Predicting the Masses of Galaxy Clusters Using Weighted Linear Regression

James Eby Advisor: Peter Freeman

Background

Galaxy clusters are the largest gravitationally bound structures in the Universe. Each can contain hundreds of galaxies, which move at high velocities under the influence of their neighbors.. In this project, we follow Ntampaka et al. (2014) and analyze data from simulated galaxy clusters to calibrate the relationship between the sample standard deviation of cluster galaxy velocities, or their velocity dispersion, and cluster mass. Velocity dispersions are observable for real clusters, while masses are not. To increase sample size, Ntampaka et al. measure velocity dispersion along several lines-of-sight through each cluster, regressing velocity dispersion onto mass ($\log_{10}(\sigma) = \beta_0 + \beta_1 \log_{10} M$). However, in their analyses they ignore the correlations between these measurements. Our goal is to determine if ignoring these correlations leads to significant biases in model parameter estimation.



Figure 1: Halo in Multidark simulation; bright regions are subhalos, and center region lies at the core of this cluster

Data

The data that we analyze come from the publicly available Multidark simulation (www.cosmosim.org), an N-body simulation that follows the evolution of 3840^3 dark matter particles. Over time, dark matter can group together in halos, regions whose average density is at least 200 times that of the Universe as a whole. Ntampaka et al. (2014) selected halos of mass greater than 1.47 x 10^{14} solar masses as their cluster sample, while retaining only those clusters with at least 20 subhalos (or individual cluster galaxies) of mass at least 1.47 x 10^{12} solar masses each. Our sample contains 2,514 distinct clusters. Each galaxy moves within its cluster, so given a line-of-sight through the cluster, the projected line-of-sight velocity for each galaxy can be determined, as can the overall velocity dispersion. Thus, for each line-of-sight through each cluster, there are two measurements: mass (the predictor variable) and velocity dispersion (the response variable).

The Correlation Between Lines of Sight

To get an idea of the correlation between such lines of sight, we ran a 2D simulation of random lines of sight and plotted a histogram of the correlation values between pairs of lines of sight. The plot shows that there is a high correlation between lines of sight of the same cluster. These results show that there can be high correlation between lines-of-sight through the same cluster, motivating us to try a weighted linear regression approach.

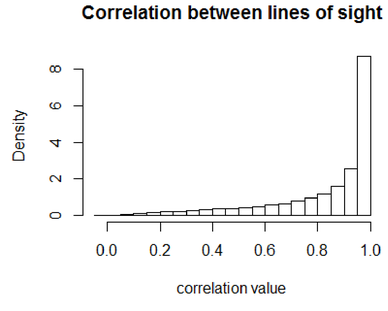


Figure 2: . The plot above corresponds to a 20,000-element cluster and 2,000 lines of sight of the same cluster.

Weighted Linear Regression

We wish to solve for β in the following equation:

$$\hat{\beta} = \left(X^T \Sigma^{-1} X\right)^{-1} X^T \Sigma^{-1} Y$$

$$Y = \begin{pmatrix} \log_{10}\sigma_1 \\ \log_{10}\sigma_2 \\ \vdots \\ \log_{10}\sigma_N \end{pmatrix} \quad X = \begin{pmatrix} 1 & \log_{10} M_1 \\ 1 & \log_{10} M_2 \\ \vdots & \vdots \\ 1 & \log_{10} M_N \end{pmatrix}$$

where X is the design matrix, Σ is the 24451x24451 block-diagonal covariance matrix, and Y is the vector of response variables (values $\log_{10}(\sigma)$ for each line-of-sight) . X is a 24,451 x 2 matrix, with the first column being 1's and the second column containing the values of $\log_{10}$ M. Σ is a block-diagonal matrix, where each block represents the covariance matrix for a single cluster. The covariance of lines of sight i and j within a single cluster was estimated by:

$$\frac{n}{n-1}\{E[V_i * V_j] - E[V_i] * E[V_j]\}$$ Note: we use V_i as a substitute for $\log_{10}(\sigma_i)$

Since we cannot do weighted linear regression on the whole sample due to the size of the covariance matrix, we explore two options for testing: (1) selecting subsets of clusters to analyze, and (2) selecting single lines-of-sight randomly through each of the 2,514 clusters. The first takes covariance explicitly into account, while the second ignores covariance by taking only one line-of-sight per cluster.

Cluster Subset Sample

To calculate the intercept (β_0) and slope values (β_1) of the regression , we took a subset of the clusters in our simulation and calculated the slope and intercepts of our linear regression.

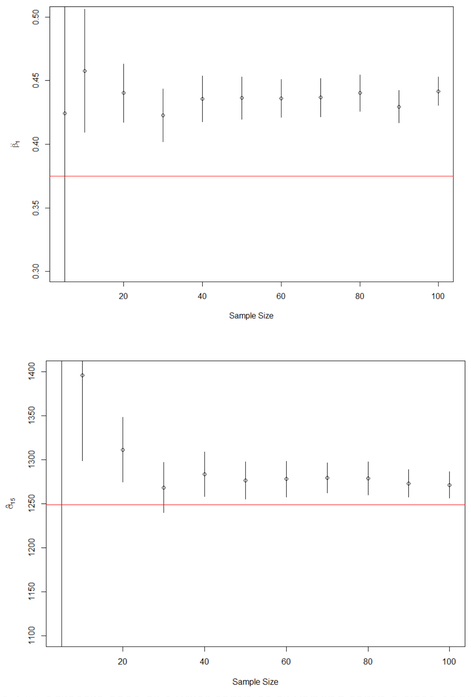


Figure 3 : We simulated 100 datasets each of cluster samples at varying sample sizes between 5 and 100 clusters. The red lines mark the value calculated by Ntampaka Et Al (β_1=.375, σ_{15} = 1249 km/s), and we provide an interval of 95% significance to illustrate the difference between values estimated using weighted linear regression as opposed to simple linear regression. We only have a single value for velocity in each line of sight, so we estimate $\log_{10}(\sigma_{Vi})$ by using V_i. This allows us to determine whether or not taking correlation into account significantly changes the regression coefficients that one calculates. Note that $\sigma_{15} = \frac{1}{100}\sum_{i=1}^{100} 10^{\beta_{0i}+15*\beta_{1i}}$

Single Line of Sight

In this simulation, we selected a single random line-of-sight through each of the 2,514 clusters and ran unweighted linear regression.

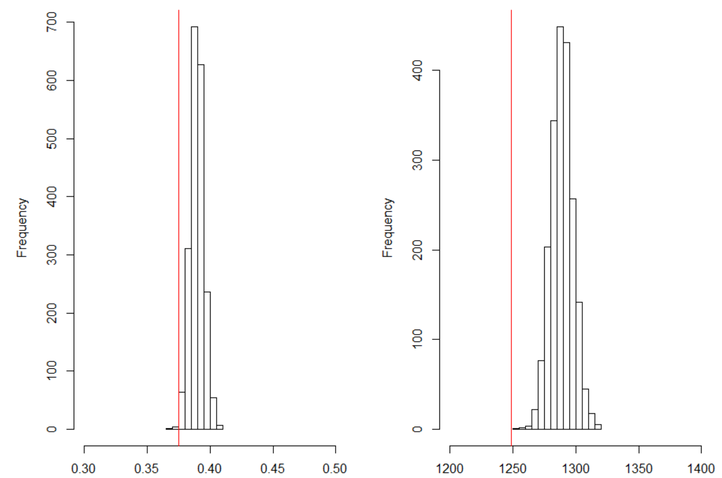


Figure 4: 2000 repetitions show significantly higher β_1 and σ_{15} values than those the estimates in Ntampaka et al shown in red (β_1 = .375 and σ_{15} = 1249 km/s). Mean and standard error for our estimate of β_1 : .38961 and .00012. Mean and standard error for our estimate of σ_{15} : 1288.869 and .19573

Conclusions

The values of β_1 and σ_{15} that we calculate are significantly different from those obtained in Ntampaka et al. Using weighted linear regression to account for the correlation between lines of sight gives a larger estimate for each regression coefficient. Although our single line of sight analysis resulted in larger estimates for both coefficients as well, the values were slightly different, which is likely due to our imperfect estimate of covariance.

References

Ntampaka, M., et al. 2014, arXiv:1410.0686

http://hipacc.ucsc.edu/Bolshoi/

CARNEGIE MELLON UNIVERSITY PITTSBURGH PENNSYLVANIA 1900

Department of Statistics