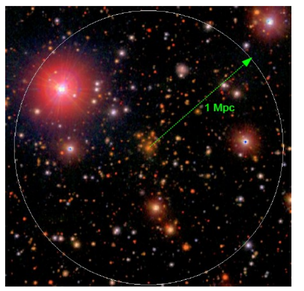
Bayesian Information Criterion (BIC) is used to determine the number of the mixture components.

$$p(z_i = j|c_i^{(m)}, \theta_j^{(t)}) = \frac{p(c_i^{(m)}|z_i = j, \theta^{(t)}) w_j^{(t)}}{\sum_{j=1}^{M} p(c_i^{(m)}|z_i = j, \theta^{(t)}) w_j^{(t)}}$$

$$p(c_i^{(m)}|z_i = j, \theta^{(t)}) = \frac{p_j(r_i)}{\sqrt{2\pi(\sigma_{cj}^2+\delta_i^2)}} \exp\left[-\frac{(c_i^{(m)}-\mu_{cj})^2}{2(\sigma_{cj}^2+\delta_i^2)}\right]$$

- **If there is a galaxy cluster, then the BIC corresponding to the two-component fit should be smaller.**
- **The bigger the difference between the two-component BIC and the one-component BIC, the more significant the detection will be.**

SDSS Coadded Data



Fitting without spatial weights:

- BIC1 = 72.75
- BIC2= 57.49

Fitting with spatial weights:

- BIC1 = 113.06
- BIC2=51.39

Including the spatial weights boosts the significance of the cluster detection.

Discussion and Future Work

The spatially weighted and measurement error corrected GMM developed in this poster provides a full description of the cluster galaxy distribution in radial and color space. It gives a consistent clustering amplitude estimate w.r.t. the assumed spatial profile and the data-driven color distribution. The spatial weights will help to reduce the false detection of clusters by penalizing the clustering amplitudes of those cluster candidates whose spatial distributions are very different from the known cluster galaxy profile, though they happen to show the bimodal feature in color space. The corresponding codes can be downloaded from [5].

We have an ongoing effort to implement the techniques mentioned here as a cluster finding algorithm and will apply it to the incoming Dark Energy Survey[6].

REFERENCES

[1] D. Titterington, et al, 1985, (John Wiley & Sons)
[2] A. Connolly, et al, 2000
[3] J. Hao, et al, Astrophys. J. 702,745 (2009)
[4] J. Hao et al, Astrophys. J. Suppl. 191, 254 (2010)
[5] https://sites.google.com/site/jiangangecgmm/
[6] http://www.darkenergysurvey.org

Funding for the SDSS and SDSS-II has been provided by the Alfred P.Sloan Foundation, the Participating institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/.The SDSS is managed by the Astrophysical Research Consortium for the Participating institutions. The Participating institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint institute for Nuclear Astrophysics,the Kavli institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-institute for Astronomy (MPIA), the Max-Planck-institute for Astrophysics (MPA),New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.