

A MACHINE LEARNING APPROACH FOR DYNAMICAL MASS MEASUREMENTS OF GALAXY CLUSTERS

M. NTAMPAKA¹, H. TRAC¹, D.J. SUTHERLAND², N. BATTAGLIA^{1, 3}, B. PÓCZOS², J. SCHNEIDER²

Draft version April 29, 2015

ABSTRACT

We present a modern machine learning approach for cluster dynamical mass measurements that is a factor of two improvement over using a conventional scaling relation. Different methods are tested against a mock cluster catalog constructed using halos with mass $\geq 10^{14} M_{\odot} h^{-1}$ from Multidark's publicly-available N -body MDPL halo catalog. In the conventional method, we use a standard $M(\sigma_v)$ power law scaling relation to infer cluster mass, M , from line-of-sight (LOS) galaxy velocity dispersion, σ_v . The resulting fractional mass error distribution is broad, with width $\Delta\epsilon \approx 0.87$ (68% scatter), and has extended high-error tails. The standard scaling relation can be simply enhanced by including higher-order moments of the LOS velocity distribution. Applying the kurtosis as a correction term to $\log(\sigma_v)$ reduces the width of the error distribution to $\Delta\epsilon \approx 0.74$ (16% improvement). Machine learning can be used to take full advantage of all the information in the velocity distribution. We employ the Support Distribution Machines (SDMs) algorithm that learns from distributions of data to predict single values. SDMs trained and tested on the distribution of LOS velocities yield $\Delta\epsilon \approx 0.46$ (47% improvement). Furthermore, the problematic tails of the mass error distribution are effectively eliminated. Decreasing cluster mass errors will improve measurements of the growth of structure and lead to tighter constraints on cosmological parameters.

Subject headings: cosmology: theory—dark matter—galaxies: clusters: general—galaxies: kinematics and dynamics—gravitation—large-scale structure of universe—methods: statistical

1. INTRODUCTION

Galaxy clusters have been utilized prominently in astrophysics and cosmology since pioneering work by Fritz Zwicky and George Abell. Clusters are the most massive gravitationally-bound systems in the Universe, with masses $\gtrsim 10^{14} M_{\odot}$, and contain scores to hundreds of galaxies embedded in dark matter halos. These objects are useful cosmological probes because halo abundance as a function of mass and redshift depend sensitively on fundamental cosmological parameters. Therefore, measures of cluster abundance can be used to constrain these parameters (e.g. Voit 2005; Allen et al. 2011; Sheth & Diaferio 2011). However, accurately measuring cluster masses for application in cosmology is a difficult endeavor.

Clusters can be identified across multiple wavelengths. They were first detected in the visible spectrum as overdensities of galaxies (e.g. Abell 1958; Zwicky et al. 1968). They are identified as overdensities of red galaxies in both visible and IR (e.g. Gladders & Yee 2005; Hao et al. 2010; Ascaso et al. 2012) and can be found as extended X-ray sources (e.g. Rosati et al. 2002; Vikhlinin et al. 2009a). Clusters are also detected by their unique signature in the cosmic microwave background, as a decrement below 218 GHz and an increment above, as predicted by Sunyaev & Zeldovich (1972) (e.g. Staniszewski et al. 2009; Marriage et al. 2011; Planck Collaboration et al. 2014c).

Once clusters are identified, mass measurements are needed to map observable cluster properties to the underlying mass. Cluster masses can be deduced from a variety of techniques including: x-ray observations from which one can infer a mass profile (e.g. Vikhlinin et al. 2009a; Mantz et al. 2010), a temperature-weighted gas mass via the Sunyaev-Zeldovich effect (e.g. Lueker et al. 2010; Hasselfield et al. 2013), mass measurement via strong and weak gravitational lensing (e.g. Johnston et al. 2007; Mandelbaum et al. 2008), the caustic technique which uses galaxy velocities to determine a mass profile (e.g. Biviano & Girardi 2003; Andreon 2010; Serra et al. 2011), the galaxy infall kinematics beyond the virial radius to determine a cluster mass profile (e.g. Zu & Weinberg 2013) and dynamical mass measurements which employ the virial theorem (e.g. Teague et al. 1990; Colless & Dunn 1996; Fadda et al. 1996; Carlberg et al. 1997; Girardi et al. 1998; Brodwin et al. 2010; Rines et al. 2010; Sifón et al. 2013; Ruel et al. 2013).

Zwicky (1933) used the dynamical mass approach. His work applied the virial theorem, using the dispersion of galaxy velocities to infer the mass of the Coma cluster. Because dynamical mass measurements probe the entire mass distribution, both regular baryonic matter as well as dark matter, he was able to use the virial theorem to conclude that dark matter outweighed luminous matter in the Coma system. The virial theorem, which relates kinetic energy to gravitational potential energy, predicts that halo mass relates to galaxy velocity dispersion, σ_v , as a power law.

The simplest approach is to treat clusters and their host halos as self-similar, dynamically-relaxed systems with the galaxy velocity dispersion, σ_v , summing up the

ntampaka@cmu.edu

¹ McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213

² School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

³ Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544

halo’s dynamics. However, evidence points toward complications that introduce scatter to the idealized case. Numerical simulations are useful tools in studying these complex dynamic systems and analyzing sources of scatter in dynamical mass predictions. Evrard et al. (2008) find that dark matter particle velocity dispersion scales with total mass as a power law, with a small lognormal scatter. They identify $\approx 10\%$ of the halo population as merger transients, exhibiting higher-than-expected dark matter particle velocity dispersions. But while a tight correlation is seen between halo mass and the velocity dispersion of simulated dark matter particles (σ_{DM}), correlations between the galaxy velocity dispersion (σ_v) and halo mass are more fraught with scatter (e.g. Saro et al. 2013). A $\lesssim 10\%$ bias between σ_{DM} and σ_v is found, though the sign of this bias is not agreed upon (e.g. Diemand et al. 2004; Faltenbacher et al. 2005; Biviano et al. 2006; Faltenbacher & Mathews 2007; Lau et al. 2010).

Also contributing to the bias and scatter are halo environment and triaxiality (e.g. White et al. 2010; Saro et al. 2013; Svensmark et al. 2014; Wojtak 2013) and projection effects (e.g. Cohn 2012; Noh & Cohn 2012), while Old et al. (2013) find a bias that is dependent on mass and galaxy selection strategy. Determining cluster parameters such as center and radii is a difficult endeavor; Skibba et al. (2011) find that the central galaxy may not be the brightest cluster galaxy and additionally is not necessarily at rest with respect to the center of the halo potential well. Determining which galaxies are cluster members is difficult, and the inclusion of interlopers also introduces scatter (e.g. Mamon et al. 2010). Old et al. (2014) compare a variety of cluster mass estimation techniques that rely on galaxy observables, including several virial-theorem-based mass estimates. They find that scatter in predicted mass increases below halo mass of $10^{14} M_\odot$, as the number of cluster members drops.

The complications of triaxiality, environment, galaxy selection, and mergers limit the accuracy with which halo masses can be directly correlated to velocity dispersion. However, in typical dynamical mass analyses, a wealth of observational information is boiled down in favor of summarizing the halo’s dynamics by a summary statistic: σ_v . This condensation of information is dictated as much by the virial theorem’s $M(\sigma_v)$ power law relationship as it is by simplicity, because taking full advantage of the wealth of information available in the full line-of-sight (LOS) velocity probability distribution function (PDF) manually is difficult. However, the application of nonparametric machine learning (ML) algorithms is a promising resource that may allow cluster masses to be predicted from the myriad data at our disposal.

Machine learning has been applied to astronomy and cosmology problems with mixed results. Object classification is a common application for machine learning in astronomy, for example: distinguishing Mira variables from semiregular and irregular red variables (Woźniak et al. 2004), classifying transient events (Mahabal et al. 2008), determining galaxy morphology (Banerji et al. 2010), and more recently, choosing clusters from simulated catalogs which resemble observed clusters (Hajian et al. 2014). Other applications range from predicting solar activity (e.g. Yu et al. 2009; Colak & Qahwaji 2009) to determining photometric redshifts (e.g. Gerdes et al. 2010; Ball et al. 2007), from cataloging impact craters on

Mars (Stepinski et al. 2009) to predicting the number of galaxies hosted by a dark matter halo (Xu et al. 2013).

In this work, we focus on improving cluster dynamical mass measurements by moving beyond a simple $M(\sigma_v)$ power law toward more fully utilizing the entire LOS velocity PDF. In Sec. 2, we discuss our simulations and mock galaxy catalogs. In Secs. 3 and 4, we lay out the methods and results for each of three approaches: first applying a virial-theorem-motivated power law to our mock catalog in Sec. 3.1, then using more of the velocity information by taking advantage of higher-order moments in Sec. 3.2. Finally, we utilize the full velocity PDF and the information contained therein by implementing Support Distribution Machines (SDMs), a machine learning algorithm that maps a full LOS velocity distribution to a halo mass prediction in Sec. 4. We report a comparison of these approaches and a discussion of the interpretation of the results in Sec. 5 and summarize our conclusions in Sec. 6. We use a Λ CDM cosmology throughout, with cosmological parameters consistent with Planck data (Planck Collaboration et al. 2014a): $\Omega_\Lambda = 0.69$, $\Omega_m = 0.31$, $\Omega_b = 0.048$, $h = 0.68$, $n = 0.96$, and $\sigma_8 = 0.82$

2. SIMULATION AND MOCK CATALOG

This work is based on massive halos from the publicly-available Multidark MDPL simulation⁴ (Hess et al., in prep.). Multidark is an N -body simulation containing 3840^3 particles in a box of length $1 h^{-1}$ Gpc and run using the L-Gadget2 code. The mass resolution of this simulation is $1.51 \times 10^9 M_\odot h^{-1}$, with cosmological parameters consistent with Planck data. Massive halos were gleaned from the BDMW halo catalog, which employs a bound density maximum (BDM) spherical overdensity halo finder with halo average density equal to 200 times the critical density of the Universe (Klypin & Holtzman 1997). These halo masses will be denoted M throughout this work.

Halos for our catalog are chosen from the $z = 0$ catalog for its large sample of massive halos. To mimic observed clusters in halo mass, observable minimum luminosity of cluster members (via subhalo mass), and richness, halos are chosen to meet the following criteria:

1. Halo minimum mass: $M \geq 10^{14} M_\odot h^{-1}$
2. Subhalo minimum mass: $M_{\text{subhalo}} \geq 10^{12} M_\odot h^{-1}$
3. Subhalo minimum number: $N_{\text{subhalo}} \geq 20$

The subhalo minimum mass of $M_{\text{subhalo}} \geq 10^{12} M_\odot h^{-1}$ corresponds to 663 particles in the smallest subhalos. Due to problems with subhalos not being matched with their host halo across the periodic boundary, clusters with centers that lie within $2.2 \text{ Mpc } h^{-1}$ of the box edge are pared from the sample.

Each subhalo and primary halo in the catalog is assigned a galaxy. This galaxy is placed at the center of each halo and subhalo, with no galaxy offset considered at this time. Each galaxy is assigned its subhalo host’s velocity. This method is intentionally simplistic, and is akin to abundance matching with the assumption

⁴ <http://www.cosmosim.org/>

of zero scatter, i.e. all subhalos above the minimum subhalo mass host an observable galaxy. The distribution of galaxies within the cluster is solely determined by the simulated substructure within the primary halo; no assumption about halo mass profiles is made. Galaxy line-of-sight velocities are calculated with respect to the mean galaxy LOS velocity of cluster members, with no assumption made concerning which of these galaxies is the brightest cluster galaxy. Because halo mass is tightly correlated with halo radius, projected subhalo radii (R_{sub}) are normalized by the halo's R_{200} .

We prepare three catalogs of these clusters. The Train Catalog is used for training and fits, and includes multiple line-of-sight views of halos. The Train Catalog has a flat mass function by design, and the number of projections per halo in 0.1 dex mass bin varies by bin to achieve 1000 training halos for each bin.

The default is to view each cluster first from the cardinal x -, y -, and z -directions and when additional LOS views are needed, these are chosen randomly on the surface of the unit sphere. While the catalog of unique halos contains 5,028 halos, the Train Catalog contains 15,000 projections of these halos, with more projections of the rare high-mass halos, creating a representative training sample that catalogs how these halos' velocities and positions might be distributed when viewed from any angle.

Smaller catalogs are used to evaluate the methods, containing only the three cardinal direction LOS views of each halo. A mass cut of $M \geq 3 \times 10^{14} M_{\odot} h^{-1}$ is applied to the Test Catalog to account for edge effects due to the hard lower mass cut of the sample and selection effects due to the N_{subhalo} constraint. A High-Mass Test Catalog, with a mass cut of $M \geq 7 \times 10^{14} M_{\odot} h^{-1}$ is also considered. These catalogs are summarized in Table 1.

It should also be noted that, despite the multiple projections of high-mass halos in the Train Catalog, there can be challenges at the high-mass end of the training sample. Training and predicting on the halos toward the more massive end of the spectrum is difficult because they are rare, leading to large statistical uncertainties.

3. POWER LAW METHODS AND RESULTS

In the next two sections, we move through successive methods, first describing the method, next reporting the results, then adding a layer of complexity with the subsequent method. Motivated by the virial theorem, we will explore improvements to dynamical mass measurements. The methods will be described in detail in the following sections, but are summarized in Table 2 for reference. They include the virial-theorem-motivated $M(\sigma_v)$ power law (PL1), a power law method that includes applying higher-order moments of the LOS velocity data (PL2), and four machine learning methods, each employing the same ML algorithm with different inputs: LOS velocities only (ML1 and ML2), and LOS velocities plus galaxy projected sky position (ML3 and ML4).

3.1. $M(\sigma_v)$ Power Law

We begin by using the virial theorem as a jumping-off point for dynamical mass measurements. The virial theorem states that, for an object that is stably bound by gravity, $2K + U = 0$, where K is the kinetic energy of the system and U is its gravitational potential energy. This can be extended to find a relationship between velocity

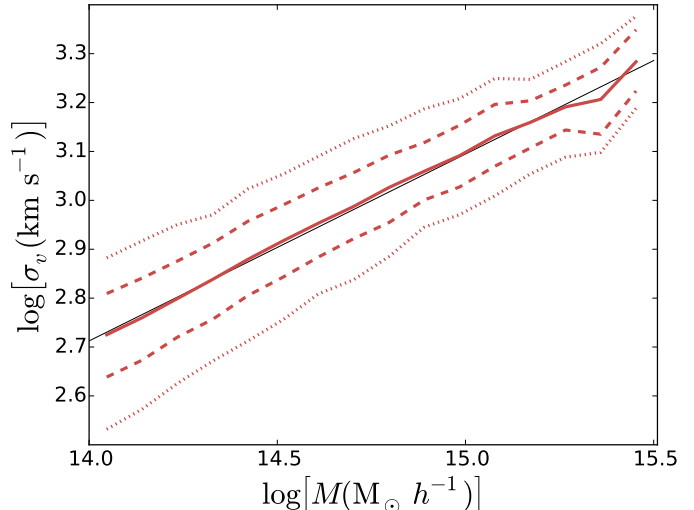


FIG. 1.— Line-of-sight galaxy velocity dispersion, σ_v , vs. halo mass, M . The binned halos are fit by a power law relation (thin solid black line). The 68% (dashed) and 95% (dotted) scatter show the substantial uncertainty in this $M(\sigma_v)$ relationship.

dispersion, $\sigma_v = \langle v^2 \rangle^{1/2}$, and mass, M :

$$\sigma_v^2 = \alpha_R \frac{GM}{R}, \quad (1)$$

where α_R is a constant of order one that depends on the cutoff definition of halo radius R and the density profile of the dark matter halo. It can be constrained by observation (e.g. Spitzer 1969). For a halo population with a known average mass density, $M \propto R^3$, and velocity dispersion relates directly to mass as

$$\sigma_v \propto M^{1/3}. \quad (2)$$

Studies of N -body simulations find that this power law generally holds with slope ≈ 0.33 (e.g. Evrard et al. 2008). Although there is a tight $M(\sigma_{DM})$ relationship for the velocity dispersion of dark matter, the halo mass is less tightly correlated with σ_v , the velocity dispersion of galaxies (e.g. Saro et al. 2013).

Halo mass M can be related to galaxy LOS velocity dispersion σ_v via the power law

$$\sigma_v(M) = \sigma_{15} \left(\frac{M}{10^{15} M_{\odot} h^{-1}} \right)^{\alpha}. \quad (3)$$

We find least-squares fit to $\log(\sigma_v) = \alpha \log(M) + \beta$ for the Train Catalog, binned in 0.1 dex mass bins. The power law best fit parameters for the Train Catalog are $\alpha = 0.382$ and $\sigma_{15} = 1244 \text{ km s}^{-1}$; this result is relatively insensitive to the catalog choice, and comparable best fit values are calculated for the Test Catalog. However, we caution that these parameters are a fit for a particular simulation and catalog, and should be applied with care to predict cluster masses for observational data.

Figure 1 shows the power law best fit and both 68% and 95% scatter. The power law is a good fit for median binned σ_v , albeit with significant scatter at all mass ranges considered.

It should be noted that the best fit to an $M(\sigma_{DM})$ relationship for this simulation has $\alpha = 0.34$, much closer to the expected value of 0.33. The steepness of our best

TABLE 1
CATALOG SUMMARY

Name	Minimum Halo Mass	Unique Halos	Projections per Unique Halo	Total Projections
Train Catalog	$1 \times 10^{14} M_{\odot} h^{-1}$	5028	Mass-dependent ¹	15000
Test Catalog	$3 \times 10^{14} M_{\odot} h^{-1}$	2278	3	6834
High-Mass Test Catalog	$7 \times 10^{14} M_{\odot} h^{-1}$	315	3	945

^aTo create a flat mass function with an equal number of training points per 0.1 dex halo mass bin.

TABLE 2
METHODS

Case	Approach	Summary Stats	Distribution Features	Color
PL1	Power Law	σ_v	—	Red
PL2	Power Law	σ_v & κ	—	Blue
ML1	Machine Learning: SDM	—	$ v_{\text{los}} $	Green
ML2	Machine Learning: SDM	—	$ v_{\text{los}} $ & $ v_{\text{los}} /\sigma_v$	Purple
ML3	Machine Learning: SDM	—	$ v_{\text{los}} $ & R_{sub}/R_{200}	Orange
ML4	Machine Learning: SDM	—	$ L_{\text{eff}} $	Brown

fit slope can be attributed to both numerical and physical effects. Saro et al. (2013) find that measurements of σ_v from small samples (e.g. $\lesssim 40$ galaxies) of very luminous galaxies tend to be smaller than the σ_v calculated with the inclusion of more plentiful, fainter galaxies. Because the halos at the low mass end of our catalog tend to have the fewest galaxies (on the order of 20 per cluster), this trend would preferentially bias the velocity dispersions down for the smallest-mass clusters. The lower-than-expected dispersions at the low-mass end thus steepen the best fit power law, giving an α greater than the expected value of 0.33. Additionally, Munari et al. (2013) find that dynamical friction and tidal disruption affect subhalo σ_v , with tidal stripping having a greater effect on higher-mass clusters, creating a velocity bias that changes with cluster mass. Both numerical and physical effects may come into play in our catalog, where a steeper α is evident for low-mass systems and a shallower one for higher-mass clusters.

The distribution of σ_v is roughly log-normal for halos of a given mass, and we define the residual, δ , as

$$\delta = \log(\sigma_v) - \log(\sigma_{v, \text{best fit}}), \quad (4)$$

where σ_v is the velocity dispersion calculated from LOS velocities and $\sigma_{v, \text{best fit}}$ is the expected velocity dispersion for a given halo mass (i.e. the velocity dispersion that would be calculated by Equation 3 with the α and σ_{15} values given).

The biweight estimator given by Beers et al. (1990) was considered for this and subsequent methods. However, a strength of the biweight estimator is in detecting and omitting interlopers' contribution to the sample variance; these interlopers have already been excluded by the catalog design. This estimator was found to have a larger range δ values compared to when the velocity dispersion was defined as the standard deviation. In order to provide a comparison to the most idealized power law method, the standard deviation was chosen as a measure of velocity dispersion, σ_v , for this catalog.

3.2. $M(\sigma_v)$ Power Law with Kurtosis

In the simple $M(\sigma_v)$ power law applied in Sec. 3.1, the full information of the LOS velocity PDF is summarized by a single statistic. But this PDF contains more information than is used. In the upcoming section, we will explore using higher-order moments of the PDF to improve dynamical mass measurements of galaxy clusters.

Figure 2 shows that stacked halos from differing populations—the full Test Catalog, large positive δ , and large negative δ —exhibit strikingly different shapes. While the full Test Catalog's shape is approximately Gaussian, the halos with a large negative δ are more sharply peaked near $v_{\text{los}}/\sigma_v = 0$ and the halos with large positive δ have a flatter distribution.

This difference in shape—the sharply peaked compared to the overly-flat curves—can be quantified by the excess kurtosis, κ , defined as:

$$\kappa = \frac{\sum_{i=1}^N (v_i - \bar{v})^4}{N\sigma_v^4} - 3, \quad (5)$$

where v_i is the line-of-sight velocity of the i^{th} galaxy, \bar{v} is the mean line-of-sight velocity of galaxies in the cluster, N is the number of galaxies in the cluster, and σ_v is the standard deviation of a cluster's galaxy line-of-sight velocities. The subtraction of 3 sets this value such that a standard normal distribution has an excess kurtosis of 0. Throughout this work, we refer to excess kurtosis simply as “kurtosis.”

Expanding on the illustration in Figure 2, Figure 3 shows the relation between δ and κ . Halos with wide, flat velocity PDFs (i.e. those with negative κ) tend to have larger δ values, while halos with sharply-peaked, high- κ PDFs have moderate-to-low δ values. The simplest reasonable fit for median δ as a function of binned mean kurtosis is a decaying exponential,

$$\delta(\kappa) = a \exp(-b * \kappa) + c, \quad (6)$$

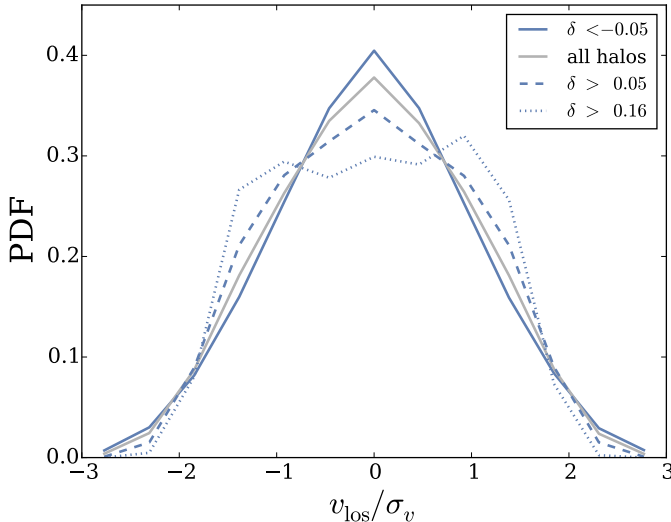


FIG. 2.— PDFs of normalized galaxy LOS velocities from stacked halos. The Test Catalog halo population (gray solid) is roughly Gaussian. However, halos with a large negative δ and underpredicted masses exhibit a sharply peaked PDF (blue solid), while those with a large positive δ and overpredicted masses have a flatter PDF (blue dashed). In the most extreme large positive δ (blue dotted), a flat, wide PDF emerges. The shape of the velocity PDF, quantified by the kurtosis, can be used to predict δ .

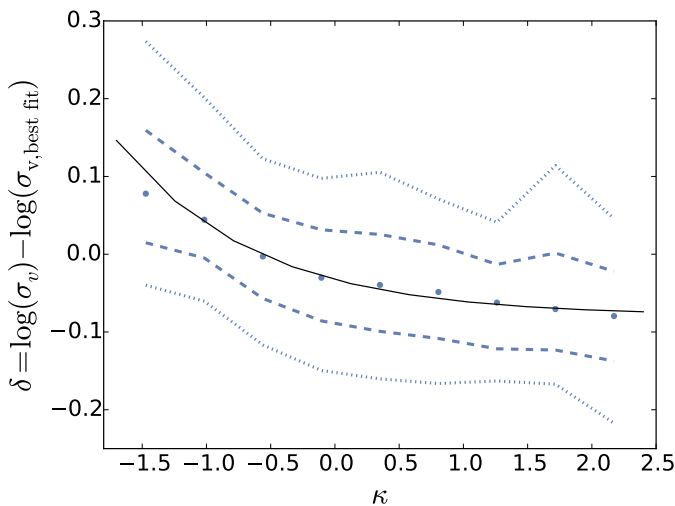


FIG. 3.— The kurtosis, κ , of the LOS velocity PDF is a predictor of δ (the residual from the $M(\sigma_v)$ power law, given by Equation 4). Individual clusters are binned, and the binned mean δ (blue points) are fitted to a decaying exponential (solid black curve). Dashed lines show binned 68% scatter; dotted show 95% scatter. Exploiting this relationship allows us to predict δ based on the kurtosis of the LOS velocity PDF and produce a κ -corrected σ_v to improve halo mass predictions.

with best fit parameters $a = 0.046$, $b = 0.93$, and $c = -0.079$. Using this fit as a predictor for residual δ in equation 4 allows us to calculate a κ -corrected velocity dispersion ($\sigma_{v, \kappa\text{-corrected}}$),

$$\log(\sigma_{v, \kappa\text{-corrected}}) = \log(\sigma_v) - \delta(\kappa), \quad (7)$$

from the measured velocity dispersion, σ_v , and the residual, $\delta(\kappa)$.

We now find a power law relating M and $\sigma_{v, \kappa\text{-corrected}}$ for the Train Catalog. Correcting the $\log(\sigma_v)$ values lin-

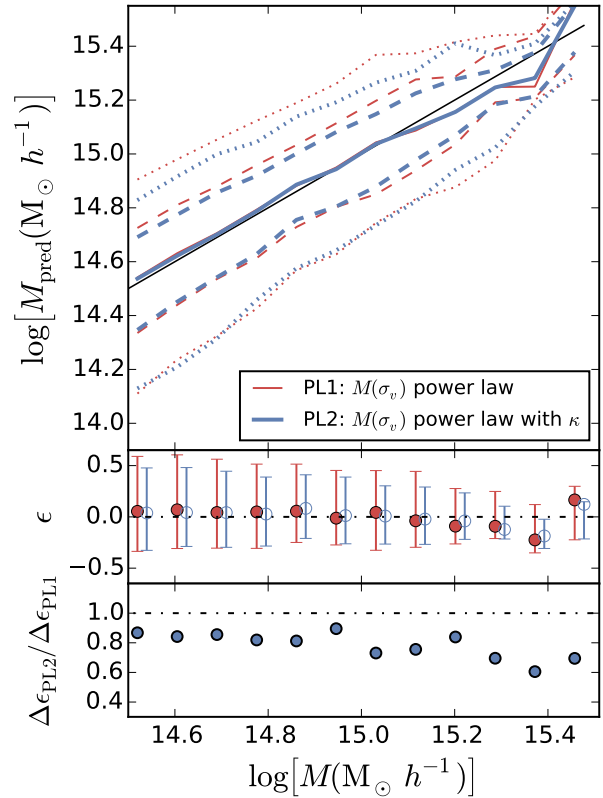


FIG. 4.— By adding kurtosis as a predictor of residual, we are able to decrease the scatter in the $M(\sigma_v)$ relationship. Top panel: predicted halo mass, M_{pred} vs. halo mass, M . PL2 mass predictions are improved by the addition of kurtosis to the $M(\sigma_v)$ power law. The PL2 method (thick blue) shows an improvement over the PL1 method (thin red). Binned median (solid), 68% (dashed) and 95% (dotted) scatter are shown for both PL1 and PL2. Middle panel: fractional mass error for PL1 (filled red circles) and PL2 (open blue circles). Points are the median ϵ , and error bars show 68% scatter. Blue points are displaced horizontally for clarity. The addition of kurtosis as a linear correction term slightly improves the negative bias at high masses, and decreases scatter at all mass scales. Bottom panel: the width of the 68% scatter of PL2 ($\Delta\epsilon_{\text{PL2}}$) is compared to that of PL1 ($\Delta\epsilon_{\text{PL1}}$). Kurtosis as a predictor of δ decreases the width of ϵ for all mass bins.

early with $\delta(\kappa)$ (as in equation 7) does not significantly alter the power law fit, with κ -corrected best fit values to Equation 3 of $\alpha = 0.384$ and $\sigma_{15} = 1241 \text{ km s}^{-1}$.

We next consider the fractional mass error

$$\epsilon = (M_{\text{pred}} - M)/M, \quad (8)$$

comparing the actual halo mass, M , to the predicted halo mass, M_{pred} . When applying the κ -corrected power law to halos in the Test Catalog, the fractional error decreases. The mean ϵ has moved closer to zero, from 0.128 to 0.079. The width of the 68% scatter in fractional mass error decreased as well, from 0.87 to 0.73. The inclusion of κ as a correction term has allowed us to better predict halo masses from our line-of-sight velocities, reducing both bias and scatter in fractional mass error.

Figure 4 shows the overall improvement of halo mass measurement of $\sigma_{v, \kappa\text{-corrected}}$ as compared to the measured velocity dispersion, σ_v . The top panel shows that at all mass ranges, the 68% and 95% scatters have decreased. The middle panel shows that the median fractional mass error, ϵ , edges closer to a zero bias for most

bins $\geq 10^{15} M_{\odot} h^{-1}$. Halos with mass $\geq 10^{15} M_{\odot} h^{-1}$ are found to have a more tightly correlated $\kappa - \delta$ relationship, which may explain why this κ correction method preferentially improves the prediction of the highest-mass halos. The bottom panel quantifies the ratio of 68% scatter of PL2 ($\Delta\epsilon_{\text{PL2}}$) compared to the 68% scatter of PL1 ($\Delta\epsilon_{\text{PL1}}$), showing that with the inclusion of κ in our $M(\sigma_v)$ power law, the scatter in ϵ decreases.

The application of kurtosis, κ , as a predictor for δ results in an improvement of mass measurements across masses considered. This result should not be surprising; kurtosis profiles have been used in conjunction with the Jeans equation to explore mass profiles (e.g. Lokas & Mamon 2003; Lokas et al. 2006); we refer to the original papers for details on this approach.

Mergers and infalling matter offer a natural explanation for the correlation between high δ and negative κ . A halo undergoing a merger or experiencing infalling matter tends to have a flat-topped or double-peaked velocity distribution (e.g. Ribeiro et al. 2011), resulting in a negative velocity PDF kurtosis. This corresponds well with what is found from simulated dark matter particle dispersion, that halos undergoing a merger tend to have masses overpredicted by an $M(\sigma_{\text{DM}})$ relationship (Evrard et al. 2008), leading to a large positive residual, δ . In the case of negative kurtosis, the relationship between κ and δ seems to be driven by the identification of halos undergoing a period of mass growth.

It should be noted that other moments of the LOS velocity PDF may be applied in a similar manner. Skewness, for example, is also a weak predictor of δ and can also be used as a correction term to the $M(\sigma_v)$ power law. For our catalog, applying skewness as a correction term to the $M(\sigma_v)$ power law decreases the mean fractional mass error by $\approx 22\%$, reducing the tendency to overpredict halo mass that is seen in PL1. However, the application of skewness as a correction term makes no significant decrease in the 68% scatter.

Though we have improved dynamical mass measurement by taking more advantage of the information encoded in the LOS velocity PDF, the method we have used here still merely summarizes the full velocity PDF: we have moved from one summary statistic, σ_v , in PL1, to two, σ_v and κ , in PL2.

4. MACHINE LEARNING METHODS AND RESULTS

Moving beyond summary statistics toward utilizing the full information encapsulated in the velocity PDF cannot easily be done manually. Taking advantage of the full LOS velocity PDF is a good candidate for moving from simple power law relationships to machine learning. The problem at hand is to take samples from a probability distribution (galaxy line-of-sight velocities and projected positions) and map these to a final numerical prediction (halo mass). Support Distribution Machines⁵ (SDMs) (Sutherland et al. 2012) are chosen as candidates to solve this problem for their virtue of learning from a distribution and predicting a scalar. In this section, we present the Support Distribution Machines framework, implementation, and the results of this machine learning approach applied to halo mass prediction.

⁵ <https://github.com/dougalsutherland/py-sdm>

4.1. SDM Theory

SDMs are built upon Support Vector Machines (SVMs). SVM is a machine learning method that, in its simplest form, takes a set of training example data, with data vectors from which to learn, and divides them such that training data with similar labels lie on the same side of the boundary. Figure 5 gives a simple illustration of how training examples with two features might be separated by a decision boundary. Test examples are then classified according to which side of the decision boundary they fall on. The algorithm we use here differs from an SVM classifier in two key ways. First, rather than simply classifying test examples into a finite number of categories, the output values are real numbers; this is called support vector regression (SVR). Second, SDMs learn from distributions rather than from data vectors.

The formal description of SVR is as follows (Drucker et al. 1997; Schölkopf & Smola 2002). In this supervised learning problem we have $\{(X_n, Y_n)\}_{n=1}^N$ (input, output) pairs, where $Y_n \in \mathbb{R}$, $X_i \in \mathbb{R}^d$. The primal form of support vector regression is formulated as a convex optimization problem:

$$\begin{aligned} \min \|w\|^2 \\ \text{subject to } \begin{cases} Y_n - w^T \phi(X_n) < \epsilon_{\text{tol}} \\ w^T \phi(X_n) - Y_n < \epsilon_{\text{tol}} \end{cases}, \end{aligned} \quad (9)$$

where $w \in \mathbb{R}^D$, $\phi: \mathbb{R}^d \rightarrow \mathbb{R}^D$ is a user-defined feature map, and $\epsilon_{\text{tol}} > 0$ is a user-defined error tolerance parameter. The intuition behind these equations is that we want to find a linear map with small weights (w) such that, in the training points, the regression error is smaller than the parameter $\epsilon_{\text{tol}} > 0$.

Depending on the data and parameters, these constraints can easily be infeasible, therefore, analogously to the ‘‘soft margin’’ loss function Bennett & Mangasarian (1992), which was adapted to SVM by Cortes & Vapnik (1995), one can introduce slack variables ξ_n, ξ_n^* to relax the equation (9). After introducing these slack variables, we arrive at the following primal convex problem stated by Vapnik (2000):

$$\begin{aligned} \min \left(\|w\|^2 + C \sum_{n=1}^N (\xi_n + \xi_n^*) \right) \\ \text{subject to } \begin{cases} Y_n - w^T \phi(X_n) < \epsilon_{\text{tol}} + \xi_n \\ w^T \phi(X_n) - Y_n < \epsilon_{\text{tol}} + \xi_n^* \\ \xi_n, \xi_n^* \geq 0, \end{cases} \end{aligned} \quad (10)$$

where $C > 0$ is a parameter. Instead of directly solving the primal quadratic problem (10), in many applications it is easier to solve its dual problem instead:

$$\begin{aligned} \max \left(-\frac{1}{2} \sum_{i,j} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(X_i, X_j) \right. \\ \left. - \epsilon_{\text{tol}} \sum_{n=1}^N (\alpha_n + \alpha_n^*) + \sum_{n=1}^N Y_n (\alpha_n - \alpha_n^*) \right) \\ \text{subject to } \sum_{n=1}^N (\alpha_n - \alpha_n^*) = 0, \text{ and } 0 \leq \alpha_n, \alpha_n^* \leq C. \end{aligned} \quad (11)$$

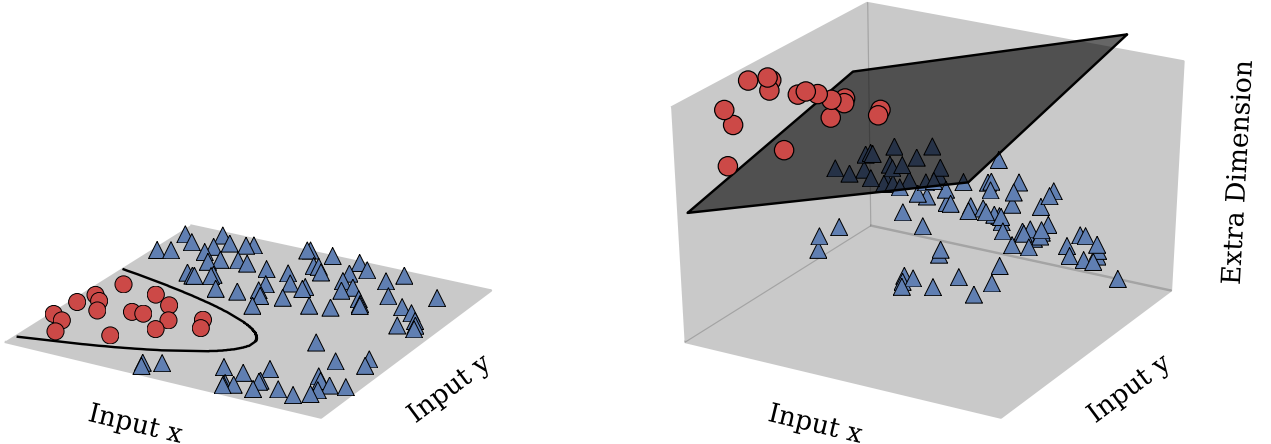


FIG. 5.— In their basic form, Support Vector Machine classifiers search for a hyperplane that separates two classes, here shown as red circles and blue triangles. In the two-feature input space (left panel), the two classes are not separable by a hyperplane. But if the inputs are re-cast into an appropriate higher-dimensional space (right panel), the red circles and blue triangles can be divided by a plane. Test examples can now be classified based on the side of the plane on which they lie; a test example lying above the plane would be classified as a red circle, and one below as a blue triangle. The “decision boundary” which divides the predicted classes (black plane, right panel) corresponds to a different shape in the original space (solid black curve, left panel). We can find the boundary in the original space without explicitly mapping to the higher-dimensional space using the “kernel trick.” Although we wish to predict a continuous-valued label rather than a binary classification, an analogous process applies.

Here, $k(X_i, X_j) \doteq \phi(X_i)^T \phi(X_j)$ is a so-called “kernel function.” The predicted value for a new input X is given by $f(X) = \sum_{n=1}^N (\alpha_n - \alpha_n^*) k(X_n, X)$. A frequently used kernel function is the Gaussian kernel $k(a, b) = \exp(-\|a - b\|^2 / \sigma^2)$, for some $\sigma > 0$ parameter, but any positive semi-definite (PSD) function can be used as kernel $k(\cdot, \cdot)$.

One crucial difference between SVR and our problem is that, in our case, the input X_n is not a finite-dimensional vector, but a distribution with density function p_n . For the kernel value between distributions p_i and p_j , we will use $k(p_i, p_j) \doteq \exp(-KL(p_i, p_j) / \sigma^2)$, where $KL(p_i, p_j) = \int p_i(x) \log(p_i(x) / p_j(x))$ is the Kullback–Leibler (KL) divergence.⁶ In our problem, of course, we do not know these densities exactly; only sample sets are available to us. We will use these sample sets to estimate the KL divergence using the estimator given by Wang et al. (2009). The KL divergence estimate, $KL_{n,m}(X_A || X_B)$, from the feature-space samples X_A from Cluster A (containing n galaxies) to the X_B samples from Cluster B (containing m galaxies) is given by

$$KL_{n,m}(X_A || X_B) = \frac{d}{n} \sum_{i=1}^n \log \frac{\nu_k(i)}{\rho_k(i)} + \log \frac{m}{n-1} \quad (12)$$

where d is the number of dimensions, i.e. the number of distribution features considered in the method, $\nu_k(i)$ is the Euclidean distance in input space from the i^{th} galaxy in X_A to its k -nearest neighbor in X_B and $\rho_k(i)$ is the distance from the i^{th} galaxy in X_A to its k -nearest neighbor in X_A . We use $k = 2$ throughout.

The training catalog is first used to select kernel parameters C and σ via 3-fold cross-validation. It is then

⁶ To make this kernel k PSD, we project the Gram matrix G of the data ($G_{ij} = k(X_i, X_j)$) to the closest PSD matrix in Frobenius norm.

used to train the regression model with the best-selected kernel, which in turn is used to predict the masses of the clusters in the Test Catalog. For further information on the Support Distribution Machine regressor, see Sutherland et al. (2012).

4.2. SDM Implementation

Before applying the SDM regressor, the unique halos are rank-ordered by mass and alternately divided into ten data sets, or folds. The Train Catalog data for nine of these folds is used to train the SDM fit, and the resulting fit is used to predict the Test Catalog halos in the tenth, unused fold. This process is repeated ten times, using multiple projections of 90% of the unique halos to train, and predicting on three LOS views of the remaining 10% of unique halos. In this way, each of the three views of each unique halo in the Test Catalog is predicted.

To explore how both the line-of-sight galaxy velocities as well as their relative positions might affect mass predictions, four sets of training features are considered. Each model uses one or two distribution features, implementing constructs of the sets of v_{los} and/or R_{sub}/R_{200} only. See Table 2 for a summary of the features used in each method. The four features considered are: absolute value of the line-of-sight velocity $|v_{\text{los}}|$, line-of-sight velocity normalized by velocity dispersion $|v_{\text{los}}|/\sigma_v$, projected plane-of-sky position relative to halo radius R_{sub}/R_{200} , and normalized effective angular momentum $|L_{\text{eff}}| = (R_{\text{sub}}/R_{200})|v_{\text{los}}|$.

The first of the four ML feature sets (ML1, with $|v_{\text{los}}|$ only) is chosen to mimic the data used in the $M(\sigma_v)$ power law (PL1), using only line-of-sight velocities to predict halo mass. However, as was shown by Evrard et al. (2008), $\approx 10\%$ of halos are merger transients. As an example, a line of sight merger of dissimilar-mass halos would exhibit notable skewness in the line-of-sight velocity PDF, with the sign of the skewness being indicative of whether the less massive halo was in the foreground

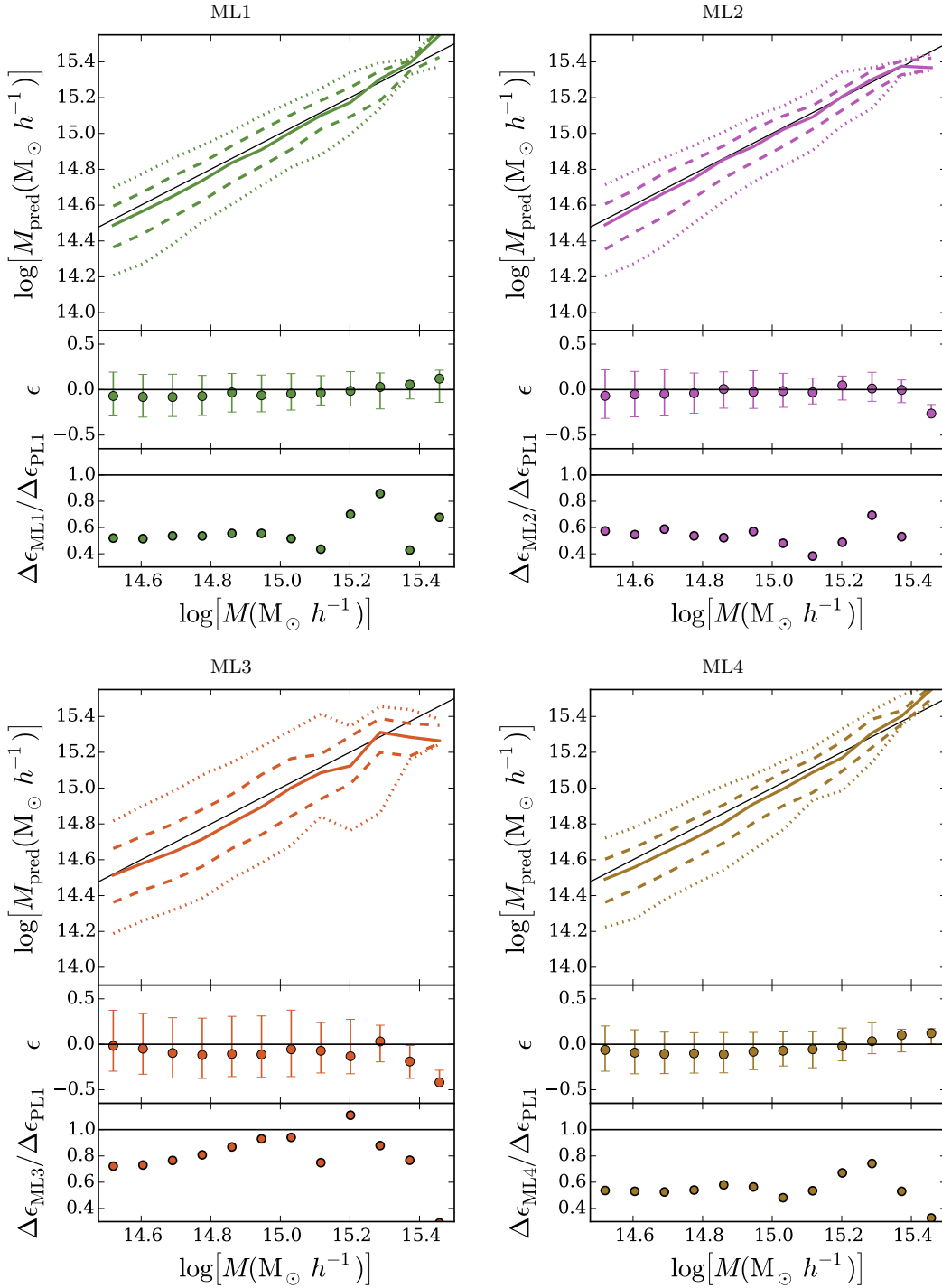


FIG. 6.— Summary of the SDM results. Four large panels show the four machine learning approaches: ML1 (top left, green), ML2 (top right, purple), ML3 (bottom left, orange), ML4 (bottom right, brown). Within each large panel, three sub panels show: Top panel: Halo predicted mass, M_{pred} , vs. actual mass, M . The binned predicted mass median, 68%, and 95% scatter (solid, dashed, and dotted, respectively) are also shown. Middle panel: Fractional mass error as a function of halo mass. Points are the median ϵ and error bars show 68% scatter. Bottom panel: Error width, $\Delta\epsilon_{\text{ML}}$ relative to that of PL1 (virial-theorem-motivated power law). Three methods—ML1, ML2, and ML4—show pronounced improvements over the $M(\sigma_v)$ power law (PL1).

or background of the more massive one. In light of this, $|v_{\text{los}}|$ is chosen as a feature over v_{los} for its virtue of making positive and negative skewness in the velocity PDF appear as identical systems. ML1 mimics PL1, though going further by now predicting from the full LOS velocity distribution rather than a single summary statistic.

ML2 also uses $|v_{\text{los}}|$, but with the addition of the second feature, $|v_{\text{los}}|/\sigma_v$. The aim of explicitly normalizing the line-of-sight velocity distribution in this second feature is to highlight differences in v_{los} PDF shapes, mimicking the κ -corrected power law that was explored in PL2.

The third and fourth ML feature sets employ addi-

tional galaxy position information. Knowing that R_{200} correlates with halo mass, we choose R_{sub}/R_{200} as a way to utilize the relative distribution of galaxies without biasing the results by training on R_{200} . ML3 uses $|v_{\text{los}}|$ and R_{sub}/R_{200} as two separate features, while ML4 combines them into one feature, a normalized effective angular momentum $|L_{\text{eff}}|$.

4.3. Results

Figure 6 compares the predicted and actual masses for the clusters in our catalog, as well as an error comparison to PL1, the $M(\sigma_v)$ power law. The bottom subpanel of each method comparison in the figure shows that the 68% scatter is substantially decreased compared to PL1.

The addition of $|v_{\text{los}}|/\sigma_v$ in PL2 highlights the difference in v_{los} PDF shape, much like the application of kurtosis did in PL2. This may explain why, for most mass bins, the mean ϵ is closer to zero, i.e. the bias is smaller compared to ML1 with the addition of this additional feature.

However, comparing ML1 to ML3 with the addition of R_{sub}/R_{200} as a second feature, the scatter in ϵ increases. To deduce the causes of this, we explore several variations to the ML3 method for comparison. We find that ML3 performs only slightly better than a feature set that uses $|v_{\text{los}}|$ and a random number in place of the R_{sub}/R_{200} value ($\epsilon \pm \Delta\epsilon = -0.04_{-0.31}^{+0.45}$). From this, we conclude that the feature R_{sub}/R_{200} , while containing information about the halo, tends to wash out the more important distribution of $|v_{\text{los}}|$ in feature space. When the normalized R_{sub}/R_{200} is replaced with an unnormalized R_{sub} , the result is similar to ML1 and ML2: $\epsilon \pm \Delta\epsilon = -0.001_{-0.22}^{+0.27}$. The quality of this fit is unsurprising because the maximum R_{sub} value is strongly correlated with R_{200} and, therefore, with halo mass. Despite the apparent failure of ML3, we remain optimistic that this feature could be of use when applied to a very large training catalog, when an appropriate simulation becomes available.

ML methods will tend to underpredict the most massive halo because the training set will include only halos less massive than this one outlier. This effect is plainly evident for ML2 and ML3. Because the mass predictions will tend to lie within the range of the training set masses, one should not overly interpret the prediction of the largest mass bin in Figure 6, as it contains only the single most massive halo.

5. METHOD COMPARISON & DISCUSSION

In this section, we will compare the six cluster mass prediction methods, using two different measures of comparison: ϵ averaged across all clusters and ϵ as a function of mass.

In Figure 7, which is a PDF of fractional mass errors, improvements in mass predictions are evident as we use more information from the LOS velocity PDF. The addition of κ as a predictor for residual δ in PL2 decreases the number of extreme overpredicted cluster masses and moderately improves mass prediction. But the machine learning methods ML1 and ML2 significantly improve the accuracy of mass predictions. With machine learning, the $\epsilon \gtrsim 0.6$ and $\epsilon \lesssim -0.6$ predictions are all but eliminated. Machine learning clearly dominates by this

measure: averaged across all halos in the catalog, ML's cluster mass predictions are significantly improved over traditional power law predictions.

Table 3 summarizes the mean fractional error ($\bar{\epsilon}$), median fractional error with 68% scatter ($\epsilon \pm \Delta\epsilon$), and width of the distributions ($\Delta\epsilon$) for both the Test and High-Mass Test Catalogs. In this table, the mean and median ϵ quantify the bias: whereas the power law methods err on the side of overpredicting cluster masses, the machine learning methods err on the side of underprediction. The $\Delta\epsilon$ value is a measure of the widths of the curves in Figure 7. While PL1 for the Test Catalog has $\Delta\epsilon = 0.87$, PL2 decreases the width of the errors to 0.74, a 16% improvement over the $M(\sigma_v)$ power law method. Machine learning methods improve $\Delta\epsilon$ further: ML3 ($\Delta\epsilon = 0.68$) improves 22% over the PL methods, while ML1 ($\Delta\epsilon = 0.46$), ML2 ($\Delta\epsilon = 0.48$), and ML4 ($\Delta\epsilon = 0.47$) have an even narrower distribution of ϵ values. ML1, ML2, and ML4 have 47%, 44% and 46% improvements, respectively. Defining a fractional log mass error as $\epsilon_{\text{log}} = [\log(M_{\text{pred}}) - \log(M)]/\log(M)$ in lieu of the fractional mass error ϵ produces similar comparative results.

Figure 8 is a direct comparison of the methods across mass bins. The comparison is quantified by

$$|\epsilon_{\text{row}}| - |\epsilon_{\text{column}}|, \quad (13)$$

where ϵ_{column} is the fractional mass error of the method indicated by the column label and ϵ_{row} is the fractional mass error of the row. A $|\epsilon_{\text{row}}| - |\epsilon_{\text{column}}|$ value below zero is indicative of the row method predicting halo mass more accurately than the column method.

The left column of Figure 8 is a comparison to the $M(\sigma_v)$ power law, and the mean values of this comparison to PL1 are summarized in Table 3. PL2 improves upon PL1 at all masses in the range considered, with an average $|\epsilon_{\text{PL2}}| - |\epsilon_{\text{PL1}}| = -0.06$. ML1 outperforms PL1 at all masses as well, but with a significantly smaller $|\epsilon_{\text{ML1}}| - |\epsilon_{\text{PL1}}| = -0.17$. ML2 and ML4 also improve on PL1 at all masses, while ML3's improvement in mass predictions is the most pronounced at low masses. At the highest masses, ML3 consistently underpredicts. Recall, however, that with machine learning methods, mass predictions typically lie within the range of the training set masses. Therefore, we expect to see an underprediction for the most massive halo. Table 3 summarizes the $|\epsilon_{\text{ML}}| - |\epsilon_{\text{PL1}}|$ values for each method.

The ML3 model underperforms at predicting most masses. This may be attributed to the inclusion of R/R_{200} , which washes out the more important $|v_{\text{los}}|$ feature, causing the significant underprediction of mass for much of the mass range considered. ML3 presents a cautionary tale: including additional distribution features with the SDM algorithm will not necessarily improve mass predictions, therefore, features should be chosen with care.

Because of its poor predictive power at most masses, ML3 is identified as a disfavored method. Both ML1 and ML2, utilizing only constructs of the line-of-sight galaxy velocities, are our preferred machine learning methods. Each has its own strength: ML1 outperforms ML2 in the measure of $\Delta\epsilon$ for the Test Catalog, whereas ML2 has a smaller $\Delta\epsilon$ for the High-Mass Test Catalog, outperforms in median ϵ , and also minimizes the tendency to under-

TABLE 3
 METHOD COMPARISON

Case	Summary	Color	Catalog	$\bar{\epsilon}^1$	$\epsilon \pm \Delta\epsilon^2$	$\Delta\epsilon^3$	$ \epsilon - \epsilon_{\text{PL1}} ^4$
PL1	$M(\sigma_v)$ Power Law	Red	Test	0.128	$0.05^{+0.51}_{-0.36}$	0.871	—
			High-Mass Test	0.093	$0.02^{+0.44}_{-0.29}$	0.731	—
PL2	$M(\sigma_{v,\kappa\text{-corrected}})$ Power Law	Blue	Test	0.079	$0.04^{+0.41}_{-0.33}$	0.735	-0.06
			High-Mass Test	0.058	$0.02^{+0.34}_{-0.27}$	0.612	-0.06
ML1	SDM with $ v_{\text{los}} $	Green	Test	-0.055	$-0.07^{+0.25}_{-0.22}$	0.460	-0.17
			High-Mass Test	-0.042	$-0.03^{+0.22}_{-0.18}$	0.402	-0.14
ML2	SDM with $ v_{\text{los}} $ & $ v_{\text{los}} /\sigma_v$	Purple	Test	-0.038	$-0.05^{+0.25}_{-0.24}$	0.484	-0.16
			High-Mass Test	-0.005	$-0.01^{+0.20}_{-0.19}$	0.386	-0.16
ML3	SDM with $ v_{\text{los}} $ & R_{sub}/R_{200}	Orange	Test	0.005	$-0.06^{+0.40}_{-0.28}$	0.679	-0.08
			High-Mass Test	-0.017	$-0.10^{+0.40}_{-0.25}$	0.651	-0.02
ML4	SDM with $ L_{\text{eff}} $	Brown	Test	-0.066	$-0.08^{+0.25}_{-0.22}$	0.468	-0.16
			High-Mass Test	-0.063	$-0.07^{+0.21}_{-0.20}$	0.410	-0.13

^aMean fractional mass error.

^bMedian fractional mass error \pm 68% scatter.

^cWidth of ϵ 68% scatter.

^dComparison of model to PL1; see equation 13.

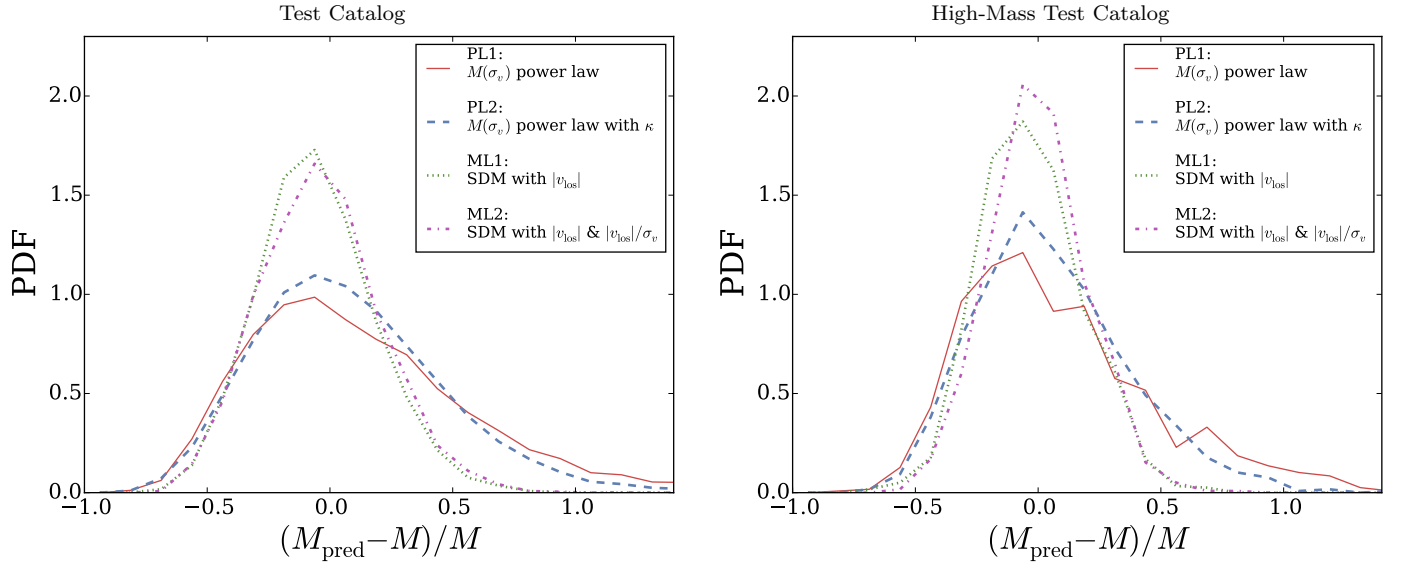


FIG. 7.— PDF of the fractional mass error, $\epsilon = (M_{\text{pred}} - M)/M$, for four methods: $M(\sigma_v)$ power law (PL1, red solid), κ -corrected power law (PL2, blue dashed), machine learning with $|v_{\text{los}}|$ (ML1, green dotted), and machine learning with $|v_{\text{los}}|$ and $|v_{\text{los}}|/\sigma_v$ (ML2, purple dash-dot). In addition to the full Test Catalog population (left panel), the High-Mass Test Catalog ($M > 7 \times 10^{14} M_{\odot} h^{-1}$, right panel) is also shown. PL2 shows a moderate improvement over PL1, while both ML methods shown outperform both power law methods, showing significantly smaller fractional mass errors.

predict masses that is seen in the other machine learning methods.

As tracers of the most massive dark matter halos, cluster counts as a function of mass and redshift are sensitive to cosmological parameters. Since they contain information about the growth of structure in the low-redshift uni-

verse, cluster richness and abundance have been used to constrain σ_8 , Ω_m , Ω_{DE} , and w (e.g. Schuecker et al. 2003; Henry et al. 2009; Vikhlinin et al. 2009b; Rozo et al. 2010; Mantz et al. 2010; Allen et al. 2011; Planck Collaboration et al. 2014b). For measurements of these types, cluster mass calculations remain a large source of

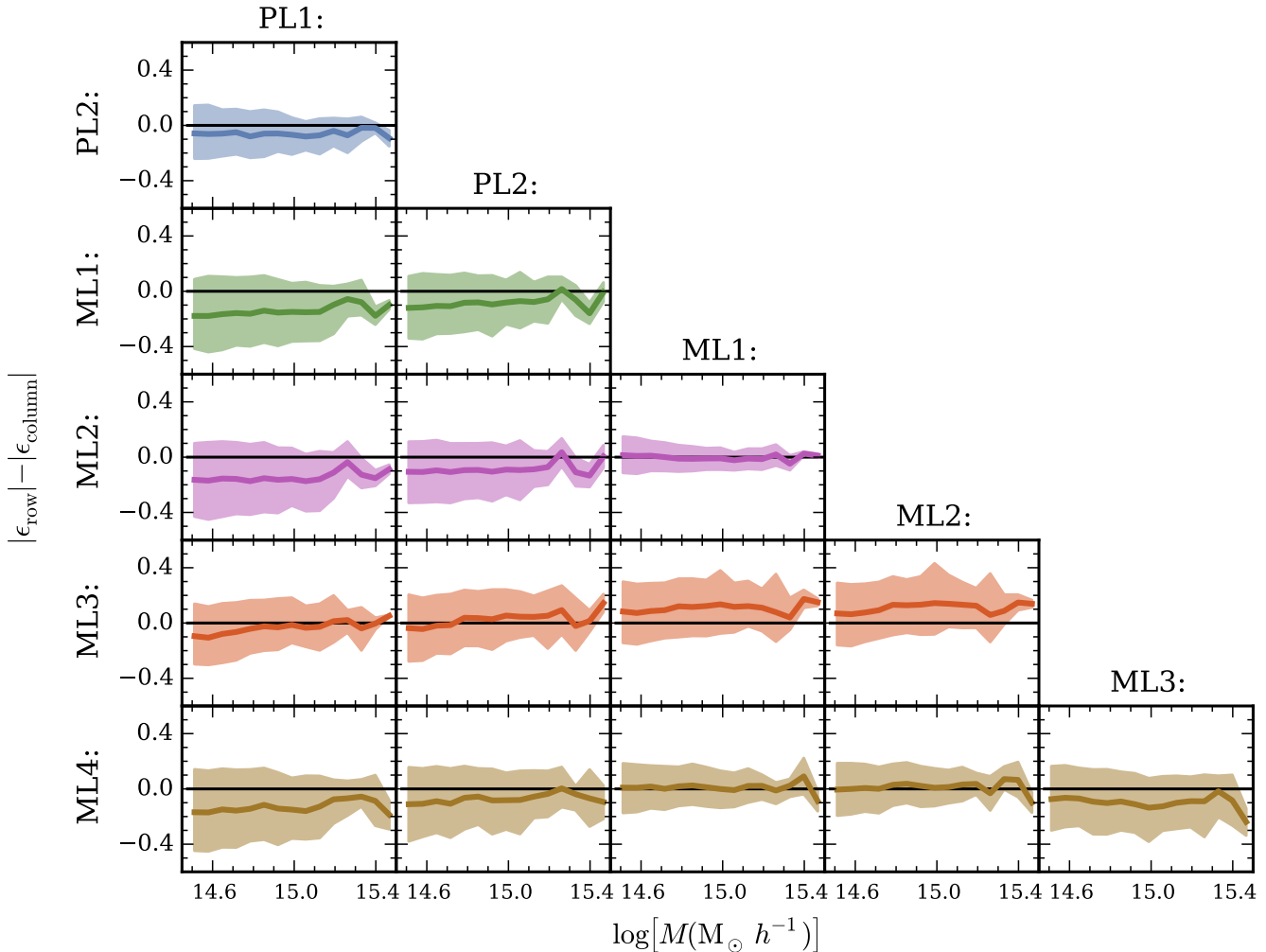


FIG. 8.— Method summary comparison: $|\epsilon_{\text{row}}| - |\epsilon_{\text{column}}|$ as a function of mass. Values below 0 indicate that the row method is performing better than the column method for a given mass bin. The left column summarizes a comparison of the five new methods to the $M(\sigma_v)$ power law: PL2 (the κ -corrected power law) improves mass predictions in all mass bins. While ML3, which includes R_{sub}/R_{200} , performs poorly compared to the other methods over much of the mass range considered, three machine learning methods—ML1, ML2, and ML4 (green, purple, and brown, respectively)—improve or maintain median accuracy of mass predictions at all masses considered.

systematic error. The improved errors, ϵ , accessible using the machine learning technique discussed here could prove to be a powerful tool for addressing this problem. In principle, tighter constraints on scaling relations and the halo mass function could be made with the same number of observed clusters. Alternatively, fewer clusters would be needed in order to have the same amount of constraining power as current techniques. Ultimately, decreasing error in cluster mass measurements should result in more accurate values of cluster properties and cosmological parameters.

6. CONCLUSIONS

We have explored dynamical mass measurements of a catalog of simulated galaxy clusters. We present methods for cluster mass measurements that extract information from the line-of-sight velocity PDF, but improve upon the $M(\sigma_v)$ power law in accuracy. All methods are trained on a catalog of simulated galaxy clusters with mass greater than $10^{14} M_{\odot} h^{-1}$ and tested on those with mass greater than $3 \times 10^{14} M_{\odot} h^{-1}$. The halos used for building the cluster catalog are gleaned from a publicly-

available halo catalog of the Multidark Simulation.

Two power law methods are considered: PL1 employs a standard $M(\sigma_v)$ power law, while PL2 takes advantage of the relationship between residual (δ) and LOS velocity PDF kurtosis (κ), using $\delta(\kappa)$ as a predictor for the amount by which the $M(\sigma_v)$ power law over- or under-predicts halo mass. In addition to the power law method, we explore four machine learning methods, all of which employ Support Distribution Machines, a machine learning algorithm that learns from a distribution and predicts a mass. For the ML methods, four different sets of distribution features are considered; all of these use only line-of-sight velocity, or LOS velocity coupled with normalized galaxy projected sky position, as raw data to construct the distribution features for training.

Our main conclusions can be summarized as follows:

1. Applying the virial-theorem-motivated $M(\sigma_v)$ power law (method PL1) to our cluster catalog results in a fractional mass error width of $\Delta\epsilon = 0.871$.
2. Kurtosis can be used to calculate a correction term

for σ_v , resulting in a κ -corrected $M(\sigma_v)$ power law. In some cases, it does so by identifying halos experiencing infalling matter by the negative κ signature of the velocity PDF and appropriately lowering the prediction of these halos' masses. This method (PL2), decreases the width of fractional mass error to $\Delta\epsilon = 0.735$, a decrease of 16% compared to the $M(\sigma_v)$ power law's predictions.

3. Support Distribution Machines trained on one distribution feature, $|v_{\text{los}}|$, (method ML1) results in $\Delta\epsilon = 0.460$, a decrease of 47% compared to the $M(\sigma_v)$ power law's predictions. When SDM is trained on two distribution features, $|v_{\text{los}}|$ and $|v_{\text{los}}/\sigma_v|$, (method ML2) results in $\Delta\epsilon = 0.484$, a decrease of 44% compared to the $M(\sigma_v)$ power law's predictions. Method ML2 is the preferred machine learning method for minimizing mean ϵ .
4. Two additional SDM methods are tested, and their fractional error (ϵ) values are summarized in Table 3. Generally, these methods outperform both the $M(\sigma_v)$ power law method and kurtosis-corrected power law method (PL1 and PL2, respectively) in terms of the width of fractional mass error $\Delta\epsilon$.

In subsequent work, we will explore several remaining challenges in applying machine learning for dynamical mass measurements. As discussed in Old et al. (2014), mass estimators perform best under the conditions for which they are calibrated; models that are calibrated to predict cluster mass in spite of the presence of interlopers actually tend to do worse when interlopers are entirely excluded from the sample and the true cluster membership is known. Because the ultimate aim is to calibrate on simulations to predict masses of observed clusters, realistic mock cluster catalogs that include known obser-

vational selection effects will have to be constructed and analyzed. Examples of major changes are: (1) including higher-redshift members to more accurately reflect the sample that will be probed by upcoming surveys, (2) assigning galaxy luminosities to subhalos and applying a galaxy luminosity cut rather than a subhalo mass cut, (3) allowing for galaxies to be offset from the center of the halo or subhalo host, and (4) incorporating observational selection criteria such as a fixed aperture with a LOS velocity cut that allows for interlopers. Before this method can be applied to observation, there remains a need to train SDM on an observationally-aligned cluster catalog built from a large-volume, high-resolution simulation. Once such a simulation becomes available, Support Distribution Machines will be a powerful tool to predict cluster masses.

We thank our referee **Ramin Skibba** for his **constructive comments and review of this manuscript**. We also thank Alex Geringer-Sameth, Shirley Ho, Paul La Plante, Rachel Mandelbaum, and Ying Zu for their valuable feedback. This work is supported in part by DOE de-sc0011114 grant. The CosmoSim database used in this paper is a service by the Leibniz-Institute for Astrophysics Potsdam (AIP). The MultiDark database was developed in cooperation with the Spanish MultiDark Consolider Project CSD2009-00064. The Bolshoi and MultiDark simulations have been performed within the Bolshoi project of the University of California High-Performance AstroComputing Center (UC-HiPACC) and were run at the NASA Ames Research Center. The MultiDark-Planck (MDPL) and the BigMD simulation suite have been performed in the Supermuc supercomputer at LRZ using time granted by PRACE.

REFERENCES

- Abell, G. O. 1958, *ApJS*, 3, 211
- Allen, S. W., Evrard, A. E., & Mantz, A. B. 2011, *ARA&A*, 49, 409
- Andreon, S. 2010, *MNRAS*, 407, 263
- Ascaso, B., Wittman, D., & Benítez, N. 2012, *MNRAS*, 420, 1167
- Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2007, *ApJ*, 663, 774
- Banerji, M., Lahav, O., Lintott, C. J., et al. 2010, *MNRAS*, 406, 342
- Beers, T. C., Flynn, K., & Gebhardt, K. 1990, *AJ*, 100, 32
- Bennett, K. P., & Mangasarian, O. L. 1992, *Optimization Methods and Software*, 1, 23
- Biviano, A., & Girardi, M. 2003, *ApJ*, 585, 205
- Biviano, A., Murante, G., Borgani, S., et al. 2006, *A&A*, 456, 23
- Brodwin, M., Ruel, J., Ade, P. A. R., et al. 2010, *ApJ*, 721, 90
- Carlberg, R. G., Yee, H. K. C., & Ellingson, E. 1997, *ApJ*, 478, 462
- Cohn, J. D. 2012, *MNRAS*, 419, 1017
- Colak, T., & Qahwaji, R. 2009, *Space Weather*, 7, 6001
- Colless, M., & Dunn, A. M. 1996, *ApJ*, 458, 435
- Cortes, C., & Vapnik, V. 1995, *Machine learning*, 20, 273
- Diemand, J., Moore, B., & Stadel, J. 2004, *Mon.Not.Roy.Astron.Soc.*, 352, 535
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. 1997, in *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 9* (MIT Press), 155–161
- Evrard, A. E., Bialek, J., Busha, M., et al. 2008, *Astrophys.J.*, 672, 122
- Fadda, D., Girardi, M., Giuricin, G., Mardirossian, F., & Mezzetti, M. 1996, *ApJ*, 473, 670
- Faltenbacher, A., Kravtsov, A. V., Nagai, D., & Gottloeber, S. 2005, *Mon.Not.Roy.Astron.Soc.*, 358, 139
- Faltenbacher, A., & Mathews, W. G. 2007, *Mon.Not.Roy.Astron.Soc.*, 375, 313
- Gerdes, D. W., Sypniewski, A. J., McKay, T. A., et al. 2010, *ApJ*, 715, 823
- Girardi, M., Giuricin, G., Mardirossian, F., Mezzetti, M., & Boschin, W. 1998, *ApJ*, 505, 74
- Gladders, M. D., & Yee, H. K. C. 2005, *ApJS*, 157, 1
- Hajian, A., Alvarez, M., & Bond, J. R. 2014, *ArXiv e-prints*, arXiv:1409.1576
- Hao, J., McKay, T. A., Koester, B. P., et al. 2010, *ApJS*, 191, 254
- Hasselfield, M., Hilton, M., Marriage, T. A., et al. 2013, *J. Cosmology Astropart. Phys.*, 7, 8
- Henry, J. P., Evrard, A. E., Hoekstra, H., Babul, A., & Mahdavi, A. 2009, *ApJ*, 691, 1307
- Johnston, D. E., Sheldon, E. S., Wechsler, R. H., et al. 2007, *ArXiv e-prints*, arXiv:0709.1159
- Klypin, A., & Holtzman, J. 1997, *ArXiv Astrophysics e-prints*, astro-ph/9712217
- Lau, E. T., Nagai, D., & Kravtsov, A. V. 2010, *ApJ*, 708, 1419
- Lokas, E. L., & Mamon, G. A. 2003, *MNRAS*, 343, 401
- Lokas, E. L., Wojtak, R., Gottlöber, S., Mamon, G. A., & Prada, F. 2006, *MNRAS*, 367, 1463
- Lueker, M., Reichardt, C. L., Schaffer, K. K., et al. 2010, *ApJ*, 719, 1045
- Mahabal, A., Djorgovski, S. G., Turmon, M., et al. 2008, *Astronomische Nachrichten*, 329, 288
- Mamon, G. A., Biviano, A., & Murante, G. 2010, *A&A*, 520, A30

- Mandelbaum, R., Seljak, U., & Hirata, C. M. 2008, *J. Cosmology Astropart. Phys.*, 8, 6
- Mantz, A., Allen, S. W., Ebeling, H., Rapetti, D., & Drlica-Wagner, A. 2010, *MNRAS*, 406, 1773
- Marriage, T. A., Acquaviva, V., Ade, P. A. R., et al. 2011, *ApJ*, 737, 61
- Munari, E., Biviano, A., Borgani, S., Murante, G., & Fabjan, D. 2013, *MNRAS*, 430, 2638
- Noh, Y., & Cohn, J. D. 2012, *MNRAS*, 426, 1829
- Old, L., Gray, M. E., & Pearce, F. R. 2013, *MNRAS*, 434, 2606
- Old, L., Skibba, R. A., Pearce, F. R., et al. 2014, *MNRAS*, 441, 1513
- Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014a, *A&A*, 571, A16
- . 2014b, *A&A*, 571, A20
- . 2014c, *A&A*, 571, A29
- Ribeiro, A. L. B., Lopes, P. A. A., & Trevisan, M. 2011, *MNRAS*, 413, L81
- Rines, K., Geller, M. J., & Diaferio, A. 2010, *ApJ*, 715, L180
- Rosati, P., Borgani, S., & Norman, C. 2002, *ARA&A*, 40, 539
- Rozo, E., Wechsler, R. H., Rykoff, E. S., et al. 2010, *ApJ*, 708, 645
- Ruel, J., Bazin, G., Bayliss, M., et al. 2013, *ArXiv e-prints*, arXiv:1311.4953
- Saro, A., Mohr, J. J., Bazin, G., & Dolag, K. 2013, *ApJ*, 772, 47
- Schölkopf, B., & Smola, A. J. 2002, *Learning with kernels: support vector machines, regularization, optimization, and beyond* (MIT press)
- Schuecker, P., Böhringer, H., Collins, C. A., & Guzzo, L. 2003, *A&A*, 398, 867
- Serra, A. L., Diaferio, A., Murante, G., & Borgani, S. 2011, *MNRAS*, 412, 800
- Sheth, R. K., & Diaferio, A. 2011, *MNRAS*, 417, 2938
- Sifón, C., Menanteau, F., Hasselfield, M., et al. 2013, *ApJ*, 772, 25
- Skibba, R. A., van den Bosch, F. C., Yang, X., et al. 2011, *MNRAS*, 410, 417
- Spitzer, Jr., L. 1969, *ApJ*, 158, L139
- Staniszewski, Z., Ade, P. A. R., Aird, K. A., et al. 2009, *ApJ*, 701, 32
- Stepinski, T., Mendenhall, M., & Bue, B. 2009, *icarus*, 203, 77
- Sunyaev, R. A., & Zeldovich, Y. B. 1972, *Comments on Astrophysics and Space Physics*, 4, 173
- Sutherland, D. J., Xiong, L., Pócosz, B., & Schneider, J. 2012, *ArXiv e-prints*, arXiv:1202.0302
- Svensmark, J., Wojtak, R., & Hansen, S. H. 2014, *ArXiv e-prints*, arXiv:1405.0284
- Teague, P. F., Carter, D., & Gray, P. M. 1990, *ApJS*, 72, 715
- Vapnik, V. 2000, *The nature of statistical learning theory* (springer)
- Vikhlinin, A., Burenin, R. A., Ebeling, H., et al. 2009a, *ApJ*, 692, 1033
- Vikhlinin, A., Kravtsov, A. V., Burenin, R. A., et al. 2009b, *ApJ*, 692, 1060
- Voit, G. M. 2005, *Reviews of Modern Physics*, 77, 207
- Wang, Q., Kulkarni, S., & Verdu, S. 2009, *Information Theory, IEEE Transactions on*, 55, 2392
- White, M., Cohn, J., & Smit, R. 2010, *MNRAS* (2010) 408, 1818, 1005.3022
- Wojtak, R. 2013, *A&A*, 559, A89
- Woźniak, P. R., Williams, S. J., Vestrand, W. T., & Gupta, V. 2004, *AJ*, 128, 2965
- Xu, X., Ho, S., Trac, H., et al. 2013, *ApJ*, 772, 147
- Yu, D., Huang, X., Wang, H., & Cui, Y. 2009, *Sol. Phys.*, 255, 91
- Zu, Y., & Weinberg, D. H. 2013, *MNRAS*, 431, 3319
- Zwicky, F. 1933, *Helvetica Physica Acta*, 6, 110
- Zwicky, F., Herzog, E., & Wild, P. 1968, *Catalogue of galaxies and of clusters of galaxies* (Pasadena, CA: California Institute of Technology)