Automatic Detection of Dark Matter via Weak Gravitational Lensing

Thomas Colvin
tcolvin@stanford.edu

Francisco Capristan
fcaprist@stanford.edu

1 Introduction

Dark Matter is an elusive form of matter that makes up approximately 25% of the known universe. Since it does not absorb or reflect light, it can never be detected via direct observation with a telescope. Instead, its existence must be inferred from the effect of its gravitational field on the background stars / galaxies. If a source of Dark Matter lies between an observer on Earth and a field of background stars, the Dark Matter's gravitational field will distort the light from the background stars to change the shape and distribution of the stars that the observer sees from their true distribution. This effect is called "gravitational lensing" and it can be split into two regimes: weak and strong. The effects of strong lensing are somewhat obvious because the distortions produced are very pronounced, i.e. affected galaxies will appear to be bent into long arcs, but this regime is a relatively rare occurrence. Weak lensing, however, produces distortions that are difficult to separate from the background "noise" of the universe (i.e. the size / ellipticity distributions of galaxies) and measurement error (i.e. thermal distortions in telescope mirrors, effects of atmosphere, etc.). This is the predominant regime for observations of the universe and automatic detection of weak lensing effects is a relatively new field of study.

2 Training Data and Signal

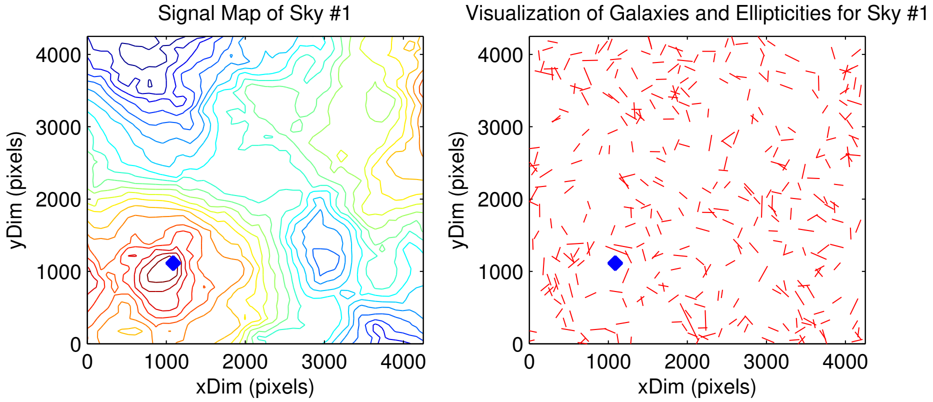

Figure 1: The figure at left shows a signal map for the sky. Notice that the area of maximum signal corresponds closely to the location of the dark matter source when there is only a single source present. The figure at right shows the galaxies in a single sky with their ellipticities and angles, measured by the line length and angle from the x-axis respectively.

The training data was taken from the data-mining-competition website Kaggle.com . The data consisted of 300 skies, each containing anywhere from 300 to 900 different galaxies specified by a position (x,y) in the sky and an "ellipticity" vector (e1,e2) which specifies its orientation. Each sky contains at least one and up to three sources of dark matter, which distort the images of the galaxies through weak gravitational lensing. Specifically, weak lensing distorts the image of a lensed galaxy by amplifying the galaxy's perceived size and by rotating such that it lies more tangent to the dark matter source. We define the tangential ellipticity of a galaxy at a position (x,y) tangential to a point (x',y') in equation 1 where ϕ is the angle from the point to the galaxy defined by equation 2.

$$e_{tangential} = -(e_1 \cos 2\phi + e_2 \sin 2\phi) \tag{1}$$

$$\phi = \arctan\left(\frac{y - y'}{x - x'}\right) \tag{2}$$

This motivates the idea of a tangential ellipticity ”signal”, i.e. a measure of the tangential ellipticity of galaxies in the sky with respect to a point, which is simply the sum of the tangential ellipticity of galaxies around the point (x',y'). High signal strength tends to be a good indicator for the location of the dominant dark matter source. The signal map could sum over all galaxies present in the sky or only sum over a smaller number of galaxies that lie near the point of interest. An example of a signal map is seen in Figures 1 and 4.

3 Methods Attempted

3.1 Logistic Regression

For this technique we attempted to cast the problem as a classification problem by discretizing the sky and asking ”does this location contain the peak of the dark matter?” The motivation for this approach was the difficulty in creating feature vectors that were meaningful across each sky using the locations and ellipticities of each galaxy . For instance, how can we ensure that the nth element of the feature vector will be meaningful across the different skies when 1.) there is no obvious ordering by which to rank the galaxies within a sky and 2.) the number of galaxies per sky varies from 300-900 so some galaxies will have to be omitted? Discretizing the sky solves this problem because nth element of the feature vector will always correspond to a specific region of the sky. Using a discretized sky we tried a few different logistic regression schemes.

3.1.1 Signal Response Surfaces

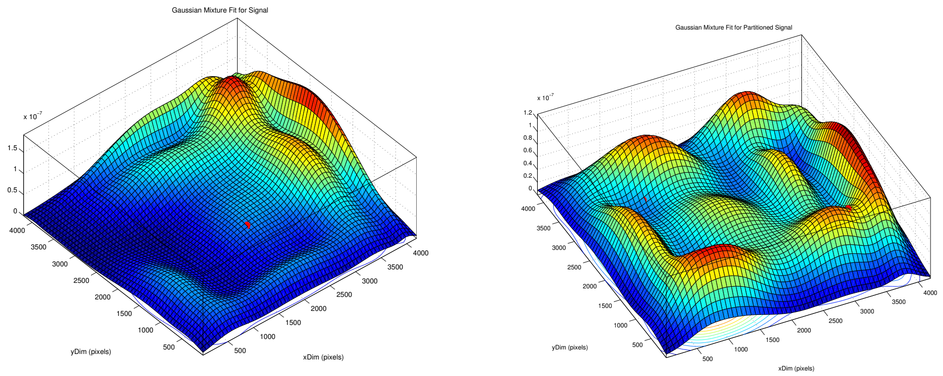

Figure 2: Left: Signal Gaussian Response Surface for Entire Sky. Right: Normalized Signal Gaussian Response Surface for Partitioned Sky.

The proposed approach involves the use of response surfaces fit to regular signal maps and modified signal maps that amplify any possible halo effect. These response surfaces are then used to create a feature vector. Since the skies could have 1-3 halos, multiple local maxima are possible. Derivative information can provide valuable insight in finding the coordinates of any local maxima, and thus could be useful in capturing some halo effects. Signal maps are not necessary smooth functions, but

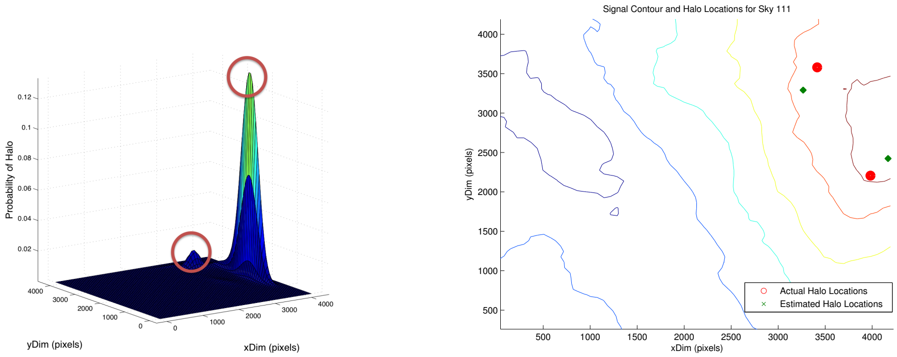

Figure 3: Left: Shows the probability of finding a halo at a given coordinate. Right: Comparison of estimated and actual halo locations

they have a tendency to have a behavior similar to that of a sum of gaussian distributions. Therefore, in order to get smooth derivative values, a response surface of the total signal (sum over all galaxies in the sky) is created by using a mixture of gaussians.

To amplify the effects of non-dominant sources of dark matter we divided the sky into 16 equal partitions and locally calculated a signal map for each partition. The signal in each partition was normalized such that the maximum signal value is the same in each of the different partition. This has the effect of amplifying other local minima (see Figure 2 on right), and thus increasing the chances of recognizing the effects of weak dark matter sources. The number of divisions was selected to ensure that a significant number of galaxies are in each partition. Derivative information is desired in this step as well, so a gaussian response surface is generated for the entire sky by using the locally normalized signal values at each partition. In theory, by using locally normalized signal maps we should be able to capture and amplify local tangential ellipticity effects due to the presence of dark matter. In practice, however, by amplifying the signal we are also amplifying the noise, thus it is expected that logistic regression will return some false positives.

To create the feature vector for logistic regression the sky is gridded into smaller cells (80x80) and the properties of the signal in each cell become elements of the vector. The feature vector thus contains values for the full signal map along with its derivatives and the locally amplified signal map with its derivatives as well. All quantities are calculated at the center of the sky cells. The response variable equaled 1 if the cell's center is located within a 3-cell radius of a halo and zero otherwise, thus a single source of dark matter will produce a 7x7 cell region in the sky that counts as a positive classification.

We trained the algorithm with 200 skies (over 1 million cells). As expected, the signal from most of the less dominant halos was amplified, but we underestimated the effects from noise amplification. Figure 3 shows a sky prediction where the noise was not dominant. After looking at several cases, we determined that the noise amplification considerably affects the logistic regression by generating a high number of false positives. Overall, few skies had acceptable halo location predictions, but in most cases the results suggests that this particular method is not appropriate to find halo locations. Other variations for this method were implemented and the results were unsatisfactory.

3.1.2 Sliding Window

We gridded the sky into a 30x30 mesh of 900 boxes and for every box asked ”is this the peak of the dark matter”? Construction of the feature vector was confounded by the discretization (things get sparse since our data is already at discrete positions) and the size of our training data and feature vector lengths are time-constrained by Newton's method.

Within the gridded sky, we construct a feature vector by looking at 3x3 groups of boxes and generate many such feature vectors per sky by sweeping the 3x3 window across the sky. The signal in each sub-box was used as an element in the feature vector and the associated response variable equaled 1 if the dark matter peak is contained within the center box, zero otherwise.

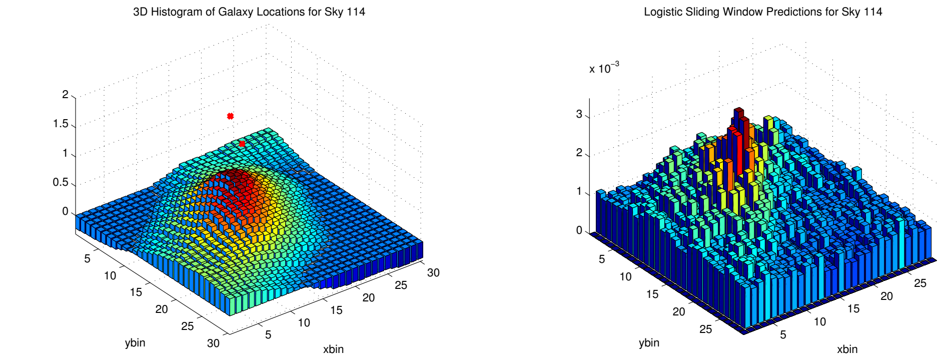

Figure 4: Left: Signal map of discretized sky where floating red dots indicate the locations of dark matter that we are trying to predict. Right: 3D histogram of probabilities for dark matter locations calculated from Logistic Sliding Window.

Figure 4 shows the predicted probabilities for dark matter locations in a sky that was not trained on. Notice that there is a lot of noise which makes it difficult to make a confident prediction. Also notice that we only seem to be locating the dominant source; the minor source does not even appear to be visible in the signal map on the left. In general, training on skies that only have a single source of dark matter seems to give better results for multiple halo testing, as opposed to training on multiple-halo skies. This is likely due to a more straightforward relationship between signal and dark matter location in single-halo skies.

3.2 Partial Least Squares Regression (PLS)

PLS regression is used to find the fundamental relations between two matrices by investigating the covariance structures in these two spaces. Specifically, a PLS model looks for the multidimensional direction in the feature-space that explains the maximum variance direction in the response-space. PLS regression is generally useful when you have more features than observations, which seems to be our case since we have only 300 training skies. This is the only method that we have seen which allows a regression on multiple dependent variables, i.e. y $\in \mathbb{R}^n$.

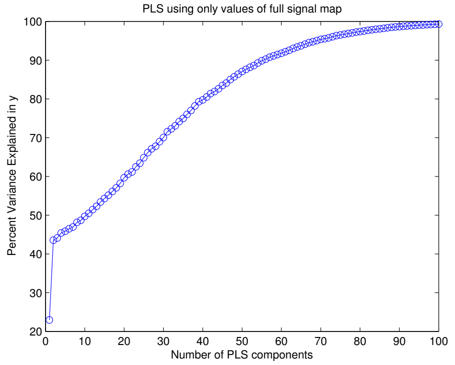

Figure 5: Typical plot showing number of PLS components required to capture the variance in the response-space.

Three types of feature vectors were considered. Type 1: We construct a feature vector using only the full signal evaluated at the discretized sky grid points, as was previously seen in the Logistic Regression approach. Type 2: We construct a feature vector using the signal in each grid cell as calculated from the nearest N = 25:25:300 galaxies (Matlab notation) as well as the signal calculated from all the galaxies in the sky. Type 3: We use the full signal from all galaxies as well as the average ellipticity, average orientation angle, and number of galaxies in each grid cell. For all types of feature vector, we used a 1x6 vector response variable that contained the grid locations for the dark matter sources, sorted by signal strength at their location. This ensures that the first two elements of the response vector always correspond to the ”strongest” dark matter source and so on.

Unfortunately, this method did not yield good results. Training on 270 of the skies and testing on the remaining 30, no feature vector scheme worked well with all of them performing similarly poorly. Type 1 does a decent job at locating dark matter in skies with a single source, but for the case of multiple sources it does poorly. The rationale behind using local calculations of the signal in Type 2 was that the perturbing influence of dark matter falls off with distance from the source so the galaxies closest to the source should feel the largest lensing effect. While this is conceptually true, in reality the noise associated with calculating the signal due to a small number of galaxies can easily overwhelm the signal due to a non-dominant source of dark matter. The hope was that by using the values calculated for many values of N, the learning algorithm could use these multiple snapshots to discern some patterns from the noise.

Type 3 suffers from the problem that as your grid gets finer, you find that most of your galaxy features are zero. For instance, if you discretize a sky containing 300 galaxies into a 30x30 grid, you will find that at least 600 of your grid cells contain zero galaxies. Doing this for 270 skies (a very small number compared to the feature vector length) produces a training set where the feature vectors themselves are very high dimensional and linearly independent. Doing a PLS regression in this case only gives a relatively small number of components (20) needed to capture all of the variance in the output. When looking at the results for the data we trained on, you can see that we are spot on for every dark matter source, but if you test on untrained data, the guesses are wild; we have severely overfit the data.

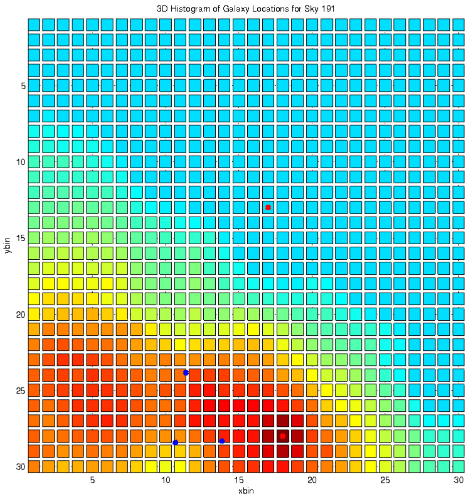

Figure 6: Example of a sky where PLS Regression predicts locations that are obviously misplaced. Red dots are true locations, blue dots are predicted locations.

3.3 Conclusions

In addition to the methods outlined above, we also explored the use of neural networks and spent a few days attempting to model the effects of distributed sources of dark matter and their weak lensing effects. These two methods failed to produce results and due to space limitations for this final writeup, are not discussed in any detail. All of the methods that we attempted failed to adequately predict the locations of dark matter in skies that had more than one source, however our solutions performed better than randomly guessing the dark matter locations. The specific failures of each method are outlined in their respective sections above.

By creating many random skies containing 600 galaxies each (with no dark matter present) and looking at their signal maps we noticed that the background noise was generally less than a factor of two less than the peak signal found from our training skies. Thus, with background noise so large, using signal strength will never be enough to properly locate minor sources of dark matter.

Through the course of this project, we became convinced that while some machine learning techniques may be helpful in solving this problem, the most appropriate way to solve this problem is to develop a physical model that connects a distribution of dark matter to its distorting effect on the background galaxies and then fit that model to the available data. With this model in hand and a testing sky, the inverse problem can be approximately solved, i.e. the distribution of the dark matter density should be optimized such that by undoing the effect of weak lensing on the testing skies, the background galaxies are seen to be distributed uniformly. This problem then becomes a matter of developing an appropriate model for weak gravitational lensing.

References

[1] Herve Abdi Partial Least Square Regression: PLS-Regression

[2] Bartelmann, Matthias, and Peter Schneider. ”Weak gravitational lensing.” Physics Reports 340.4 (2001): 291-472.

[3] All data and competition background provided by http://www.kaggle.com/c/DarkWorlds/