Dark Matter Halo Detection in Weak Lensing Regimes

Matt Anderson
Stanford University
andersme@stanford.edu

Phil Chen
Stanford University
pcchen@stanford.edu

Dustin Janatpour
Stanford University
dustinj@stanford.edu

INTRODUCTION

The Observing Dark Worlds Competition found on Kaggle, a website aggregating machine learning competitions, posed the following challenge: Given an image of some fraction of a sky with information about the location and observed ellipticities of hundreds of galaxies, can we predict the location of dark matter halos in that sky?

Although current methods exist to predict the location of these centers given comprehensive amounts of data such as three-dimensional position, mass, etc., there are not many approaches to finding the centers given two-dimensional images of a sky. However, for many regions of space, the only data collected are two-dimensional images, and thus the dark matter structures in those regions of space are as of yet unknown.

From only two-dimensional images, dark matter centers can be detected by the ellipticity distortion they cause on each galaxy. However, the incomplete and inherently noisy nature of the data complicates the problem. This distortion cannot be observed directly, especially under weak lensing assumptions, as a galaxy's natural ellipticity is drawn from a complex, unknown random distribution. In addition, these background ellipticities tend to be large relative to the dark matter signal, especially for galaxies far from the dark matter. Only by considering every galaxy and enforcing the assumption that average natural ellipticity is close to zero can we begin to look for signs of dark matter halos.

After ruling out many of the methods discussed in the course, we adopted several distinct approaches to the problem. The most successful was a variation on the k-means algorithm that used a modified clustering update rule and scoring mechanism for clusters that selected the best candidates. An alternative we considered was applying a batch gradient descent to an objective function formulated directly from the physical model. However, the complexity of this model prevented this method from producing global minima, leading to poor results.

CHI 2009, April 4 - 9, 2009, Boston, Massachusetts, USA.

DATA AND EVALUATION

We are given both training data and test data. The data consists of several hundred skies (three hundred for the training set, one hundred and twenty for the test set), each comprised of three hundred to seven hundred and twenty galaxy observations. Additionally, the training data contains the locations of true halos.

Each of the galaxies is specified by its (x,y)-position and its ellipticities e_1 and e_2. e_1 corresponds to elongation along the x and y axes. Positive e_1 corresponds to a horizontally elongated galaxy, and negative e_1 corresponds to a vertically elongated galaxy. e_2 corresponds to elongation along the axes 45 degrees from the x-axis. Positive e_2 corresponds to elongation along the 45 degree axis. Negative e_2 corresponds to elongation along the negative 45 degree axis.

The quality of a set of predictions is measured by a combination of distance error, the average radial distance between the predicted halo to the true halo, and positional bias, the average angle between a prediction and the line generated by a reference point and the true halo.

Benchmarks

The competition provides four benchmarks with publicly available algorithms, described briefly below.

- Lenstool Maximum Likelihood - Uses Lenstool software to estimate maximum likelihood for the halo positions.
- Gridded Signal Benchmark - Divides the sky into a grid. For each grid tile, this algorithm calculates the average signal (sum of tangential ellipticities) for that tile relative to the tile center, then outputs grid tiles with the highest signal.
- Randomly Placing Halos - Randomly places halos in the sky.
- Single Halo Maximum Likelihood - Divides the sky into a grid. For each grid tile, calculates the maximum likelihood that a halo will be at the center of the grid tile. Outputs the single most likely halo.

ALGORITHMS

Variations of K-Means Clustering

In attempting to tackle this problem, we considered many algorithms. We believed that the problem lent itself to the K-means clustering algorithm in the sense that cluster centers

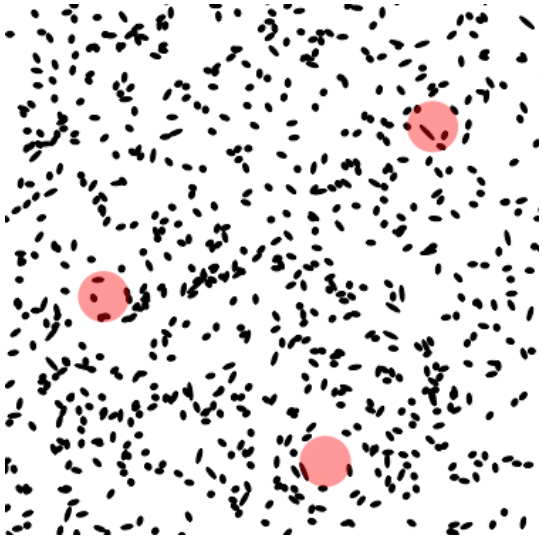

Figure 1. An example sky with the halos being shown

might somehow be correlated to halos. Since halos typically affect the set of nearby nodes, we examined how we might vary the K-means clustering to achieve such a correlation. To begin, we analyzed the k-means clustering to determine what difficulties we might have.

Difficulties

- Susceptible to local minima: Since calculating optimal clusters is an NP-hard problem, the algorithms typically used for k-means are heuristic and converge to local optima. Therefore, we would not be able to calculate global optima
- "Clusters": The galaxy and dark matter halo positions are uncorrelated (rather, the ellipticities and the dark matter halo positions are). Thus, the idea of clusters (as traditionally defined) does not perfectly apply to the data.
- Centroids: Using the positional centroid of the points assigned to a particular cluster does not provide any information regarding the existence of a halo.
- Number of Clusters: For predicting a single halo, the algorithm would converge after only one iteration under traditional k-means, as all points would be assigned to a single cluster.

Approach

To deal with these difficulties, we attempted to modify the k-means algorithm as follows.

- Assignment rule: We attempted to use a modified distance metric for calculating assignments based on the tangential ellipticity. This distance metric was

  $$d_{pdm}(h, g) = (e_1 \cos(2\phi) + e_2 \sin(2\phi))(r)$$

  where the angle from the galaxy center, ϕ, was defined as follows:

  $$\phi = \arctan\left(\frac{g_y - h_y}{g_x - h_x}\right)$$

  We calculate r as the euclidean distance.

  $$r = \sqrt{(g_y - h_y)^2 + (g_x - h_x)^2}$$

- Centroid updating rule: In place of the usual centroid calculation, we modified the centroid update rule in an attempt to maximize local tangential ellipticity. After we calculated the centroid for a particular cluster, we perturbed it by using neighboring points to generate an update vector.

  We know that the effect of a dark matter halo on surrounding galaxies is inversely related to the distance to the galaxy. Specifically, dark matter halos cause elongation of galaxies along the tangential axis. Thus, for each neighboring point, we construct a vector from the galaxy, g, to the halo, h.

  $$\overrightarrow{v} = \mathbf{h} - \mathbf{g}$$

  We calculate the length of this vector (distance from galaxy to halo). We take the vector normal, $\overrightarrow{n}$, to the galaxy's major axis (based on its ellipse). We then project $\overrightarrow{v}$ onto $\overrightarrow{n}$ as follows:

  $$\overrightarrow{p} = \frac{\overrightarrow{v} \cdot \overrightarrow{n}}{\overrightarrow{n} \cdot \overrightarrow{n}} \overrightarrow{n}$$

  We then calculate the update vector:

  $$\overrightarrow{u} = \overrightarrow{v} - \overrightarrow{p}$$

  We average the update vectors for all of the neighboring points. Then, we apply this final update vector to the centroid to perturb it. We do not consider convergence until either the cluster center has stopped changing or no new assignments have been calculated.

- Scoring function: To deal with the convergence to local minima, we initialize the algorithm with a k parameter that is greater than the number of halos. Once the clusters have converged, we score each cluster and return the top

clusters. After running this algorithm several times, we take the average of the clusters.

We used two types of scoring functions, one based on signal (tangential ellipticity) and one based on maximum likelihood. The signal-based scoring function was simply the sum of all tangential ellipticities of its neighbors. The maximum likelihood approach was based on single maximum likelihood. That is,

$$f = \frac{1}{r}$$

$$f_1 = -\cos 2\phi \cdot f$$

$$f_2 = -\sin 2\phi \cdot f$$

$$c = (f_1 - e_1)^2 + (f_2 - e_2)^2$$

$$s = e^{-\frac{c}{2}}$$

Batch Gradient Descent

Introduction and Formulation

We wish to formulate the halo-center finding problem as a minimization problem in order to apply the batch gradient descent algorithm to an objective function and hopefully obtain a good set of halos. Given a sky with m galaxies and n halos, we write

$$g^{(i)} = \begin{bmatrix} x^{(i)} \\ y^{(i)} \\ e_1^{(i)} \\ e_2^{(i)} \end{bmatrix}, h^{(j)} = \begin{bmatrix} y^{(j)} \\ x^{(j)} \\ \theta^{(j)} \end{bmatrix} \quad (1)$$

for $i = 1, ..., m$, $j = 1, ..., n$. For each galaxy and halo, x and y represent its x and y coordinates. $e_1^{(i)}$ and $e_2^{(i)}$ are the real and imaginary ellipticities of the i'th galaxy, respectively, and $\theta^{(j)}$ is the Einstein radius of the jth halo. The positional angle of the vector from the jth halo center to the ith galaxy is

$$\phi^{(ij)} = \arctan\left(\frac{y^{(i)} - y^{(j)}}{x^{(i)} - x^{(j)}}\right)$$

and the observed (complex) ellipticity of each galaxy is $e_c^{(i)} = e_1^{(i)} + ie_2^{(i)}$. In the weak lensing limit, the observed ellipticity for galaxy i is

$$e_c^{(i)} = e_s^{(i)} + \sum_j \gamma^{(ij)}$$

where $e_s^{(i)}$ is the natural ellipticity of the ith galaxy, $\gamma^{(ij)} = \exp[2\phi^{(ij)}i]\frac{\theta^{(j)2}}{\theta_r^{(ij)2}}$, where $\theta_r^{(ij)}$ is the angle between the galaxy and the halo with respect to the observer. Since the observer is very far away, this approximates to $\theta_r^{(ij)} = r^{(ij)} = [(x^{(i)} - x^{(j)})^2 + (y^{(i)} - y^{(j)})^2]^{1/2}$, the distance between them in the image. Thus, we may write

$$e_c^{(i)} = e_s^{(i)} + \sum_j \exp[2\phi^{(ij)}i]\frac{\theta^{(j)2}}{[(x^{(i)} - x^{(j)})^2 + (y^{(i)} - y^{(j)})^2]}$$

Now, the assumption given is that the magnitude of the average ellipticity should be close to zero, so for the objective function, we wish to minimize

$$\left|\sum_i e_s^{(i)}\right|^2.$$

From the above equations, we can write

$$e_s^{(i)} = e_c^{(i)} - \sum_j \exp\left[2i\phi^{(ij)}\right]\frac{\theta^{(j)2}}{[(x^{(i)} - x^{(j)})^2 + (y^{(i)} - y^{(j)})^2]}$$

Then, we have

$$J(h) = \left|\sum_i e_s^{(i)}\right|^2 =$$

$$\left|\sum_i e_c^{(i)} - \sum_j \exp\left[2i\phi^{(ij)}\right]\frac{\theta^{(j)2}}{[(x^{(i)} - x^{(j)})^2 + (y^{(i)} - y^{(j)})^2]}\right|^2$$

We then wish to minimize this value with respect to $\theta^{(j)}, x^{(j)}, y^{(j)}$ for $j = 1, ..., n$. Since we wish to perform a batch gradient descent with constraints (in this case, $0 \leq x^{(j)}, y^{(j)} \leq 4200$, $\theta^{(j)} \geq 0$), we could add a log-barrier function to penalize values that are too close to the barriers. However, for the sake of simplicity, we broke the function into two parts. Since the log-barriers are added linearly to the objective function, we first apply the gradient descent rule under the assumption of no log barriers, then add the barriers at the end.

Using Euler's identity and the fact that $e_c^{(i)}$ is complex, we rewrite the above as $J(h) = a^2 + b^2$, with

$$a = \sum_i e_1^{(i)} - \sum_j \cos\left[2\phi^{(ij)}\right]\frac{\theta^{(j)2}}{[(x^{(i)} - x^{(j)})^2 + (y^{(i)} - y^{(j)})^2]}$$

$$b = \sum_i e_2^{(i)} - \sum_j \sin\left[2\phi^{(ij)}\right]\frac{\theta^{(j)2}}{[(x^{(i)} - x^{(j)})^2 + (y^{(i)} - y^{(j)})^2]}$$

Noting the similarities in the equations above, we write

$$a = \sum_i e_1^{(i)} - \sum_j \cos\left[f_0^{(ij)}\right] f_1^{(ij)}$$

$$b = \sum_i e_2^{(i)} - \sum_j \sin\left[f_0^{(ij)}\right] f_1^{(ij)}$$

where

$$f_0^{(ij)} = 2\arctan\left(\frac{y^{(i)} - y^{(j)}}{x^{(i)} - x^{(j)}}\right)$$

$$f_1^{(ij)} = \frac{\theta^{(j)2}}{[(x^{(i)} - x^{(j)})^2 + (y^{(i)} - y^{(j)})^2]}$$

Update Rule

With learning rate α, we give the batch gradient descent rule for our objective:

$$\theta^{(j)} := \theta^{(j)} - \alpha \frac{\partial}{\partial \theta^{(j)}} J(h)$$

$$x^{(j)} := x^{(j)} - \alpha \frac{\partial}{\partial x^{(j)}} J(h)$$

$$y^{(j)} := y^{(j)} - \alpha \frac{\partial}{\partial y^{(j)}} J(h)$$

Where

$$\frac{\partial}{\partial \theta^{(j)}} J(h) = 2a \frac{\partial a}{\partial \theta^{(j)}} + 2b \frac{\partial b}{\partial \theta^{(j)}}$$

$$\frac{\partial}{\partial x^{(j)}} J(h) = 2a \frac{\partial a}{\partial x^{(j)}} + 2b \frac{\partial b}{\partial x^{(j)}}$$

$$\frac{\partial}{\partial y^{(j)}} J(h) = 2a \frac{\partial a}{\partial y^{(j)}} + 2b \frac{\partial b}{\partial y^{(j)}}$$

and

$$\frac{\partial a}{\partial \theta^{(j)}} = -\sum_i \cos(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial \theta^{(j)}}$$

$$\frac{\partial b}{\partial \theta^{(j)}} = -\sum_i \sin(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial \theta^{(j)}}$$

$$\frac{\partial a}{\partial x^{(j)}} = -\sum_i \left[\cos(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial x^{(j)}} - \sin(f_0^{(ij)}) \frac{\partial f_0^{(ij)}}{\partial x^{(j)}} f_1^{(ij)} \right]$$

$$\frac{\partial b}{\partial x^{(j)}} = -\sum_i \left[\sin(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial x^{(j)}} + \cos(f_0^{(ij)}) \frac{\partial f_0^{(ij)}}{\partial x^{(j)}} f_1^{(ij)} \right]$$

$$\frac{\partial a}{\partial y^{(j)}} = -\sum_i \left[\cos(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial y^{(j)}} - \sin(f_0^{(ij)}) \frac{\partial f_0^{(ij)}}{\partial y^{(j)}} f_1^{(ij)} \right]$$

$$\frac{\partial b}{\partial y^{(j)}} = -\sum_i \left[\sin(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial y^{(j)}} + \cos(f_0^{(ij)}) \frac{\partial f_0^{(ij)}}{\partial y^{(j)}} f_1^{(ij)} \right]$$

with

$$\frac{\partial f_1^{(ij)}}{\partial \theta^{(j)}} = \frac{2\theta^{(j)}}{r_y^{(ij)2} + r_x^{(ij)2}}$$

$$\frac{\partial f_0^{(ij)}}{\partial x^{(j)}} = \frac{2 r_y^{(ij)}}{\left(1 + \left[\frac{r_y^{(ij)}}{r_x^{(ij)}}\right]^2\right) r_x^{(ij)2}}$$

$$\frac{\partial f_1^{(ij)}}{\partial x^{(j)}} = \frac{2 r_x^{(ij)} \theta^{(j)2}}{\left(r_y^{(ij)2} + r_x^{(ij)2}\right)^2}$$

$$\frac{\partial f_0^{(ij)}}{\partial y^{(j)}} = -\frac{2}{\left(1 + \left[\frac{r_y^{(ij)}}{r_x^{(ij)}}\right]^2\right) r_x^{(ij)}}$$

$$\frac{\partial f_1^{(ij)}}{\partial x^{(j)}} = \frac{2 r_y^{(ij)} \theta^{(j)2}}{\left(r_y^{(ij)2} + r_x^{(ij)2}\right)^2}$$

and $r_x^{(ij)} = x^{(i)} - x^{(j)}$, $r_y^{(ij)} = y^{(i)} - y^{(j)}$.

Adding a log-barrier

To penalize values that approach the boundaries, we add a function $B(h)$ to the objective function, where $B(h)$ is

$$-\sum_j [\frac{1}{2}\log([4200 - x^{(j)}]^2) + \frac{1}{2}\log([0 - x^{(j)}]^2) +$$

$$\frac{1}{2}\log([4200 - y^{(j)}]^2) + \frac{1}{2}\log([0 - y^{(j)}]^2) + \frac{1}{2}\log([0 - \theta^{(j)}]^2)$$

We can see that as any of the parameters approaches a boundary, B goes to infinity. To keep B from dominating the objective function far away from the barrier, we add a scaling parameter μ, so that our new objective function is $J'(h) = a^2 + b^2 + \mu B(h)$. The new gradient updates are

$$\theta^{(j)} := \theta^{(j)} - \alpha \frac{\partial}{\partial \theta^{(j)}} J'(h)$$

$$x^{(j)} := x^{(j)} - \alpha \frac{\partial}{\partial x^{(j)}} J'(h)$$

$$y^{(j)} := y^{(j)} - \alpha \frac{\partial}{\partial y^{(j)}} J'(h)$$

with

$$\frac{\partial a}{\partial \theta^{(j)}} = \frac{1}{\theta^{(j)}} - \sum_i \cos(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial \theta^{(j)}}$$

$$\frac{\partial b}{\partial \theta^{(j)}} = \frac{1}{\theta^{(j)}} - \sum_i \sin(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial \theta^{(j)}}$$

$$A_X = -\sum_i \left[\cos(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial x^{(j)}} - \sin(f_0^{(ij)}) \frac{\partial f_0^{(ij)}}{\partial x^{(j)}} f_1^{(ij)} \right]$$

$$B_X = -\sum_i \left[\sin(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial x^{(j)}} + \cos(f_0^{(ij)}) \frac{\partial f_0^{(ij)}}{\partial x^{(j)}} f_1^{(ij)} \right]$$

$$A_Y = -\sum_i \left[\cos(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial y^{(j)}} - \sin(f_0^{(ij)}) \frac{\partial f_0^{(ij)}}{\partial y^{(j)}} f_1^{(ij)} \right]$$

$$B_Y = -\sum_i \left[\sin(f_0^{(ij)}) \frac{\partial f_1^{(ij)}}{\partial y^{(j)}} + \cos(f_0^{(ij)}) \frac{\partial f_0^{(ij)}}{\partial y^{(j)}} f_1^{(ij)} \right]$$

$$\frac{\partial a}{\partial x^{(j)}} = \frac{1}{x^{(j)}} - \frac{1}{4200 - x^{(j)}} + A_X$$

$$\frac{\partial b}{\partial x^{(j)}} = \frac{1}{x^{(j)}} - \frac{1}{4200 - x^{(j)}} + B_X$$

$$\frac{\partial a}{\partial y^{(j)}} = \frac{1}{y^{(j)}} - \frac{1}{4200 - y^{(j)}} + A_Y$$

$$\frac{\partial b}{\partial y^{(j)}} = \frac{1}{y^{(j)}} - \frac{1}{4200 - y^{(j)}} + B_Y$$

The rest of the algorithm proceeds as without the barrier.

RESULTS AND DISCUSSION

Modified K-Means

We tested the variations of the K-means algorithm with all possible combinations of traditional and different factors.

Using a combination of the traditional assignment and modified update rule with the maximum likelihood scoring function produced results that were better than the random benchmark and the single maximum likelihood benchmark. The reason for the improvement was its ability to distinguish multiple clusters more effectively than the single maximum likelihood. However, it performed worse than the gridded signal benchmark.

Ultimately, the performance of the k-means algorithm was slow computationally and poor relative to the benchmarks. The computational complexity was high due to the numbers of iterations, clusters, and averaging. The main issue was the susceptibility to local optima, which is an intrinsic weakness of the algorithm. For example, on a typical run, eight out of the ten runs would be relatively accurate, but the other two had high enough variation to introduce significant error. One possible cause is the fact that the clusters could be surrounding the halos, but the individual centers of the clusters would not be close to true halo.

One potentially interesting approach would be to construct a mixture of Gaussians model that attempted to learn the density of dark matter halos with regard to galaxies. We would then use the model to generate potential dark matter halo locations.

Batch Gradient Descent

We tested batch gradient descent on each training sky by discretizing and choosing many sets of initialization points and returning the set of parameters that converged to the best objective. Unfortunately, application of this algorithm performed worse than the random benchmark. This is likely due to a number of factors which we outline below.

First and foremost, our objective function is non-convex. Thus, even in instances in which the gradient converged (which was not always the case), we could at best hope for locally optimal results. The objective function is also highly sensitive to the behavior of galaxies that are very close to the halos, and thus gradient descent behaves poorly.

In addition, the signal this objective function reflects is very weak. It is quite reasonable that, even in the absence of dark matter, the average over the galaxy's ellipticities will be nonzero.

Finally, though we are in the weak lensing regime, the assumption that we are operating in the weak lensing limit may have been inaccurate. However, the full formalism for weak-lensing regime-observed ellipticity is

$$e_c^{(i)} = \frac{e_s^{(i)} + \frac{\gamma}{1-\kappa}}{1 + \frac{\gamma}{1-\kappa} e_s^{(i)}}$$

where κ is convergence, which we are not given. Using this to develop the objective, however, yields a function that is even less conducive to gradient descent.

While initial progress was promising, neither our modified clustering approach nor our batch gradient descent rule yielded robust results. We believe that, despite our best efforts, the former model was too coarse to adequately respond to the highly nuanced lensing perturbations caused by the presence of halos, and the latter was subject to the limitations of non-convexity on the objective function. Though we had hoped to test alternate approaches, for instance binary classification indicating the approximate presence of halos in sky localities or factor analytical models that acknowledge variation in mass, density, and position of halos as unobserved variables, we struggled to adequately featurize and linearly separate the data. We also found it difficult to formulate linear approximations for nonlinearity intrinsic to the problem. Given more time, however, we believe progress could be made through the development of techniques for solving nonlinear factor analytical models.

REFERENCES

1. Narayan, R., & Bartelmann, M. (1996). Lectures on gravitational lensing. arXiv preprint astro-ph/9606001.
2. Bernstein, G. M., & Jarvis, M. (2007). Shapes and shears, stars and smears: Optimal measurements for weak lensing. The Astronomical Journal, 123(2), 583.
3. Bartelmann, M., & Schneider, P. (2001). Weak gravitational lensing. Physics Reports, 340(4), 291-472.
4. Padmanabhan, N., Seljak, U., & Pen, U. L. (2003). Mining weak lensing surveys. New Astronomy, 8(6), 581-603.
5. Munshi, D., Valageas, P., Van Waerbeke, L., & Heavens, A. (2008). Cosmology with weak lensing surveys. Physics Reports, 462(3), 67-121.