Detecting Dark Matter Halos

Sam Beder, Brie Bunge, Adriana Diakite

December 14, 2012

Introduction

Dark matter's gravitational pull affects the positions of galaxies in the universe. If galaxies were circular, detecting the locations of dark matter halos would be directly based on the ellipticity of the surrounding galaxies. However, galaxies are inherently elliptical, and this property is random. Our challenge is to account for this randomness when finding the center of dark matter halos. Our data set includes simulated skies with one, two, or three halos and our task was to find their centers. We found that using the sum of the tangential ellipticities[1] of surrounding galaxies was the most salient signal for determining halo locations. While this can be directly determined for one-halo skies, two and three-halo skies required more sophisticated methods due to interactions between multiple halos.

Data

Kaggle provides all of the data for this problem, which consists of 300 simulated skies with 300-720 galaxies in each. Each galaxy has an x, y coordinate, and ellipticity components, e1 and e2. Each sky has a given number of one to three halos.

Proposed Methods

The following initial methods were overall unsuccessful, but were essential in helping us gain an understanding of our data and for leading us towards our eventual proposed solution:

Histograms of Angle and Magnitude

Description for Single Histogram Approach

In this method we divide the sky into m rectangles $\{x_1, x_2, \dots, x_m\}$. For each bin/ training example we built a k x k histogram of angles by magnitudes (both calculated from ellipticity). With this representation of the data, we created our design matrix X such that each row of the matrix is a training example and each column is an angle/magnitude bucket of the histogram corresponding to that training example.

Results

This initial method was a largely unsuccessful strategy for identifying our cluster locations. We found that the bin containing the halo did not have very indicative features since we would also had very few galaxies per bin. Furthermore, we suspected that a bin might actually be better at indicating halos in adjacent bins since the galaxies would essentially point in that direction. This insight inspired us to use a multi-histogram approach to try to identify the general location of the halos.

Description for Multi-Histogram Approach

We modified our design matrix X such that each row/training example had nine histograms concatenated together (the histogram for the rectangle of interest and the histograms of the eight surrounding rectangles).

Results

[1] The ellipticity of a galaxy at a position (x,y) tangential to a point (x',y') is $e_{tangential}=-(e_1\cos(2\phi)+e_2\sin(2\phi))$, where the angle of the galaxy with respect to the dark matter center is given by $\phi = \text{atan}((y-y')/(x-x'))$.

Unfortunately, this approach also performed disappointingly. Using an SVM on our data yielded a model that predicted that no halos were present in our test skies. Therefore with 100 bins we had at least a 97% accuracy rate on training skies due to our skewed data, but this result is clearly not helpful because it gives us no indication of the halo location, and our model's least sure negative predictions were only rarely the correct positive predictions.

Conclusion

The evidence reveals the largest difficulty in analyzing our data - the massive amount of noise in the ellipticity of our galaxies. The effect of a halo on its surrounding galaxies seems to be so slight that using a subset of our galaxies would make the effects statistically insignificant. We therefore decided to pursue more holistic approaches that attempt to get rid of the noise of individual galaxies.

Clustering Galaxy Positions with K-means

Description

The general strategy of this approach is to use clusters in an attempt to get rid of the noise of individual galaxies. This approach can be divided into three steps:

1. Cluster the galaxies into k clusters based on their x,y position in the 2D sky
2. Calculate the average ellipticity and magnitude of each cluster
3. Weight the position of each cluster based on its average ellipticity/magnitude to calculate a predicted halo position

The number of clusters used significantly affected the success of our guesses. With too few clusters (<10) the average ellipticity of clusters were telling, but we would not have enough clusters to make an accurate guess. With too many clusters (>40) the large amount of noise from galaxies was not averaged out so making accurate guesses from these results proved difficult. We found a medium amount of clusters (~14-25) gave us the best results.

Results

We were able to generate features by using a fixed number of clusters and using our various average cluster centroids, average ellipticities, and average magnitudes to predict the halo location. Because k-means clustering can generate different results depending on the cluster starting location, we decided that we should generate multiple guesses and average our results. We did 20 iterations on each of the 100 skies (we applied this method to skies with only one halo). And found that on average we were about 755 pixels away from the actual location of the halo (the sky is 4200 x 4200 pixels).

Conclusion

This approach is able to more accurately guess on some skies, but fails significantly on skies with halos to the sides of the sky and with multi-halo skies. Although this approach got rid of some of the randomness of galaxy ellipticity, it does not give features that can accurately predict halo location.

Reflection on previous methods

Our failed attempts pushed us to approach the problem from a different perspective. Instead of looking at individual or clusters of galaxies and seeing where they guess a halo is, we considered a specific location and assigned it a likelihood of being the position of the dark matter halo center. The best indicator was tangential ellipticity.

Final Methods

Subtraction Approach

In this approach, we define our signal of the most likely halo location to be the sum of the tangential ellipticities with respect to that location. If there is only one halo, we take the location with the strongest signal. If there are multiple halos, the effect of the prior halos is subtracted out before determining the strongest signal (this is possible because tangential ellipticities from different halos add linearly).

To determine the signal, we divide the sky up into bins. For each bin, we calculate the tangential ellipticity. The center of the bin with the highest tangential ellipticity is set as the initial point for a smoothed annealing algorithm that refines the peak signal location. The result becomes the guess for the dark matter halo. On average, our guesses for the initial halo position in all skies were 144.92 pixels from the actual position. This high accuracy is due to one halo almost always dominating the sky with its effects, while the remaining less massive halos became harder to locate precisely.

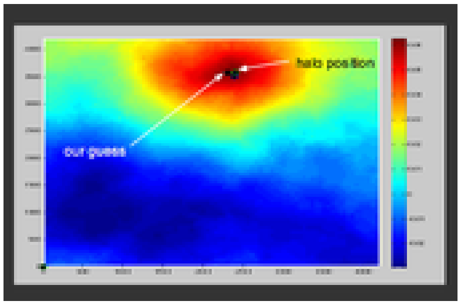
To determine how to subtract the effect of a halo, we model the force of a halo on the ellipticities of surrounding galaxies as $\frac{a}{r^b}$, where r is the radial distance to a halo. Since the cross-component of the tangential ellipticity is zero, we can find this force in terms of the ellipticity components, e1 and e2, and subtract it from the e1 and e2 of each galaxy. The variables a and b are learned by finding the best fit between the force's e1 and e2 and the actual sky's e1 and e2.

For two-halo skies, we locate the first halo. Then, we calculate the 'best' parameters a_1, b_1 and a_2, b_2 to eliminate the first and second halos, respectively. The 'best' parameters are those that eliminate both halos in such a way that the resulting sky appears to have no halos. This happens in two layers of optimization. The top level optimizes a_1, b_1. For each iteration, a_1, b_1 are used to subtract out the effect of the first halo. The highest seven resulting signal peaks are chosen as potential second halo locations. For each of these choices, an inner optimization layer tries to find the optimal a_2, b_2. After optimizing on both layers, the second halo location with the minimum cost becomes the second halo guess. This two-layer optimization performed better than finding the first halo location, subtracting its effect based on a and b learned from the one-halo skies in the dataset, and choosing the second halo location based on the highest signal from the resulting sky.

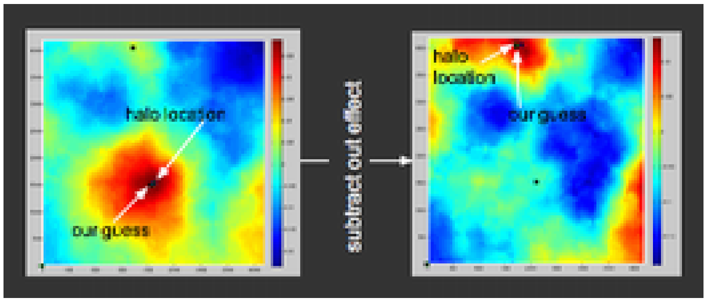
For three-halo skies, we built on the two-halo procedure with a third layer of optimization. We find the optimal choice of a_3 , b_3 which depends on the third halo location, which comes from the optimal choice of a_2 , b_2 which in turn depends on the second halo location, which is ultimately based on our optimal a_1 , b_1. This strategy worked successfully, but was far too time consuming, so for our final results we subtracted the first halo's effect using the result of a neural network learned on the one-halo training data. This method was very successful and efficient on most of our training data. On average, our proposed halo locations were off by 753 pixels from the actual halo location.

Below, we see this method successfully identifying halos in skies with one, two, and three halos using the methods described above:

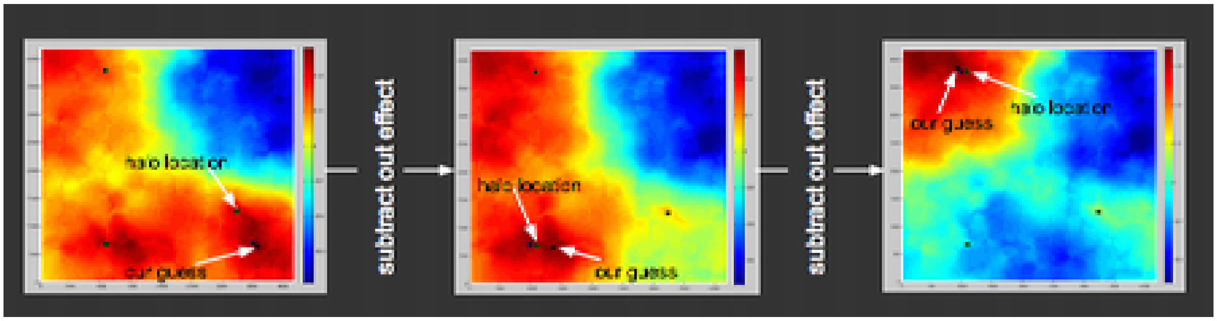
One-Halo Sky



Two-Halo Sky



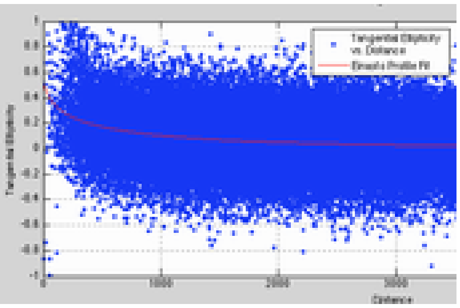
Three Halo Sky



We tried a variety of methods to improve our results. The galaxies' ellipticities almost perfectly fit a gaussian distribution with sigma 0.22, so we tried gaussian filtering to help get rid of some of the noise in the data. We also tried restricting our calculations of tangential ellipticity to a certain radius around the point of the proposed halo to cope with interfering effects. Neither of these improved our results; we believe that this is because it ended up losing more information than it clarified.

Modeling Approach

This approach assumes an Einasto profile[2] to model tangential ellipticity as a function of distance. It was suggested that this equation might be a good fit (Chivers). We confirmed this by fitting the equation, and learning its coefficients, from the tangential ellipticity and distance of one-halo skies; the fit is shown in the figure to the right. We used this model to determine what a noise-free sky would look like with halos in certain locations. The optimization objective is to find halo locations that result in a sky model that has the least-mean-squared difference from an actual sky.

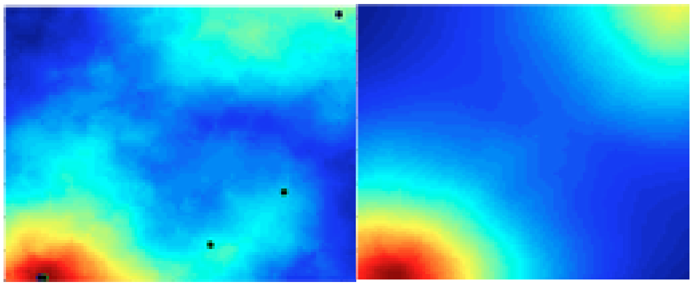


For one halo skies, we find the halo location via the same procedure as in the subtraction approach. For two halo skies, we find the first halo in the same way and incorporate its effect into the model for that sky. Then, we optimize for the second halo location using a multi-start minimization algorithm to find the best sky model. We determine the starting locations by subtracting out the effect of the first halo and creating a matrix that represents the signal at the centers of cells in a 15 by 15 grid of the sky. The top ten peaks are used as starting locations. Choosing the top peaks, rather than the top bins, prevents starting locations from all being clustered together in particularly high-signal areas. The procedure for three halo skies builds on

[2] The equation for the Einasto profile is a*exp(-b*r^c), where r is radial distance ("Einasto Profile").

this by optimizing instead on the locations of both the second and third halos. We determine the starting locations for the third halo by subtracting out the effect for each of the second halo start points.

The resulting sky models reflected the sky well, as seen below (the actual sky is on the left and our model is on the right). However, there were other cases that satisfied the goal of the algorithm, but resulted in halo location guesses that were in corners or on edges.



Conclusion and Future Work

We are proud of our algorithm's performance, but see further work that can potentially improve our results. For example, we began exploring the use of neural nets to learn different ways of subtracting halos based on whether it was dominant, secondary, or minor in the sky. So far this approach showed promise on a couple of difficult skies, but we were not able to get it fully functional.

We also observed that particular algorithms worked better on certain skies than others, so it would be beneficial to use different algorithms depending on the situation. As a result, we believe that we could improve our performance by running several of our most successful algorithms, and then have each of them evaluate their confidence based on their mean squared difference, an evaluation of the sky, and their prediction locations. Then we could use a voting system to choose the predictions of the most confident algorithms.

We also hoped to apply more physics to apply the true lensing effect that dark matter halos have on surrounding galaxies instead of our simplified version of the lensing equations. We should be careful of using this method because the simulated skies in the dataset also do not use the true physics, but to make this program truly applicable in finding real dark matter halos, incorporating this physics would be a necessary step.

References:

Chivers, Corey. "Simulating Weak Gravitational Lensing)." *Bayesianbiologist*. N.p., 24 Nov. 2012. Web. 14 Dec. 2012. <http://bayesianbiologist.com/2012/11/24/simulating-weak-gravitational-lensing/>.

Harvey, David. "No Free Hunch." *Observing Dark Worlds: A Beginners Guide to Dark Matter & How to Find It*. N.p., 12 Oct. 2012. Web. 16 Nov. 2012. <http://blog.kaggle.com/2012/10/12/observing-dark-worlds-a-beginners-guide-to-dark-matter-how-to-find-it/>.

"Einasto Profile." *Wikipedia*. Wikimedia Foundation, 11 Dec. 2012. Web. 14 Dec. 2012. <http://en.wikipedia.org/wiki/Einasto_profile>.