Detecting Dark Matter Halos using Principal Component Analysis

William Chickering, Yu-Han Chou
Computer Science, Stanford University, Stanford, CA 94305
(Dated: December 15, 2012)

Principal Component Analysis (PCA)-based pattern recognition is employed to predict the locations of dark matter halos within simulated galaxy distributions. Local maxima in the net tangential force produced by galaxy positions and ellipticities forms a feature space upon which PCA is performed to distinguish halo from non-halo maxima. Within a dimensionally reduced space large halo types are successfully clustered into two types, facilitating smart subtraction from the tangential force landscape in order to better resolve minor halos. The strengths and weaknesses of this novel approach are examined.

INTRODUCTION

No community can claim greater big data challenges than that of astronomers. For example, the Large Synoptic Survey Telescope (LSST) is scheduled to come online in 2018 and will produce on the order of 60 petabytes of data during its ten year run [1]. In response to this impending deluge of information, astronomers are seeking new approaches for analyzing their observational data. The Kaggle.com competition Observing Dark Worlds (http://www.kaggle.com/c/DarkWorlds) is an example of this. The challenge is to locate the center of one to three dark matter halos given the positions and ellipticities of galaxies produced by a simulation based on a model of general relativity. To illustrate this, in Fig. 1 we show the galaxies of a typical sky (ignore the intensity data for now). Galaxies are represented as short lines with orientations representing their ellipticities. A large dark matter cloud can alter the apparent positions and ellipticities of galaxies, such that they appear to circulate around the center of the dark matter. This effect, known as weak gravitational lensing, is exploited by astronomers to locate dark matter. Historically, model fitting is used to predict halo center locations [2]. Our approach is different. In this article we outline an image processing technique, based on principal component analysis, that could potentially complement standard physics-based techniques to improve the precision of dark matter detection.

FEATURE SPACE

Our dataset consists of a simulated distributions of 300-800 galaxies spread over a square field referred to as a *sky*. A training set is provided consisting of 300 skies, each of which includes one to three dark matter halos. Within a sky, each galaxy is fully described by position coordinates x and y along with ellipticity[4] vector components e_1 and e_2. We choose not to use these raw data but rather analyze a derivative quantity e_{tf} called tangential force which is known to be correlated with dark matter and is defined as

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

where ϕ is the angle formed between a galaxy and a field point (x',y') given by

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

That is, a particular galaxy exerts a e_{tf} upon a field point that depends only on ϕ. We construct feature vectors using the net tangential force $E_{tf}(x',y') = \sum e_{tf}$ obtained by summing the contributions exerted by all galaxies upon a particular field point. Unlike the galaxy distribution itself, E_{tf} has the virtue of being continuous and differentiable across a sky.

The true distribution of galaxy positions and ellipticities is essentially random such that on average $E_{tf} = 0$. When a dark matter cloud is positioned between a group of galaxies and a distant observer, the collective mass of the cloud bends the light, elongating galaxy shapes and correlating galaxy positions such that regions of large E_{tf} form. In Fig. 1 we show a typical sky simulation illustrating the relationship between the galaxy distribution and E_{tf} where dark regions indicate positive E_{tf} and light green regions indicate negative E_{tf}. The darkest region in the figure correlates with a circulating pattern of galaxies and indicates the general vicinity of a large dark matter halo. What is not obvious from inspecting the figure is that a second, much smaller, halo is positioned roughly half way between the large halo and the right edge of the sky. The dominance of a single large halo and near invisibility of one or two minor halos is characteristic of the competition's dataset.

METHOD

The Observing Dark Worlds competition has attracted a great deal attention. It is clear from online forums that contestants are applying a number of techniques. While many, including the authors, have had success in predicting the locations of dominant halos, it is the virtually

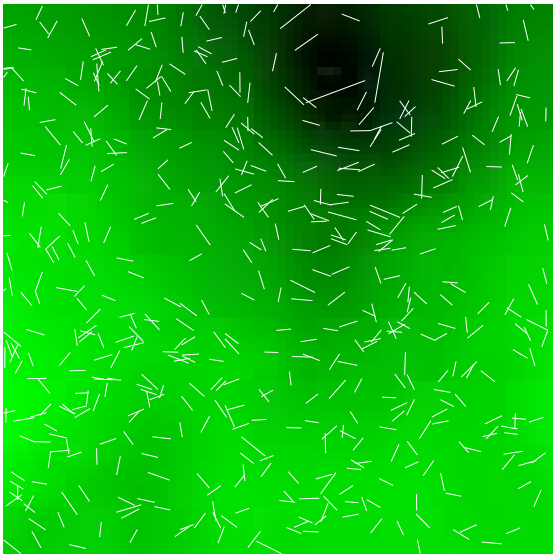

FIG. 1. An example galaxy distribution from the Kaggle.com Observing Dark Worlds competition. The intensity data represents net tangential force E_{tf} with dark regions indicating relatively large positive values that are consistent with the presence of a dark matter halo.

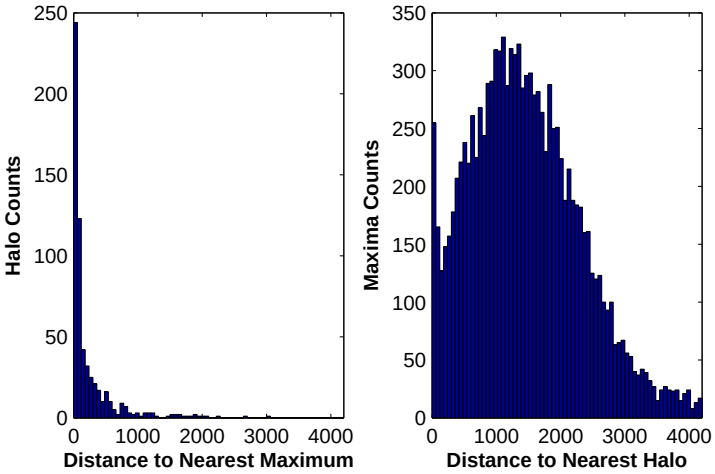

FIG. 2. Left) Histogram of dark matter halos binned according to distance to nearest local maximum in E_{tf}. Right) Histogram of local maxima in E_{tf} binned according to distance to nearest halo.

imperceptible minor halos that present the greatest challenge. A viable approach to this problem would seem to be to study the training data to learn the functional form of the E_{tf} resulting from a large halo. Indeed the standard framework used for numerical analysis of gravitational lensing effects employs a variety of models to which data is fit[2]. Alternatively, one could take a more naive approach, without the aid of physics, and apply fitting algorithms such as a weighted linear regression or estimation maximization to fit the gross features of a sky's big halo.

Our initial efforts consisted of such fitting techniques. In particular, an estimation maximization algorithm in which the tangential force is modeled as having a Gaussian profile produces reasonable predictions of large halos, but proved ineffective at identifying their minor counterparts. In addition, several weighting strategies were applied to construct quantities derivative of the tangential force (e.g. replacing E_{tf} with $\sum e_{tf}/r^a$, where $0 < a < 1$) in an effort to achieve greater resolution of the minor halos. But again, the results were disappointing. Next, clustering techniques were employed, which yielded somewhat better results. In particular, a divisive hierarchical clustering approach, which breaks apart the E_{tf} landscape at local minima effectively reduces the original problem of finding minor halos in the sky to evaluating the likelihood that a particular cluster, with its respective features of size, shape, etc., contains a halo. Ultimately, none of these results were satisfactory. Resigned to the fact that we faced a very difficult challenge, we decided to take a radically different approach and treat the problem of halo prediction as that of image recognition. It is the authors' belief that an optimal solution to a problem as challenging as the detection of secondary halos will ultimately consist of many techniques working in tandem. In this spirit, we focus our report on the use of principal component analysis (PCA) to develop a pattern recognition algorithm for detecting dark matter halos.

Using PCA

In developing an image processing technique, windowing of the data is helpful in order to isolate relevant features in the data. One approach to the present problem could be a sliding-window in which the algorithm scans through training data recording features along with labels indicating the presence or absence of a halo. The problem with such an approach lies in the fact that halos are so uncommon, and therefore, the amount of negative $y^{(i)} = 0$ training data dwarfs that of the positive $y^{(i)}$ data. A critical observation helps us here. In the left panel of Fig. 2 we show a histogram of dark matter halos binned according to the distance to the nearest local maximum in E_{tf}. The data clearly show that the preponderance of halos are located near such a maximum. An overly optimistic interpretation of this data might suggest that halo detection is not so difficult after all. The right panel of Fig. 2 reveals the sobering reality that while nearly all halos are found near maxima, the vast majority of local maxima are nowhere near a halo. So, while peaks in E_{tf} do not solve the problem of halo detection, confining our algorithm to the analysis of regions around maxima dramatically reduces the problem's

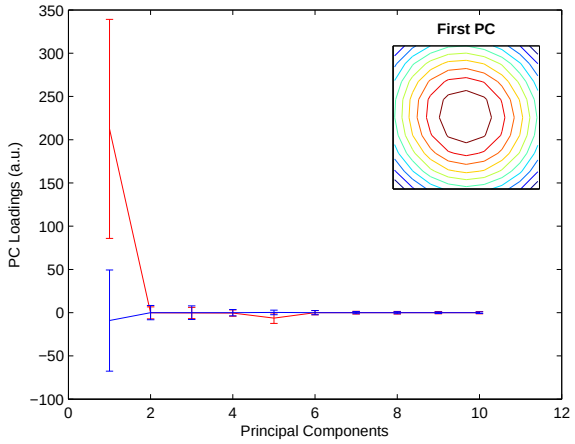

FIG. 3. The first ten principal component loadings within the space formed via PCA performed on windowed. The red curve corresponds to halos while the blue data corresponds to non-halos. Error bars correspond to ± one standard deviation. The data clearly shows separation of halo from non-halos in the 1D projection onto first principal component. The inset shows the first principal component.

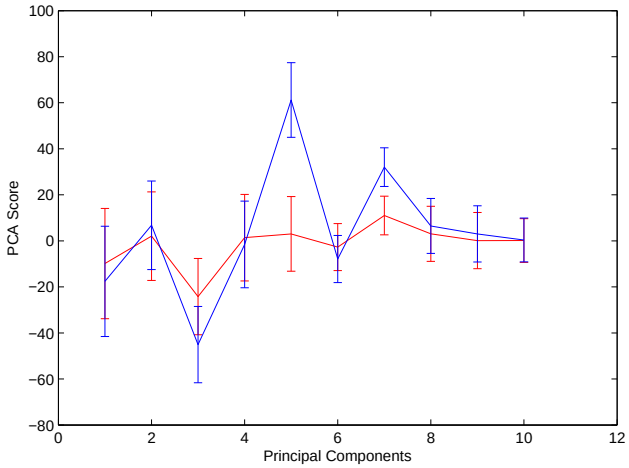

FIG. 4. The first ten principal components of big halos ($y^{(i)} = 1$) in the PCA space formed from full skies centered about local maxima in E_{tf}. The data clearly identifies two types of big halos, represented by the blue and red curves.

phase space.

To train the algorithm, a square region (with side 10% that of the full sky) surrounding each local maxima in E_{tf} forms our feature vector $x^{(i)}$. PCA is performed on 80% of all skies resulting in $m \approx 10,000$ training vectors among which about 300 are halos (i.e. $y^{(i)} = 1$). 20% of the skies are excluded to allow for testing of the algorithm. Once in the basis formed by PCA only the first ten principal components are used such that our data is dimensionally reduced. When testing a sky the region within a window centered at each local maxima in E_{tf} is projected into this ten dimensional space. Using a naive Bayes model, the principal component loadings reveal a probability that a local maxima is a halo.

The utility of PCA is that it provides a basis in which the variance of the data is maximized. Even using our local maxima technique, the ratio of positive-to-negative training data is 0.03. The basis formed through PCA therefore provides a representation in which the variance among local maxima is maximized. It is not clear *a priori* that halos would be distinguishable from non-halos in such a basis. Fortunately, the distinction can be made. In Fig. 3 shows the first ten principal component loadings for halos $y^{(i)} = 1$ versus non-halos $y^{(i)} = 0$. The error bars correspond to ± one standard deviation. The data makes clear that separation is achieved by more than one sigma in the first principal component, which is depicted in the inset of Fig. 3. Note that the first principal component captures the radial symmetry that characterized a large halo. Importantly, the degree of separation achieved within this dimensionally reduced space is largely dependent on how well the data is vetted prior to the PCA. That is, if a larger number of negative results were included the principal components chosen would correspond to those that maximize the variance among negative results and less so between positive and negative. Simply considering all local maxima in E_{tf} with values greater than zero, this approach effectively predicts major halos.

Two Halo Types

To predict the position of minor halos the algorithm attempts to subtract the effect of the major halo. This process is complicated by several factors, not least of which is the fact that the E_{tf} resulting from big halos varies in shape. Once again PCA helps us. To learn about big halo types we process our training data in a separate manner from that described above. For each local maximum in E_{tf} in each sky, we circularly shift the sky such that the maximum is centered. PCA is then performed on a collection of *full skies*; that is, a separate training set is formed consisting of 10,000 full sky feature vectors. As was true in the case of the windowed algorithm described above the big halos can be distinguished from non-halo maxima, although not as well as before. But distinguishing between halos and non-halos is not the goal here. Rather, we find that big halos cluster quite convincingly within this full-sky local maximum PCA space! This fortuitous discovery is represented by Fig. 4, which shows the first ten principal component loadings for the mean of the two halo types. Once again, the errors correspond to ± one standard deviation. Particular components 5 and 7 show good separation. K-means clustering was used to

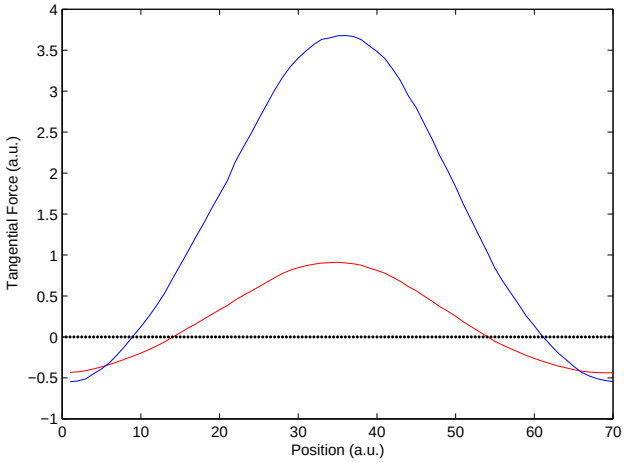

FIG. 5. The mean profiles of the two big halo types discovered within the PCA space formed from full skies centered about local maxima in E_{tf}. Type I is sharp is characterized by a tall sharp peak, while type II is shallow with minima of 50% the magnitude of its peak.

separate the big halos within this dimensionally reduced space. Note that larger cluster numbers did not yield consistent results during K-mean clustering but rather were very sensitive to initial conditions; thus, we can only convincingly identify two clusters or types of big halo. In Fig. 5 the E_{tf} profiles of the two halo types are shown. The two types have very different min-max ratios indicating that a subtraction of the wrong type would dramatically distort a sky's E_{tf} landscape.

Subtraction

The discovery of two big halo types in this space allows for a smarter subtraction. Having identified the location of the big halo using the windowed PCA algorithm, we then circularly shift the sky to center the halo and then project the full sky into the ten dimensional PCA space represented by Fig. 4. As in the windowed PCA algorithm, we use a naive Bayes model of this space to determine probabilities of the test halo being type I or type II. We then subtract the average halo of the corresponding type from the test sky. Fig. 6 demonstrates this process. In the topmost panel is a contour plot of the sky's original E_{tf}. The red, roughly circular contours indicate the position and magnitude of the major halo. Two minor halos also exist in this sky and are circled in the figure. The middle panel show the average big halo type identified with the sky. The E_{tf} of the middle plot is then subtracted from the original sky yielding the contours shown in the bottom plot. The minor halos are now more easily resolved with the same technique used for identifying major halos.

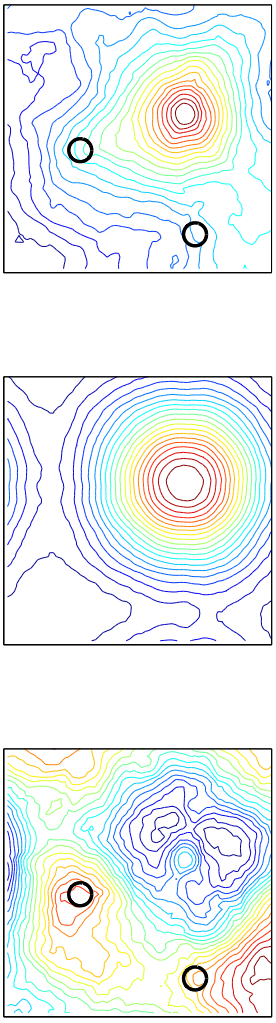

FIG. 6. Full sky contour plots of E_{tf} demonstrating the subtraction process. The upper panel shows the original E_{tf}; the large red peak corresponds to a major halo. Two minor halos that are difficult to resolve within the original landscape are circled. The middle panel shows the average type I big halo identified as most resembling the sky, represented with the ten dimensional space created via PCA, which has been aligned with the major halo of the sky. The bottom panel shows the result of subtraction, after which the minor halos are much better resolved.

RESULTS

Using PCA we successfully predict the locations of big halos. The Observing Dark Worlds competition includes several benchmarks to which our algorithm compares favorably. In particular, two benchmarks in the form of algorithms referred to as *Gridded_Signal_benchmark* and *Maximum_likelihood_benchmark* average distance errors of 1646 units and 632 units, respectively, when predicting major halo locations on the first 100 training skies. Our PCA-based algorithm, meanwhile, achieves

292 units. While our algorithm performs much better than the provided benchmarks, it should be noted that these algorithms are clearly quite poor and that other techniques are known to outperform the PCA algorithm in its present form. Nonetheless, we feel our technique has value in that it is a novel approach to the problem and offers avenues toward composite techniques.

Reliable prediction of minor halos remains elusive. The subtraction method described here works on some skies, but far from all. It is likely that the present approach is well suited for only a subset of dark matter halo configurations. As a consequence, using PCA may be best employed as part of a suite of techniques when performing halo detection.

Moreover, this PCA approach is far from optimized. Forming predictions on particular principal components rather than simply the first ten significantly improves performance for certain skies. Similarly, feature selection could be further optimized by modulating the window size and/or shape during halo detection; as well as using different window size/shapes when searching for minor versus major halos. Other key parameters such as bin size and the distance tolerance for determining label during training could be optimized. Finally, a key defect in our present algorithm is the use of circular shifts which result in significant artifacts when subtracting halos near a sky edge. Finally, the authors also speculate that performing PCA exclusively on big halos—that is, omitting non-halo E_{tf} local maxima—one might successfully cluster major halos into more than the two classifications identified in the present work.

CONCLUSION

By employing PCA and other machine learning algorithms, we can successfully predict the positions of major halos. By confining the feature space to local maxima in tangential force we dramatically reduce the problem's phase space and significantly improve the effectiveness of PCA at distinguishing halos from non-halos. Using a separate PCA technique, in which clustering technique are employed on a dimensionally reduced space, two distinct halo types are identified facilitating the categorization of test skies. By identifying major halo type, smarter subtractions are achieved providing greater resolution of the effect of minor halos on the tangential force landscape. By applying image pattern recognition techniques to a new problem space, a novel approach is developed and explored.

[1] LSST and Technology Innovation. [Online]. Available: http://www.lsst.org/lsst/about/technology.

[2] E. Jullo *et al.*, New J. Phys. **9**, 447 (2007).

[3] M. Turk, A. Pentland, Journal of Cognitive Neuroscience **3**, 71 (1991).

[4] The ellipticity describes the elongation of a galaxy along the x and y axes. See http://www.kaggle.com/c/DarkWorlds/details/an-introduction-to-ellipticity for details.