The Astrophysical Journal, 831:135 (16pp), 2016 November 10

doi:10.3847/0004-637X/831/2/135

DYNAMICAL MASS MEASUREMENTS OF CONTAMINATED GALAXY CLUSTERS USING MACHINE LEARNING

M. Ntampaka[1], H. Trac[1], D. J. Sutherland[2], S. Fromenteau[1], B. Póczos[2], and J. Schneider[2]

[1] McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA; ntampaka@cmu.edu

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

Received 2015 September 30; revised 2016 June 16; accepted 2016 June 16; published 2016 November 3

ABSTRACT

We study dynamical mass measurements of galaxy clusters contaminated by interlopers and show that a modern machine learning algorithm can predict masses by better than a factor of two compared to a standard scaling relation approach. We create two mock catalogs from Multidark's publicly available *N*-body MDPL1 simulation, one with perfect galaxy cluster membership information and the other where a simple cylindrical cut around the cluster center allows interlopers to contaminate the clusters. In the standard approach, we use a power-law scaling relation to infer cluster mass from galaxy line-of-sight (LOS) velocity dispersion. Assuming perfect membership knowledge, this unrealistic case produces a wide fractional mass error distribution, with a width of $\Delta\epsilon \approx 0.87$. Interlopers introduce additional scatter, significantly widening the error distribution further ($\Delta\epsilon \approx 2.13$). We employ the support distribution machine (SDM) class of algorithms to learn from distributions of data to predict single values. Applied to distributions of galaxy observables such as LOS velocity and projected distance from the cluster center, SDM yields better than a factor-of-two improvement ($\Delta\epsilon \approx 0.67$) for the contaminated case. Remarkably, SDM applied to contaminated clusters is better able to recover masses than even the scaling relation approach applied to uncontaminated clusters. We show that the SDM method more accurately reproduces the cluster mass function, making it a valuable tool for employing cluster observations to evaluate cosmological models.

Key words: cosmology: theory – dark matter – galaxies: clusters: general – galaxies: kinematics and dynamics – gravitation – large-scale structure of universe – methods: statistical

1. INTRODUCTION

Galaxy clusters are the most massive gravitationally bound systems in the universe. They are dark matter dominated, and have halos of mass $\gtrsim 10^{14}\ M_{\odot}\ h^{-1}$. The majority of multiple-wavelength observations do not directly probe the dark matter distribution, but the baryonic component of clusters: the hot gas and tens to thousands of galaxies contained within the halo. Clusters have complex substructure and internal dynamics, and grow through hierarchical merging and the accretion of matter from the cosmic web. Cluster abundance as a function of mass and redshift is sensitive to the underlying dark matter and dark energy content of the universe and can be used to test cosmological models. See Voit (2005) and Allen et al. (2011) for a review.

While measurements of cluster masses can be employed to constrain cosmological parameters (e.g., Schuecker et al. 2003; Henry et al. 2009; Vikhlinin et al. 2009; Mantz et al. 2010; Rozo et al. 2010; Vanderlinde et al. 2010; Allen et al. 2011; Sehgal et al. 2011; Planck Collaboration et al. 2014b; Mantz et al. 2015), capitalizing on clusters as cosmological probes requires a large, well-defined sample of cluster observations, a connection linking the observations of the baryonic component to the underlying dark matter, and a good understanding of the intrinsic scatter in the mass-observable relationship. A variety of methods connecting observables to cluster mass exist, utilizing observations across multiple wavelengths. A subset of these techniques, broadly labeled dynamical mass measurements, are based on measurements of galaxy kinematics. Dynamical mass measurements utilize line-of-sight (LOS) velocities of the galaxies within the virial radius of the cluster and may also take advantage of the unvirialized matter falling toward the cluster.

The virial theorem approach considers cluster members' LOS velocity dispersion, σ_v. This method scales halo mass, M, with σ_v as a power law and famously led to Zwicky's (1933) discovery of dark matter in the Coma cluster. Dynamical mass measurements based on the virial theorem continue to be used to determine cluster masses (e.g., Brodwin et al. 2010; Rines et al. 2010; Sifón et al. 2013; Ruel et al. 2014; Bocquet et al. 2015). Old et al. (2014) and Old et al. (2015) provide a comparison of several dynamical mass techniques based on galaxy observables. Even when cluster membership is perfectly and fully known, there is scatter in the $M(\sigma_v)$ scaling relation. This can be attributed to both physical effects and selection effects, including halo environment and triaxiality (e.g., White et al. 2010; Saro et al. 2013; Wojtak 2013; Svensmark et al. 2014), projection effects (e.g., Cohn 2012; Noh & Cohn 2012), mass-dependent tidal disruption (e.g., Munari et al. 2013), the degree of relaxedness of the cluster (e.g., Evrard et al. 2008; Ribeiro et al. 2011), and galaxy selection strategy (e.g., Old et al. 2013; Saro et al. 2013; Wu et al. 2013). Halos undergoing mergers or matter accretion possess a telltale wide, flat velocity probability distribution function (PDF) (Ribeiro et al. 2011). Impure, incomplete cluster membership catalogs increase scatter in the $M(\sigma_v)$ relationship further. Reducing errors in cluster mass measurements is essential for applying clusters as cosmological probes.

The galaxy dynamics beyond the virial radius of the cluster is likewise informative, and nearby, unvirialized matter can also be used for cluster mass measurements. The caustic technique employs infalling matter and galaxy velocities to determine a mass profile (e.g., Biviano & Girardi 2003; Serra et al. 2011; Gifford & Miller 2013) and can be applied to determine cluster masses (e.g., Rines & Diaferio 2006; Geller et al. 2013; Rines et al. 2013), performing well even in the case of merging halos (e.g., Rines et al. 2003). Furthermore, the nonvirialized infalling matter beyond the virial radius provides cues that can be used to infer a cluster's mass (e.g., Zu & Weinberg 2013; Falco et al. 2014).

A machine learning (ML) approach to dynamical mass measurements was explored in Ntampaka et al. (2015). Here, we built on the virial theorem's simple $M(\sigma_v)$ power law to take advantage of the entire LOS velocity PDF for mock observations with pure and complete cluster membership information, using all relevant substructure within the R_{200c} of each cluster. Taking full advantage of the velocity PDF was achieved by applying a nonparametric ML approach to a PDF of LOS velocities from a mock cluster catalog. By employing support distribution machines (SDMs), an ML class of algorithms that learns from a distribution to predict a scalar, the full velocity PDF was used to improve mass predictions. A traditional power-law scaling relation yielded a wide fractional mass error distribution (see Equation (3)) and extended high-error tails. SDMs trained on LOS velocities resulted in almost a factor-of-two reduction in mass errors compared to the traditional approach, substantially reducing the number of severely over- and underestimated halo masses in the ideal case with pure and complete cluster membership information.

However, the idealized catalog used in this case did not account for a primary source of error in dynamical mass measurements: interloper galaxies in the fore- or background of the true cluster, appearing to be cluster members. In an ideal cluster catalog, all cluster members are known (complete) and the observations contain only true members (pure). Cluster observations that are impure due to contamination by interlopers are subject to additional scatter in the $M(\sigma_v)$ relationship (e.g., Mamon et al. 2010), and a variety of methods have been developed to remove interloper galaxies from the sample (e.g., Fadda et al. 1996; von der Linden et al. 2007; Mamon et al. 2013; Pearson et al. 2015)

In this follow-up paper, we explore how a more realistically prepared mock catalog influences both the $M(\sigma_v)$ scaling relation as well as the SDM predictions of cluster mass. Cluster members are selected within a cylinder defined by a projected radius in the plane of the sky and a radial velocity along the LOS. This technique produces a catalog of spectroscopic member catalogs that are impure, containing interloping galaxies that appear to be cluster members but do not reside within the virial radius of the cluster. They are also incomplete, excluding some true cluster members from the sample.

In Section 2, we discuss our methods: the simulation (2.1), mock observation (2.2), power-law scaling relation (2.3), and SDM implementation (2.4). Results are presented in Section 3 and discussed in Section 4. We present a summary of our findings in Section 5. Finally, we explore how changes to our mock catalog affect power-law and ML results in the Appendix.

2. METHODS

2.1. Simulation

The mock cluster catalog is created from the publicly available Multidark MDPL1 simulation.[3] Multidark is an N-body simulation containing 3840^3 particles in a box of length $1\ h^{-1}$ Gpc and a mass resolution of $1.51 \times 10^9\, M_\odot\, h^{-1}$. Multidark was run using the L-Gadget2 code. It utilizes a ΛCDM cosmology, 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 $\sigma_8 = 0.82$.

Halos are identified by Multidark's BDMW algorithm, which uses a bound density maximum (BDM) spherical overdensity halo finder with a halo average density equal to 200 times the critical density of the universe, denoted as M. All halos and subhalos at redshift $z = 0$ with mass $M \geqslant 10^{12}\, M_\odot\, h^{-1}$ are included in our sample. For more information on the Multidark simulation and BDMW halo finder, see Klypin & Holtzman (1997), Riebe et al. (2013), Klypin et al. (2014), and references therein.

2.2. Mock Observations

Two mock observations are created: pure and contaminated. For each of these two mock observations, a train sample and a test sample are made. The pure catalog is ideal, in that all cluster members above $M_{\rm sub} = 10^{12}\, M_\odot\, h^{-1}$ within R_{200} are included in the catalog. The train catalog has a flat mass function, with 5028 unique halos with $M \geqslant 10^{14}\, M_\odot\, h^{-1}$. Halos in this catalog each contribute multiple lines of sight, such that low- and high-mass clusters are represented in equal measures. The test catalog has 2278 unique halos with a lower-mass cut of $M \geqslant 3 \times 10^{14}\, M_\odot\, h^{-1}$, and each unique halo contributes exactly three lines of sight. It is discussed in further detail in Ntampaka et al. (2015).

In contrast with the pure catalog, the contaminated catalog includes more realistic observational selection effects. It employs a simple, cylindrical cut around each cluster, allowing interlopers to contaminate the sample. As with the pure catalog, the contaminated catalog has both a train catalog with a flat mass function, as well as a test catalog that uses three lines of sight per cluster.

The contaminated catalog is constructed in the following way. Each halo and subhalo is assumed to represent an observable galaxy, with the galaxy inheriting its host's position and velocity. A simple cut is made around each cluster, allowing for interlopers to contaminate the cluster observation. To allow for interlopers across the box edge, the entire simulation box is padded with a 200 Mpc h^{-1}-thick slice from across the periodic boundary to make a cube with length 1.4 Gpc h^{-1}. This cubic mock observation will be used to create a mock cluster catalog that incorporates known observational selection effects.

An intentionally simplistic cylindrical cut is made around each cluster center. Only halos with $M \geqslant 10^{14}\, M_\odot\, h^{-1}$ with centers that reside within the original 1 Gpc h^{-1} box volume are considered to be "cluster candidates." Following Old et al. (2014), true cluster centers are assumed to be known by the observer. Following Wojtak et al. (2007), the observer is placed 100 Mpc from the center of the cluster along the chosen LOS.

The full 3D galaxy velocity and position information is reduced, then, to what can be observed along this LOS: plane-of-sky x'- and y'-positions and LOS velocities. A galaxy's net velocity, v, is given by the sum of the peculiar velocity plus the Hubble flow. An initial cylindrical cut defined by a circular aperture with radius $R_{\rm aperture}$ about the cluster center in the plane of the sky and an LOS initial velocity cut of $v_{\rm cut}$ about the

[3] http://www.cosmosim.org/

Table 1
Catalog Summary

Catalog Name	Type	Min. Halo Mass ($M_\odot\ h^{-1}$)	$R_{\rm aperture}$ (Mpc h^{-1})	$v_{\rm cut}$ (km s^{-1})	$s_{\rm cut}$	Projections per Unique Halo	Total Projections	σ_{15} (km s^{-1})	α
Pure	Train	1×10^{14}	⋯	⋯	⋯	varies	15000	1244	0.382
Pure	Test	3×10^{14}	⋯	⋯	⋯	3	6834	⋯	⋯
Pure	High-mass Test	7×10^{14}	⋯	⋯	⋯	3	945	⋯	⋯
Contaminated	ML Train	1×10^{14}	1.6	2500	2.0	varies	15000	⋯	⋯
Contaminated	PL Train	3×10^{14}	1.6	2500	2.0	varies	10213	753	0.359
Contaminated	Test	3×10^{14}	1.6	2500	2.0	3	7449	⋯	⋯
Contaminated	High-mass Test	7×10^{14}	1.6	2500	2.0	3	951	⋯	⋯

Note. For the pure catalogs, cluster radius and member galaxies are known. For further details on the creation of this catalog, see Ntampaka et al. (2015).

expected hubble flow velocity of an object located at a distance of 100 Mpc from the observer.

The cylinder $R_{\rm aperture}$ and $v_{\rm cut}$ values are chosen to correspond with the radius and $2\sigma_v$, respectively, of a $1 \times 10^{15}\ M_\odot\ h^{-1}$ cluster. The radius of a cluster of this mass is 1.6 Mpc h^{-1}. The $2\sigma_v$ is informed by the best-fit power law found in Ntampaka et al. (2015), giving twice a typical velocity dispersion of true cluster members of $2\sigma_v \approx 2500$ km s^{-1} for a cluster of mass of $1 \times 10^{15}\ M_\odot\ h^{-1}$. These parameters are noted in Table 1. A more thorough exploration of how $R_{\rm aperture}$ and $v_{\rm cut}$ choices affect cluster mass predictions is presented in the Appendix.

This initial cylinder is pared iteratively in velocity space, with outliers beyond $2\sigma_v$ of the mean velocity being omitted from the sample. Here, σ_v denotes the standard deviation of all LOS velocities of the galaxies that reside in the cylinder. This paring occurs until convergence is reached or until fewer than 20 members remain. Clusters with at least 20 members remaining are added to the cluster catalog.

In order to create a representative training sample of how the rare, high-mass clusters might appear when viewed from any direction, the entire box is rotated and this process is repeated. The first three rotations are chosen so that the observer views along the box x-, y-, and z-directions. The remaining rotations are chosen randomly on the surface of the unit sphere. To create the contaminated train catalog, 1000 such rotations are performed.

The train catalog includes halos with $M \geqslant 1 \times 10^{14}\ M_\odot\ h^{-1}$. It is created with a flat mass function, such that there are exactly 1000 training clusters in each 0.1 dex mass bin. In bins with fewer than 1000 clusters, this is done by assembling many LOS views of rare halos. In mass bins with more than 1000 clusters, clusters are rank-ordered by mass and evenly removed from the training sample.

In contrast with the contaminated train catalog, the contaminated test catalog contains exactly three LOS views of every halo: the box x-, y-, and z-directions. Because boundary effects are expected near the edge of the training sample, a minimum mass cut of $M \geqslant 3 \times 10^{14}\ M_\odot\ h^{-1}$ is applied to the test catalogs. The single most massive halo has a mass that will necessarily lie outside of the training sample, and therefore is omitted from the test catalogs as well.

In summary, the contaminated catalog is created in the following manner.

1. All halos and subhalos with masses greater than $10^{12}\ M_\odot\ h^{-1}$ are assumed to represent a galaxy, with the galaxy inheriting its host's position and velocity.
2. Halos with masses greater than $10^{14}\ M_\odot\ h^{-1}$ are considered "cluster candidates."
3. A cluster candidate's center is assumed to be known, and an observer is placed 100 Mpc from the cluster.
4. All galaxies in the box are given an appropriate velocity that includes both Hubble flow and peculiar velocities.
5. A cylinder is cut around the cluster candidate center; this cylinder is defined by an aperture radius, $R_{\rm aperture}$, and an LOS velocity cut, $v_{\rm cut}$.
6. Galaxies outside of mean galaxy velocity $\pm 2\sigma_v$ are iteratively removed from this cylinder until convergence is reached.
7. This is repeated for all massive halos in the box, and those with at least 20 members remaining are kept in the sample.
8. The box is rotated, and steps 3–7 are repeated.
9. The contaminated train catalog is made of multiple LOS projections, up to 1000 for the highest-mass cluster. The number of projections per unique halo is chosen to create a flat mass function for the train catalog.
10. The contaminated test catalog is made of the first three (x-, y-, and z-directions) views of all halos above $M = 3 \times 10^{14}\ M_\odot\ h^{-1}$. The most massive halo is also excluded from the test catalog.

Figure 1 shows the average $v_{\rm los}$ and R distributions for the train catalogs, divided into three $\log[M\,(M_\odot\ h^{-1})]$ bins. The pure catalog is pure, in that there are no interlopers contaminating the galaxy clusters. It is also complete, in that all galaxies within the cluster R_{200} are known. In contrast, the contaminated catalog includes interlopers and excludes some true cluster members. The shape of $v_{\rm los}$ and R distributions are mass-dependent, and this dependence on cluster mass can be utilized in mass predictions. In Section 2.4, we will explore ways to predict cluster mass by exploiting these mass-dependent distributions using a distribution-to-scalar ML technique.

2.3. Power Law

In a typical power-law scaling relation, one starts with the virial theorem to find a relationship between the velocity dispersion, σ_v, and halo mass, M. This power law is given as $\sigma_v \propto M^{1/3}$, but can be rewritten more generally as

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

where σ_{15} is the typical velocity dispersion of galaxies residing within a $10^{15}\ M_\odot\ h^{-1}$ halo and the parameter α is allowed to vary from the theoretically predicted $\alpha = 1/3$ and is instead fit to data. The best fit is then be used to predict cluster mass from

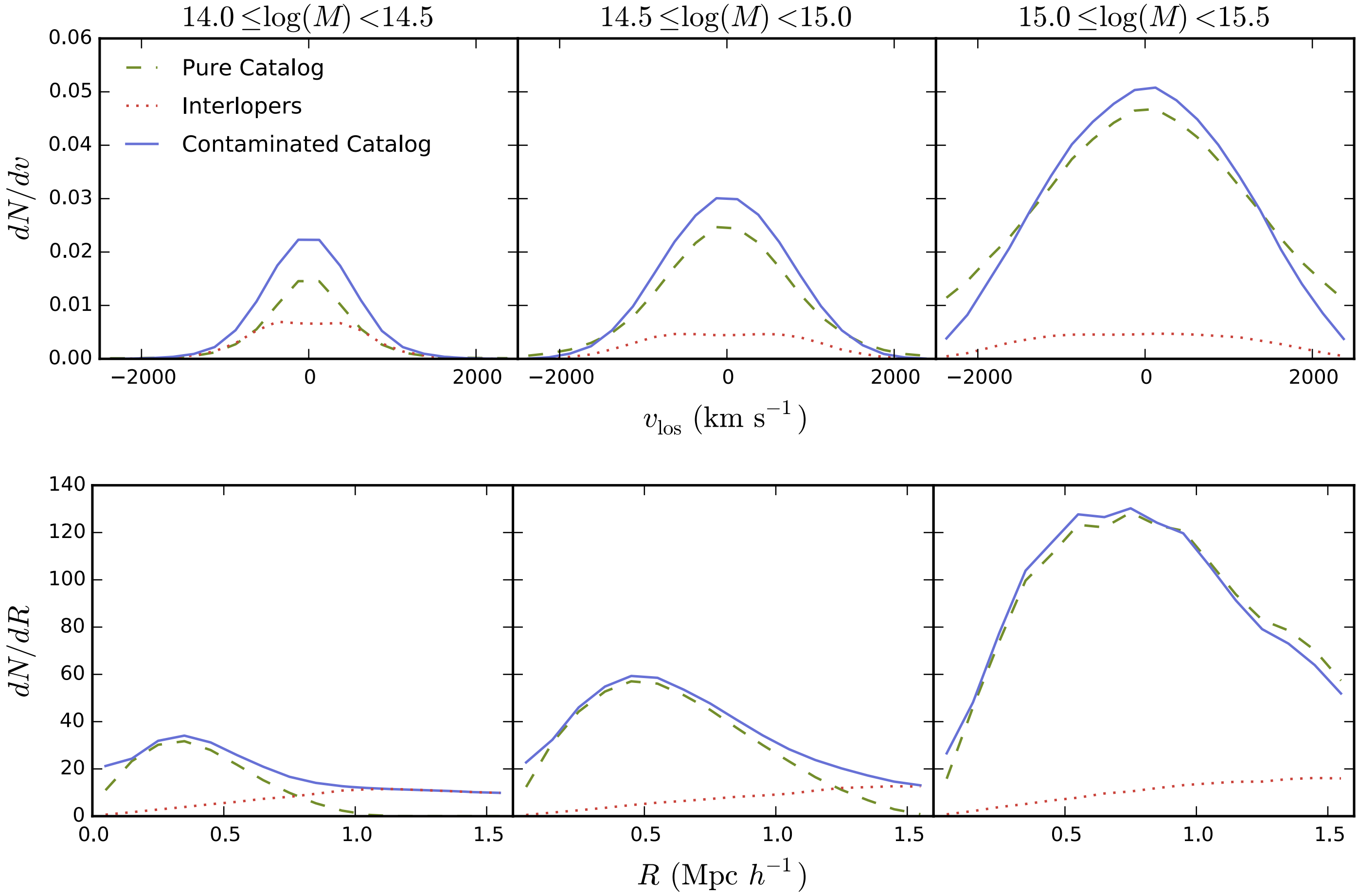

Figure 1. Top: average distribution of galaxy LOS velocities from stacked clusters in three $\log[M\ (M_\odot\ h^{-1})]$ bins, in increasing mass from left to right. While the pure catalog (green dashed) consists solely of galaxies residing within the virial radius of the cluster, the contaminated catalog (blue solid) contains contaminating interlopers (red dotted) and excludes some true cluster members. In the top right panel, the exclusion of true cluster members is evident where the blue solid line dips below the green dashed line. Bottom: average distribution of galaxy projected radii from the cluster center. Both $v_{\rm los}$ and R distributions change shape and amplitude with cluster mass, even for the contaminated catalog; this mass-dependent shape can be exploited by a distribution-to-scalar ML technique to learn cluster masses from distributions of data like the examples shown here.

a velocity dispersion of galaxies. When applied to the pure catalog, this method will be denoted as $\rm PL_P$, and when applied to the contaminated catalog, it will be denoted as $\rm PL_C$.

To account for a potentially changing slope caused by the cylindrical cut used for the contaminated catalog, a lower-mass cut of $3 \times 10^{14} M_\odot\ h^{-1}$ will be applied to the data used to fit the power law. We find a least-squares fit to $\log(\sigma_v) = \alpha \log(M) + \beta$ for the PL train catalog.

While $\rm PL_P$ is well-described by $\alpha = 0.382$, $\sigma_{15} = 1244$ km s^{-1}, $\rm PL_C$ has a shallower slope and smaller velocity dispersion expected for a $10^{15}\ M_\odot\ h^{-1}$ halo, $\alpha = 0.359$ and $\sigma_{15} = 753$ km s^{-1}, respectively. These best-fit parameters to the $M(\sigma_v)$ power law (Equation (1)) for each catalog are noted in Table 1. The scaling relation best fit for the contaminated catalog is shallower and has a smaller σ_{15} compared to that of the pure catalog, therefore, applying the $\rm PL_P$ fit to observed clusters with interlopers can introduce additional errors. We additionally caution that these parameters are a fit for a particular simulation and cylindrical cut and should be applied to observational data with care.

The introduction of interlopers is a large source of scatter in $M(\sigma_v)$. Figure 2 shows a two-dimensional histogram of σ_v versus M for the contaminated catalog. Overlaid is a best fit with 1σ and 2σ lognormal errors calculated for clusters with masses above $3 \times 10^{14}\ M_\odot\ h^{-1}$ and extrapolated down to lower masses. This lognormal scatter, $\sigma_{\rm gauss}$, is determined by the standard deviation of the residual, δ, defined as

$$\delta = \log(\sigma_{\rm measured}) - \log(\sigma_{\rm expected}). \tag{2}$$

Here, $\sigma_{\rm measured}$ is the velocity dispersion of the galaxies within the pared cylinder and $\sigma_{\rm expected}$ is the typical velocity dispersion expected for a cluster of a given mass, found by applying Equation (1) with true cluster mass M and best-fit parameters σ_{15} and α. Of halos with $M \geqslant 3 \times 10^{14}\ M_\odot\ h^{-1}$, 1% reside above the $+2\sigma$ dotted line and 4% reside below the -2σ dotted line. However, of halos with $1 \times 10^{14}\ M_\odot\ h^{-1} \leqslant M < 3 \times 10^{14}\ M_\odot\ h^{-1}$, 8% reside above $+2\sigma$ and 4% below -2σ. The scatter found for the higher-mass clusters is clearly not descriptive of the lower-mass clusters; this is explored further in the Appendix.

The $\rm PL_P$ and $\rm PL_C$ approaches rely on a single summary statistic, σ_v, to describe the dynamics of the cluster members. However, mergers and infalling matter, for example, can distort the shape of the velocity PDF and cause the cluster's mass to be overpredicted by a traditional power-law approach. Next, we will explore a ML approach for predicting cluster masses that learns from a distribution, rather than from a single summary statistic.

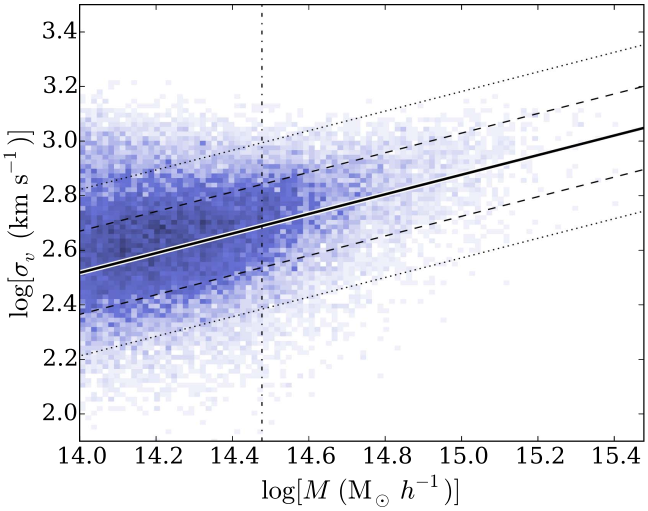

Figure 2. Velocity dispersion, σ_v, vs. cluster mass, M, for a simple cylindrical cut with iterative 2σ paring. Clusters above $3 \times 10^{14}\,M_\odot\,h^{-1}$ (vertical black dash dotted) inform the fit (black solid) and determine the lognormal scatter (68% and 95%, dashed and dotted, respectively). The presence of interlopers introduces significant scatter, particularly at low masses, where the effect of interlopers is more pronounced.

2.4. Support Distribution Machines

SDMs (Sutherland et al. 2012) are a class of ML algorithms built upon support vector machines (SVMs; Drucker et al. 1997; Schölkopf & Smola 2002). Given a training set of (distribution, scalar) pairs, the goal of SDM is to learn a function that predicts a scalar from a distribution. They will be applied here to learn from distributions of galaxy observables such as galaxy LOS velocity and projected distance from cluster center. These distributions of galaxy observables will then be implemented to predict the log of the cluster mass, $\log(M)$.

The SDM method applied requires the divergence between pairs of distributions in the training and test sets. For this purpose, we employ the Kullback–Leibler divergence, and estimate the divergence via the estimator from Wang et al. (2009). This is a k-nearest-neighbor-based estimator. In practice, we use $k = 3$. The relative divergences from training data are used to select SDM best-fit kernel parameters C and σ, the loss function parameter and Gaussian kernel parameter, respectively, via three-fold cross-validation. These are used to train the regression model with the selected best-fit kernel, which in turn is used to predict masses for the test data. For a full discussion of SVM formalism as well as a discussion of how SDM deviates from the SVM base case, see Sutherland et al. (2012) and Ntampaka et al. (2015).

In order to take full advantage of the available data, we cyclically learn from 90% of the clusters and predict masses from the remaining, independent 10%; this is repeated 10 times until the masses of all clusters in the contaminated catalog have been predicted. To prepare the mock cluster catalog for SDM implementation, clusters are rank-ordered by mass and sequentially assigned to one of 10 folds. Multiple LOS views of a unique cluster are all assigned to the same fold, ensuring that each time SDM is implemented, a unique cluster is used either for training or for predicting, but never both.

Of the 10 folds, 9 from the contaminated train catalog are used to select SDM best-fit kernel parameters C and σ and subsequently train the regression model with the selected kernel. This regression model is then used to predict the masses of the clusters in the tenth fold of the contaminated test catalog. The process is repeated 10 times, training on 9 train catalog folds and predicting the tenth test catalog fold, until masses for the entire contaminated test catalog have been predicted.

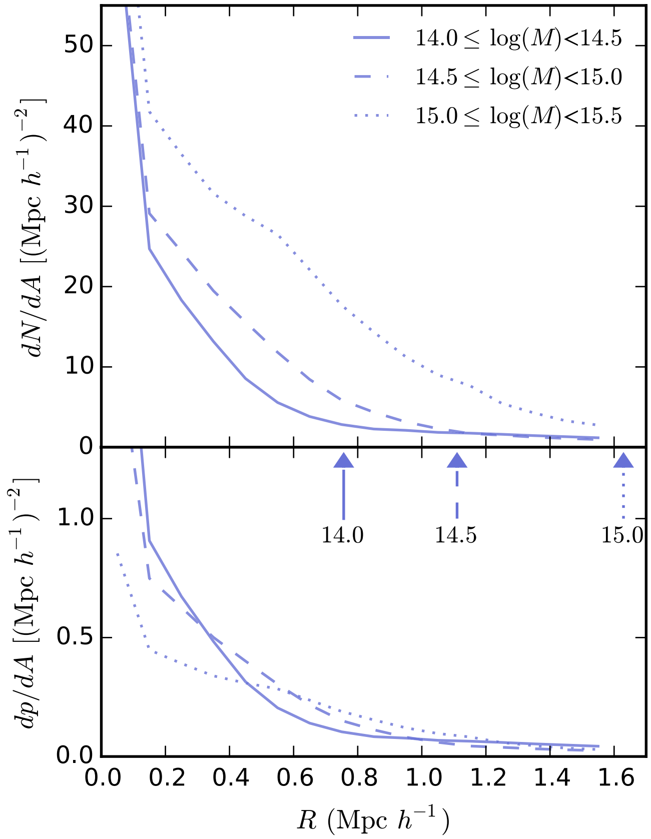

Figure 3. Top: average number of galaxies per unit plane-of-sky area, dN/dA, vs. projected distance from the center of the cluster, R, for three $\log[M\ (M_\odot\ h^{-1})]$ ranges in the contaminated test catalog, in 0.1 Mpc h^{-1} bins. The shape and amplitude of this effective column density vary with the mass of the primary halo. Bottom: probability of finding a galaxy per unit area, dp/dA, vs. R. The shape and amplitude of this measure also varies with primary halo mass. Arrows denote the characteristic radius of a halo with $\log[M\ (M_\odot\ h^{-1})]$ indicated. SDM trained on the feature R takes advantage of how the distribution of subhalo radius changes with mass to predict a halo mass based on the distribution of R.

We implement SDM with four sets of galaxy features: the PDF of galaxy LOS absolute velocity ($|v_{\mathrm{los}}|$), the PDF of normalized velocity ($|v_{\mathrm{los}}|/\sigma_v$), the PDF of projected distance from the cluster center (R), and combinations thereof. As discussed in Ntampaka et al. (2015), features must be chosen with care because features uncorrelated with mass tend to wash out the effects of the more important features. The motivation for features implemented here is as follows.

1. ML$_v$: the use of velocities is motivated by the virial theorem, as we have seen in Figure 2 that velocity dispersion of galaxies, σ_v, relates to mass as a power law, albeit with significant scatter. The ML$_v$ catalog uses absolute value of galaxy LOS velocities, $|v_{\mathrm{los}}|$, as a single feature for training and testing by means of SDM.
2. ML$_R$: even in the presence of interlopers, galaxy density profiles can be used to determine cluster mass (e.g., Hansen et al. 2005; Pearson et al. 2015). This is motivated by Figure 3, which shows stacked halos from

Table 2
Feature Summary

Case	Approach	Train and Test Catalogs	Summary Stats	Distribution Features	Color
$\mathrm{PL_P}$	Power Law	Pure	σ_v	⋯	Red
$\mathrm{PL_C}$	Power Law	Contaminated	σ_v	⋯	Blue
ML_v	Machine Learning: SDM	Contaminated	⋯	$\lvert v_{\mathrm{los}}\rvert$	Green
ML_R	Machine Learning: SDM	Contaminated	⋯	R	Orange
$\mathrm{ML}_{v,R}$	Machine Learning: SDM	Contaminated	⋯	$\lvert v_{\mathrm{los}}\rvert$ and R	Brown
$\mathrm{ML}_{v,\sigma,R}$	Machine Learning: SDM	Contaminated	⋯	$\lvert v_{\mathrm{los}}\rvert$, $\lvert v_{\mathrm{los}}\rvert/\sigma_v$, and R	Purple

the contaminated test catalog divided into three $\log[M\,(M_\odot\,h^{-1})]$ bins. Despite the fixed aperture, the number of galaxies per unit plane-of-sky area (dN/dA) in concentric rings has a markedly different distribution for the low-, middle-, and high-mass halos. The probability of finding a galaxy per unit plane-of-sky area (dp/dA) also exhibits a unique shape for each mass bin. For this reason, we will consider an ML_R catalog, with the galaxy radii from the halo center, R, as the sole feature.

3. $\mathrm{ML}_{v,R}$: decreasing velocity dispersion profiles have been noted in clusters (e.g., Rines et al. 2003). Because v_{los} and R individually can provide information about cluster mass, it seems reasonable that the joint probability distribution of $|v_{\mathrm{los}}|$ and R may be informative as well. $\mathrm{ML}_{v,R}$ will learn from the joint distribution of the LOS velocity feature, $|v_{\mathrm{los}}|$, and the galaxy radius feature, R, in a two-dimensional feature space.
4. $\mathrm{ML}_{v,\sigma,R}$: the shape of the velocity PDF can be indicative of mass accretion and mergers (Evrard et al. 2008; Ribeiro et al. 2011). As found in Ntampaka et al. (2015), explicitly normalizing v_{los} by its width, σ_v, can emphasize these shape differences and improve mass predictions, particularly at the high-mass end. We will consider a training set, $\mathrm{ML}_{v,\sigma,R}$, that employs $|v_{\mathrm{los}}|$, $|v_{\mathrm{los}}|/\sigma_v$, and R in a three-dimensional features space.

These ML method names and corresponding distribution features are summarized in Table 2 for reference and will be used by SDM to predict cluster masses. Next, we will explore how the PL's scaling relation and ML's distribution-to-scalar approach predicted masses of clusters from the mock cluster catalog.

3. RESULTS

3.1. Power Law

Figure 4 shows the predicted versus true cluster masses for the pure and contaminated catalogs. When a power law is applied to the pure catalog, there is significant scatter in mass predictions. The bottom panel of Figure 4 shows the median and 68% scatter in the fractional mass error, ϵ, given by

$$\epsilon = (M_{\mathrm{pred}} - M)/M, \tag{3}$$

where M is the true cluster mass and M_{pred} is the predicted cluster mass. The scatter in $\mathrm{PL_P}$ errors can be attributed to both physical and selection effects. For example, infalling matter tends to create a velocity PDF with negative kurtosis, tending to overpredict the mass. Cluster mergers (Evrard et al. 2008), galaxy selection effects (Saro et al. 2013), and dynamical friction and tidal disruption (Munari et al. 2013) can each play a role in contributing to this scatter.

Figure 4 also shows results for the power-law scaling relation applied to the contaminated catalog. Impure and incomplete clusters introduce further scatter and errors increase significantly. This scatter is most notable at the low-mass end, where the inclusion of interlopers is most prominent.

$\mathrm{PL_P}$ and $\mathrm{PL_C}$ serve as upper and lower bounds for errors for a power-law scaling relation: $\mathrm{PL_P}$'s pure and complete clusters show the level of scatter that remains when interlopers are completely eliminated, while $\mathrm{PL_C}$'s simplistic interloper removal technique highlights how interlopers can affect scatter in an extreme case. More effective interloper removal methods are available, applying more discriminating statistical techniques (e.g., Fadda et al. 1996; von der Linden et al. 2007; Mamon et al. 2013), with some considering only red elliptical galaxies, which preferentially reside in clusters (e.g., Saro et al. 2013). We expect a more refined interloper removal scheme to reside between the two benchmark cases shown in Figure 4.

One may consider the possibility of improving mass predictions by extending mass range for training. However, due to the existence of many high-error, high-σ_v clusters shown in Figure 2, decreasing the lower-mass limit may not improve mass predictions. Even without this high-error population, the power-law dynamical mass approach has significant scatter exacerbated by the presence of interlopers. Furthermore, the potentially informative infalling galaxy observations have not been considered, nor have the baseline LOS velocity PDF shapes indicative of a nonvirialized or merging system. Next, we will explore the results of learning on full distributions with an ML approach.

3.2. Machine Learning

Figure 5 shows the SDM predictions for each of the four feature sets: ML_v, ML_R, $\mathrm{ML}_{v,R}$, and $\mathrm{ML}_{v,\sigma,R}$. As in Figure 4, the top panel shows predicted versus true mass median with 68% and 95% scatter. Each of the ML methods reduces scatter significantly compared to $\mathrm{PL_C}$, the power law that is applied to the same catalog as these ML methods. One should not overly interpret the fluctuations in the two largest mass bins, as they contain only six unique clusters, a small fraction of the total clusters in the sample. The bottom panel shows median error ϵ (see Equation (3)) with 68% scatter. The 68% scatter is dramatically reduced compared to the power-law relation with the same catalog, $\mathrm{PL_C}$, and is comparable to the power-law relation with a catalog of pure and complete clusters, $\mathrm{PL_P}$. $\mathrm{ML}_{v,\sigma,R}$ has median binned mass predictions that are closest to the true mass, while ML_R has the smallest error width, but all four ML methods outperform $\mathrm{PL_C}$ by a large margin.

A comparison of mass predictions is presented in Figure 6. PL provides two benchmarks: while the $\mathrm{PL_C}$ error shows what we might expect from a impure and incomplete interloper

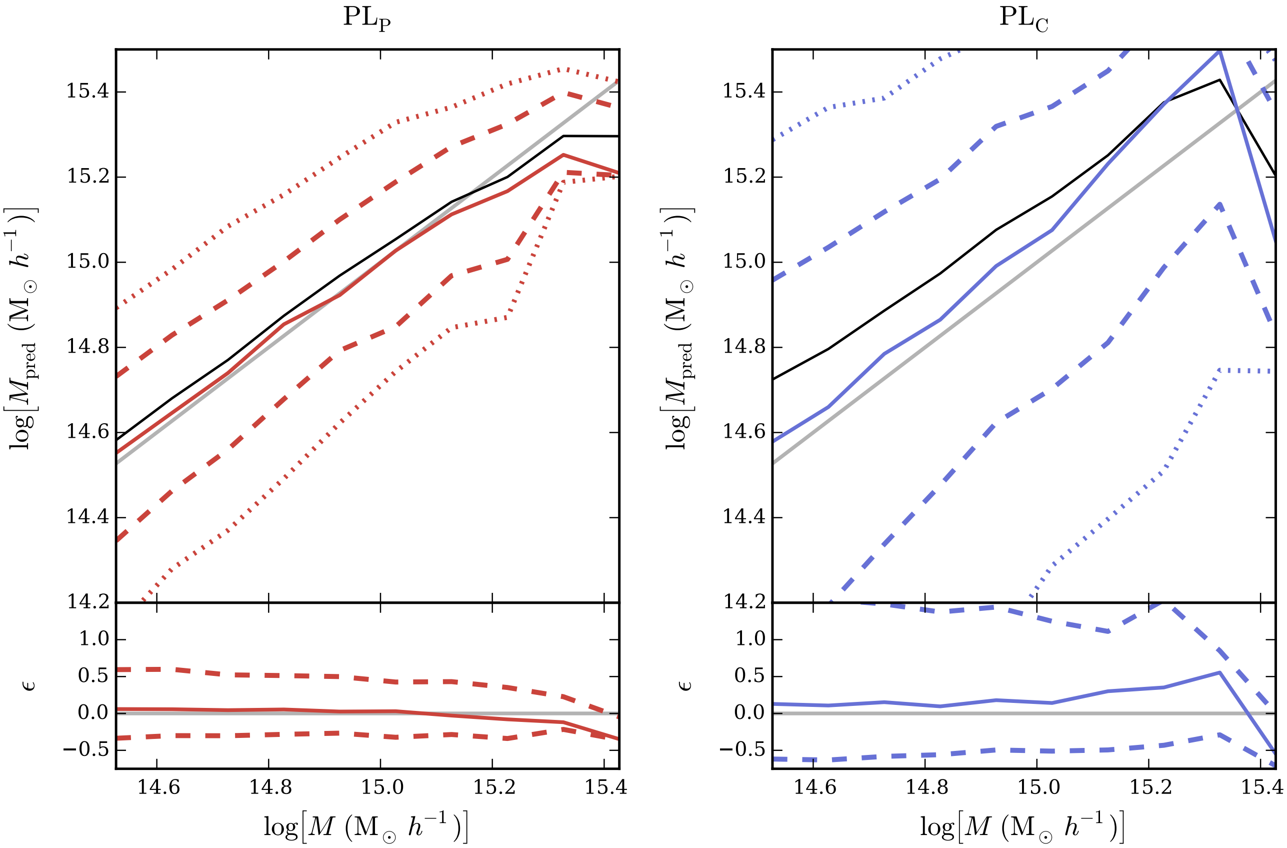

Figure 4. Left: power-law scaling relation applied to the pure catalog (method PL$_P$). Predicted vs. true mass, binned in 0.1 dex $\log[M\ (M_\odot\ h^{-1})]$ bins, with mean (black solid), median (red solid), 68% (dashed), and 95% (dotted) scatter shows that significant scatter exists even when applying a scaling relation to a catalog of pure and complete clusters (top). Though the mass error median (red solid) is nearly zero (gray solid), it has significant 68% scatter (red dashed; bottom). Right: power-law scaling relation applied to the contaminated catalog, which contains impure and incomplete clusters (method PL$_C$). The imperfect catalog introduces additional scatter in ϵ compared to the PL$_P$ case, most notably at low masses where the sample impurity is particularly pronounced. These two plots provide best (left) and worst (right) case scenario benchmarks for applying an $M(\sigma_v)$ power-law scaling relation to cluster observation.

catalog, PL$_P$ gives a best-case scenario where cluster members are perfectly known and interlopers are entirely excluded. Across the entire mass range considered, ML$_v$ and ML$_{v,\sigma,R}$ exhibit a dramatically tighter error distribution than a power law applied to the contaminated catalog. Even in comparison to the pure catalog, SDM produces a tighter error distribution.

Figure 7 shows a PDF of errors for all clusters above $3\times10^{14}\,M_\odot\,h^{-1}$ and for those above $7\times10^{14}\,M_\odot\,h^{-1}$. The PL$_C$ curve shows the PDF of errors associated with $M(\sigma_v)$ power law with the contaminated catalog's simple cylindrical cut about cluster centers. In contrast, the PL$_P$ curve shows the PDF of erros associated with the $M(\sigma_v)$ power law of the pure catalog, built from perfect knowledge of cluster members. For both ML$_v$ and ML$_{v,\sigma,R}$, the number of extreme overpredicted masses with $\epsilon \gtrsim 0.6$ is dramatically reduced over even the PL$_P$ power law. The extreme underpredicted masses with $\epsilon \lesssim -0.6$ are reduced compared to PL$_C$.

The mean error ($\bar{\epsilon}$) and median with central 68% width ($\epsilon \pm \Delta\epsilon$) of these PDFs are summarized in Table 3. Here we see PL's tendency to overpredict (positive ϵ and $\bar{\epsilon}$) in contrast with ML's tendency to underpredict (negative ϵ and $\bar{\epsilon}$). ML's underpredictions are caused by the hard upper mass limit and dearth of unique training halos at the high-mass end. The resulting underprediction is most conspicuous in ML$_v$ (both the contaminated test and contaminated high-mass test) and in ML$_{v,R}$ (contaminated high-mass test only). ML$_{v,R}$ has the smallest error offset (−0.04), but does so at the cost of underpredicting the highest-mass clusters. This bias is most evident at the higher-mass end, where halos' masses are systematically underpredicted. Because of this pronounced bias, ML$_{v,R}$ is therefore identified as a disfavored method.

The relative error widths ($\Delta\epsilon$) for all ML methods for all methods are more than a factor of two smaller than PL$_C$ (69%, 69%, 58%, and 64% for ML$_v$, ML$_R$, ML$_{v,R}$, and ML$_{v,\sigma,R}$, respectively). Even compared to PL$_P$, which is applied to the pure catalog, SDM produces a smaller relative error width (23%, 23%, 3%, and 12% for ML$_v$, ML$_R$, ML$_{v,R}$, and ML$_{v,\sigma,R}$, respectively).

As we saw in Figures 2 and 4, there is a wide scatter in σ_v associated with the contaminated test catalog. Shown in the right panel of Figure 7 are the clusters for which PL$_C$ severely overestimated cluster mass. These objects are particularly worrisome, as are predicted by PL$_C$ as being much more massive than they truly are, appearing to be rare, high-mass clusters. These outliers are isolated by their residual, δ (Equation (2)); each has $\delta \geqslant 1.5\times\sigma_{\rm gauss}$. We find that the ML error PDF for these objects is centered on zero, with a PDF width only slightly wider than the one shown in the left panel of Figure 7 for the full catalog. Furthermore, while the PL$_C$ method over predicts catastrophically, the ML methods predict much more reasonable masses.

Figure 8 shows a comparison of the five methods applied to the contaminated catalog: PL$_C$, ML$_v$, ML$_R$, ML$_{v,R}$, and ML$_{v,\sigma,R}$. The difference in absolute errors, denoted as

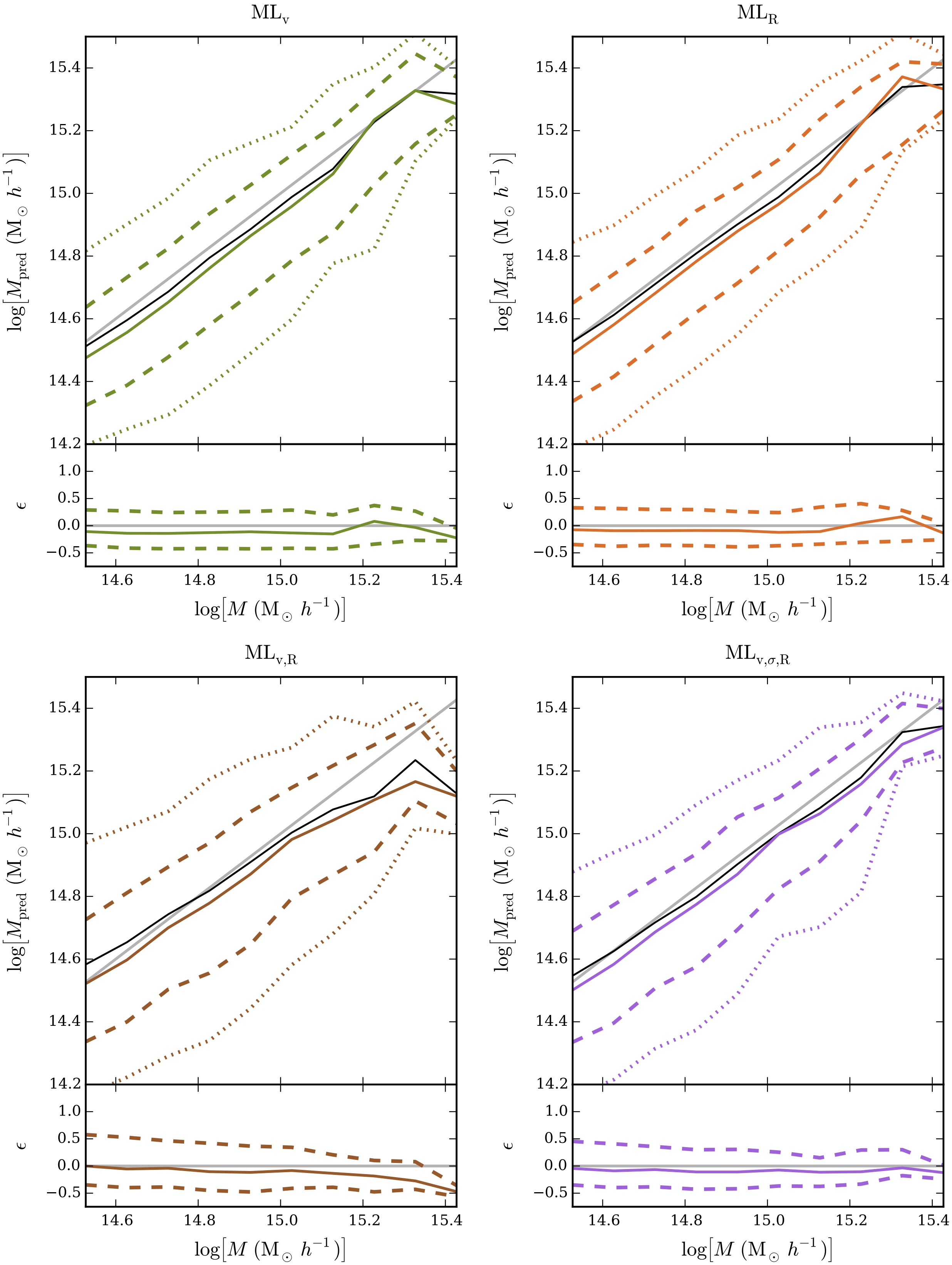

Figure 5. Top left: SDM results for ML_v (green). The predicted vs. true mass is binned in 0.1 dex $\log[M\,(M_\odot\,h^{-1})]$ bins. Mean (black solid), median (colored solid), 68% (dashed), and 95% (dotted) scatter are shown (top). The median error (solid) and error 68% scatter (dashed) are also shown (bottom). ML_v gives better than a factor-of-two reduction in the width of error compared to a standard scaling relation applied to the same catalog. Top right: SDM results for ML_R (orange). ML_R and ML_v minimize the width of the error distribution. Bottom left: SDM results for $\mathrm{ML}_{v,R}$ (brown). $\mathrm{ML}_{v,R}$ underpredicts at high masses and is therefore identified as a disfavored method. Bottom right: SDM results for $\mathrm{ML}_{v,\sigma,R}$ (purple). $\mathrm{ML}_{v,\sigma,R}$ minimizes the tendency to underpredict across mass range.

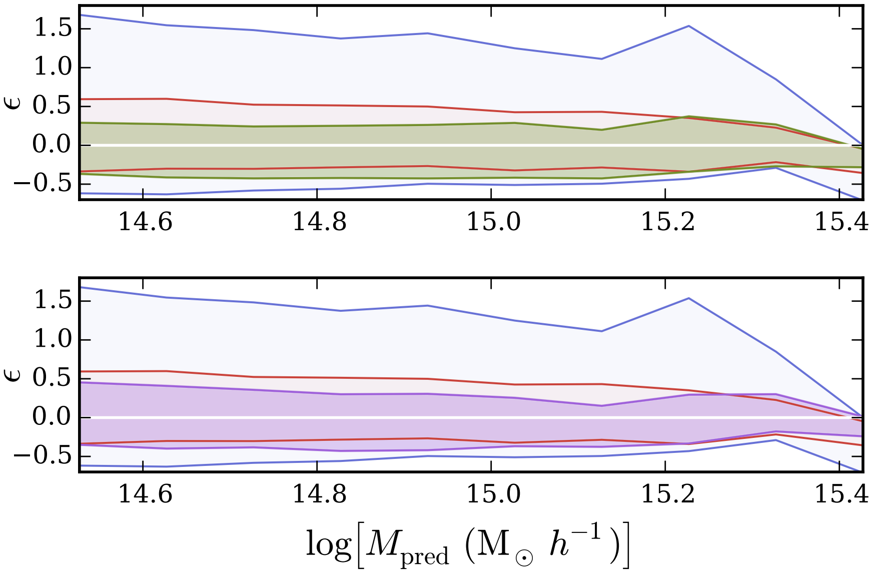

Figure 6. Top: error 16th and 84th percentiles (i.e., 68% scatter) as a function of mass for ML_v (green) as compared to a power-law approach applied to the pure catalog ($\mathrm{PL_P}$, red) and to the contaminated catalog ($\mathrm{PL_C}$, blue). Bottom: error scatter as a function of mass for $\mathrm{ML}_{v,\sigma,R}$ (purple) compared to $\mathrm{PL_P}$ and $\mathrm{PL_C}$. The errors of a dynamical mass power-law approach with a more refined interloper removal scheme should be bounded by $\mathrm{PL_C}$ and $\mathrm{PL_P}$. However, even when trained on the impure and incomplete catalog that produced the blue $\mathrm{PL_C}$ results, ML_v and $\mathrm{ML}_{v,\sigma,R}$ have ϵ widths comparable to or smaller than the best-case $\mathrm{PL_P}$ power law.

$|\epsilon_{\rm row}| - |\epsilon_{\rm column}|$, gives a measure of how well the row method predicts compared to the column method; values below zero are indicative of the row method predicting more accurately. The left column of this plot shows a comparison of ML to PL; all four ML methods consistently predict masses with a much smaller error than $\mathrm{PL_C}$. The mean difference in the absolute value of errors, denoted as $|\epsilon| - |\epsilon_{\mathrm{PL_I}}|$, is summarized in Table 3. This summary statistic quantifies the mean value shown in the left column of Figure 8. The more negative this value is, the more reduced a model's errors are compared to $\mathrm{PL_C}$. Model ML_R decreases error ϵ by an average of 0.61 compared to $\mathrm{PL_C}$; ML_R is the best ML method by this measure. The right three columns of Figure 8 compare the ML techniques to one another. $\mathrm{ML}_{v,R}$ is shown here to be the weakest of the ML methods; though it outperforms $\mathrm{PL_C}$ by a large margin, SDM produces more accurate mass predictions when applied with other feature sets.

As in Ntampaka et al. (2015), pairing $|v_{\rm los}|$ with the feature $|v_{\rm los}|/\sigma_v$ accentuates differences in velocity PDF shape and highlights, for example, the wide, flat hallmark PDF of a halo experiencing infalling matter. As a result of this additional feature, the mean and median errors edge closer to the desired values of zero. This offers an explanation as to why the three-feature set of $\mathrm{ML}_{v,\sigma,R}$ shows a mean error closer to zero (0.01) compared to ML_v and ML_R. $\mathrm{ML}_{v,\sigma,R}$ is identified as the preferred feature set for minimizing error bias.

Though $\mathrm{ML}_{v,R}$ employs two features that are highly correlated with mass, these features reside in a two-dimensional feature space. The joint distribution of $|v_{\rm los}|$ and R is likely too sparsely sampled by the galaxies in an individual cluster to make a strong correlation between this joint distribution and cluster mass. This effect becomes particularly pronounced for rare, massive clusters, which are underpredicted by $\mathrm{ML}_{v,R}$.

$\mathrm{ML}_{v,\sigma,R}$, however, predicts the masses of these clusters well. This may be explained by the nature of the third feature, $|v_{\rm los}|/\sigma_v$. Though the probability distribution employed by $\mathrm{ML}_{v,\sigma,R}$ resides in a three-dimensional feature space, the combination of $|v_{\rm los}|$ with $|v_{\rm los}|/\sigma_v$ constrains individual clusters' distributions to lie on a plane. These planes are sorted in the three-dimensional space by their slope, σ_v. This sorting effectively isolates high-σ_v clusters from low-σ_v ones. As we have seen with $\mathrm{PL_C}$, σ_v is a predictor of mass, albeit with significant scatter.

By taking advantage of the full LOS velocity and projected radius distributions, the SDM approach to determining cluster mass from galaxy observables reduces the distribution of errors by roughly a factor of two, and also predicts masses well even in the cases where $\mathrm{PL_C}$ catastrophically over predicts, making it a valuable tool for probing cosmological models with observations of galaxy clusters.

4. DISCUSSION

Reducing errors and eliminating biases in cluster mass measurements are crucial to utilizing clusters to discern and constrain cosmological models. The halo mass function and its evolution are sensitive to cosmological parameters such as σ_8, Ω_M, $\Omega_{\rm DE}$, and w (e.g., Schuecker et al. 2003; Henry et al. 2009; Vikhlinin et al. 2009; Mantz et al. 2010; Rozo et al. 2010; Allen et al. 2011). Therefore, accurate measurements of cluster abundance as a function of mass and redshift can be used to understand the underlying cosmology. The limiting factor in constraining parameters and evaluating cosmological models with cluster counts, however, is in accurately connecting galaxy observables to halo mass to reproduce the halo mass function.

Figure 9 shows how the scatter and biases in each model affect the halo mass functions recovered by $\mathrm{PL_P}$, $\mathrm{PL_C}$, ML_v, and $\mathrm{ML}_{v,\mathrm{P}}$ (SDM applied to the pure catalog with feature $|v_{\rm los}|$, as in Ntampaka et al. 2015) in comparison to the simulation's true mass function. The scatter about the scaling relation in $\mathrm{PL_P}$ coupled with the rapidly declining shape of the mass function causes the abundant, low-mass clusters with high δ to populate the high-mass bins in the mass function, causing the upscattering at high masses. This effect is exacerbated in $\mathrm{PL_C}$, where the scatter about the scaling relation is much larger and

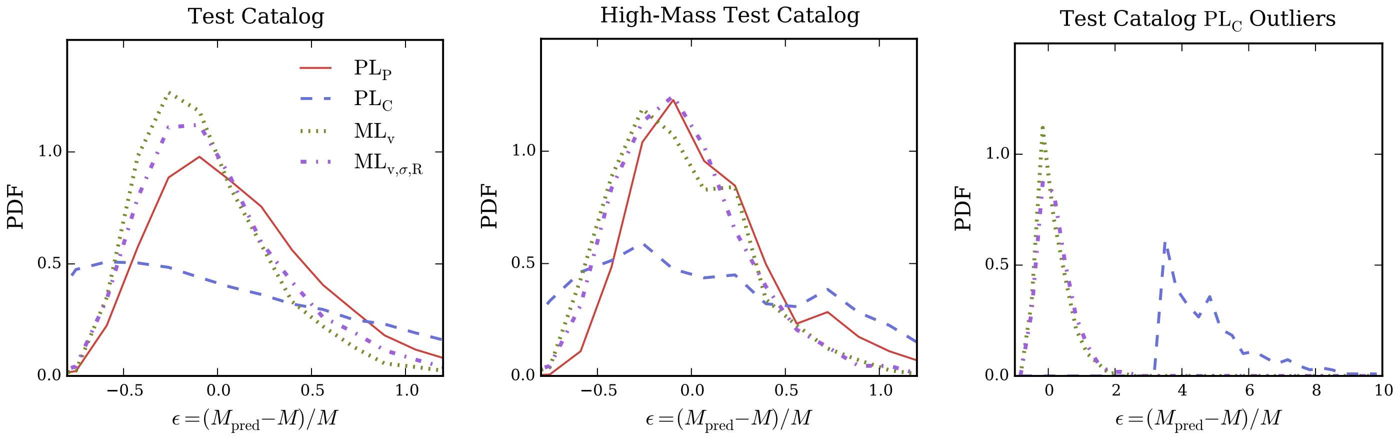

Figure 7. Left: PDF of fractional mass errors for the full test catalogs. A power-law $M(\sigma_v)$ scaling relation for a catalog of pure and complete clusters shows significant errors (PL_P, red solid). The error distribution widens further when interlopers contaminate the clusters (PL_C, blue dashed). Remarkably, SDM (ML_v, green dotted, and $ML_{v,\sigma,R}$, purple dash dotted) applied to the contaminated catalog outperform the $M(\sigma_v)$ scaling relation applied to the pure catalog. Center: PDF of errors for the high-mass test catalogs ($M \geqslant 7 \times 10^{14}\,M_\odot\,h^{-1}$) shows a similar trend for rare, high-mass halos; the ML approaches minimize error significantly over a power-law scaling relation applied to the same catalog. Right: PDF of the high-δ, high-PL_C-error population of clusters. While the power law catastrophically overestimates the masses of these outlying objects, ML approaches perform well, with a PDF of fractional mass errors for these outliers that is only slightly wider than is found for the full catalog.

Table 3
Method Comparison

Case	Summary	Color	Catalog	$\bar{\epsilon}$[a]	$\epsilon \pm \Delta\epsilon$[b]	$\Delta\epsilon$[c]	$\lvert\epsilon\rvert - \lvert\epsilon_{PL_C}\rvert$[d]
PLM	$M(\sigma_v)$ Power Law, Pure	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	⋯
PL_C	$M(\sigma_v)$ Power Law, Contaminated	Blue	Test	0.508	$0.13^{+1.40}_{-0.73}$	2.131	⋯
			High-mass Test	0.409	$0.18^{+1.15}_{-0.68}$	1.829	⋯
ML_v	ML with v_{los}	Green	Test	−0.052	$-0.12^{+0.40}_{-0.27}$	0.670	−0.63
			High-mass Test	−0.059	$-0.10^{+0.38}_{-0.31}$	0.686	−0.47
ML_R	ML with R	Orange	Test	−0.016	$-0.08^{+0.39}_{-0.28}$	0.670	−0.64
			High-mass Test	−0.040	$-0.10^{+0.37}_{-0.26}$	0.635	−0.49
$ML_{v,R}$	ML with $\lvert v_{los}\rvert$ and R	Brown	Test	0.078	$-0.04^{+0.56}_{-0.34}$	0.899	−0.54
			High-mass Test	−0.032	$-0.11^{+0.45}_{-0.33}$	0.783	−0.42
$ML_{v,\sigma,R}$	ML with $\lvert v_{los}\rvert$, $\lvert v_{los}\rvert/\sigma_v$, & R	Purple	Test	0.011	$-0.07^{+0.46}_{-0.31}$	0.763	−0.61
			High-mass Test	−0.044	$-0.09^{+0.36}_{-0.29}$	0.649	−0.49

Notes.
[a] Mean fractional mass error.
[b] Median fractional mass error ±68% scatter.
[c] Width of ϵ 68% scatter.
[d] Mean difference between model and PL_C errors.

the high-δ clusters may be catastrophically overpredicted (as shown in Figure 7). This effect, known as Eddington bias (Eddington 1913), alters the shape and amplitude of the measured halo mass function from the true value. This results in PL_C's measured mass function dramatically overreporting the number of high-mass clusters.

Any cosmological analysis of the HMF that employs such mass measurements must correct for this upscatter at high masses. Understanding the nature of the intrinsic scatter and observational selection effects is a crucial step to correct the observed HMF for Eddington bias. Analytic approaches exist to correct for the simple case of lognormal scatter (e.g., Mortonson et al. 2011; Evrard et al. 2014), while a more complicated scatter may be more difficult to correct. Before correction for Eddington bias, the large scatter and errors associated with traditional power-law mass measurements lead to the failure to recover the true mass function, which limits the constraining power of dynamical mass measurements of galaxy clusters. PL_P's altered shape mimics the mass function of a simulation with a higher σ_8 and Ω_M. This is particularly pronounced in the fractional difference, $\Delta y/y$, between the Multidark and mock HMFs, which shows that the presence of interlopers causes the PL HMF to deviate from the simulation HMF, particularly at high masses.

At the low-mass end, the underabundance of clusters is not caused by Eddington bias, but is an artifact of the hard lower-mass limits of the test catalogs. This downscattering should not be interpreted as a dearth of low-mass clusters predicted by the PL and ML methods, but rather as a limitation of the test catalogs.

In addition to the halo mass functions from the methods highlighted in this work, mock HMFs that include scatter of other common cluster mass measurement techniques are included for comparison. Cluster masses can be deduced from a variety of techniques, and here we show three different methods for determining cluster mass: the Sunyaev–Zel'dovich

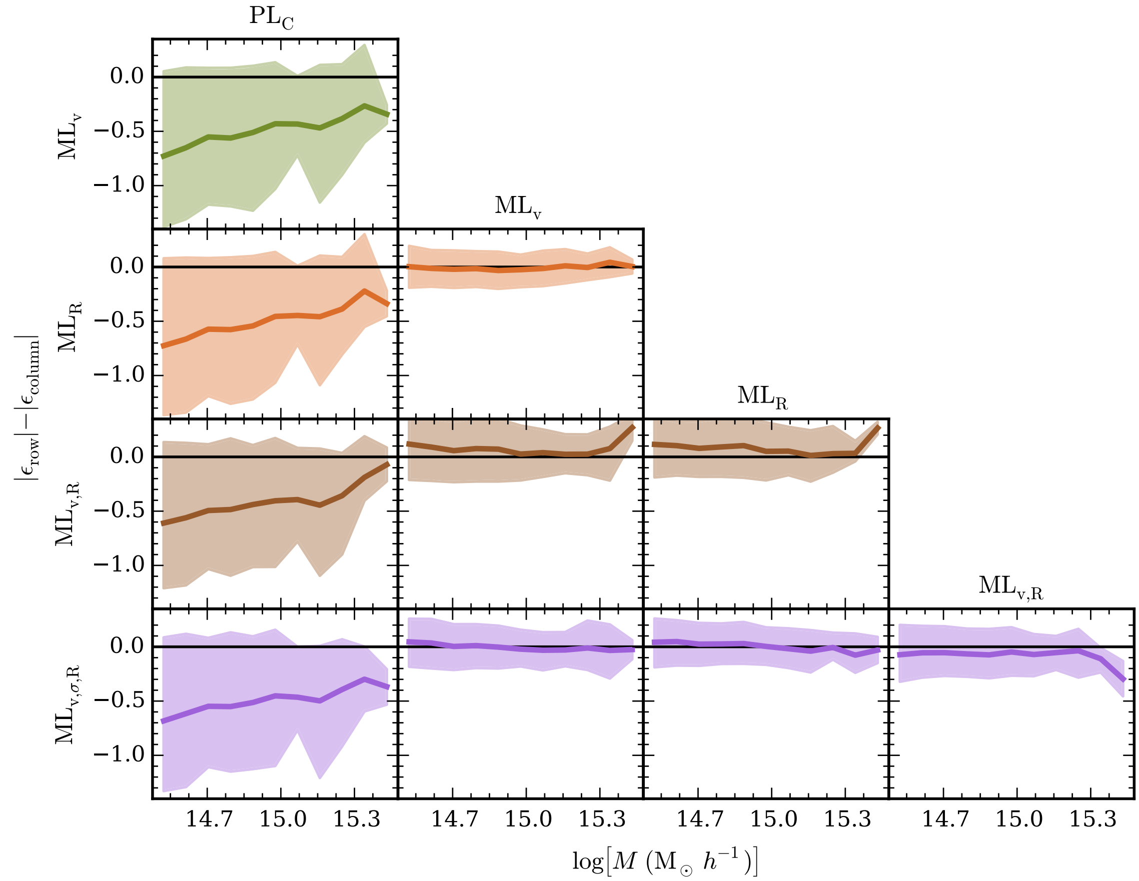

Figure 8. Summary comparison of the five methods trained and tested on the contaminated catalog, with difference in absolute error, $|\epsilon_\mathrm{row}| - |\epsilon_\mathrm{column}|$, as a function of mass (see Equation (3)). Values below the solid black 0 line indicate that the row method is performing better than the column method for a given mass bin. The left column summarizes a comparison of the four new SDM methods to the PL$_\mathrm{C}$ power law; SDM with any of the four feature combinations improves mass predictions in all mass bins. While ML$_{v,R}$ outperforms PL$_\mathrm{C}$, it performs poorly at high masses compared to the other ML methods.

(SZ) effect, weak gravitational lensing (WL), and X-ray. The SZ effect, first proposed by Sunyaev & Zeldovich (1972) can be used to determine a temperature-weighted gas mass, and we model its intrinsic scatter according the Battaglia et al. (2012) scaling relation for $z = 0$ with AGN feedback. Weak gravitational lensing probes structure along the LOS, and we model scatter in this technique according to the Becker & Kravtsov (2011) prescription for $z = 0.25$, $M_{500c} \geqslant 2.0\times 10^{14}\ M_\odot\ h^{-1}$ clusters. X-ray observations can be used to infer a gas mass profile; scatter in this M–Y_X relation of $\sigma_{\ln M} = 0.06$ is adopted from Fabjan et al. (2011), and it should be noted that this is intrinsic scatter and does not include observational effects. The mass–concentration relation from Bhattacharya et al. (2013) and the NFW density profile from Navarro et al. (1996) are implemented to convert all masses to M_{200c} for comparison.

Figure 9 shows the halo mass functions recovered by SZ, WL, and X-ray methods compared to the range of scatters achievable with SDM: ML$_{v,\mathrm{P}}$ with a pure and complete cluster membership catalog and ML$_v$ with a large cylindrical cut around each cluster allowing many interlopers. It should be noted that the HMF presented assumes a complete large mock observation of 6834 (7449) clusters in the pure (contaminated) catalog. Figure 9 also shows the Poisson error associated with a more reasonable observation of 500 clusters. Current cluster surveys (e.g., de Haan et al. 2016) contain on the order of hundreds of clusters, and the choice of 500 clusters is chosen to show the errors accessible through current catalogs. Note that the small number of high-mass objects limit the accuracy with which the tail end of the HMF can be determined. As is shown in, e.g., Ntampaka et al. (2016), a binned HMF has the most power to resolve σ_8–Ω_m models at the lowest masses because, while high-mass clusters are sensitive to changes in these cosmological parameters, the Poisson error bars on these rare objects dominates. For the mass ranges where the HMF can best resolve changes in σ_8 and Ω_m, SDM produces a competitive HMF to these other mass proxies, though it has a larger deviation from the true HMF at the high-mass tail.

However, it should be noted that these cluster mass methods utilize different wavelength observations with different systematic errors, biases, and limitations. Therefore, while Figure 9 shows that five different cluster mass techniques—PL, ML, SZ, X-ray, and WL—in a direct comparison, it should not be overly interpreted as a definitive guide to cluster mass measurement. For example, weak lensing is difficult and expensive to apply to high-redshift clusters due to a lack of adequate background galaxies. Biases in X-ray and SZ cluster masses may arise because of nonthermal pressure support (e.g., Evrard 1990; Rasia et al. 2004; Lau et al. 2009; this bias is not modeled in Figure 9 because this effect is typically corrected for, though uncertainty in the bias may produce further disagreement between observed and true HMF). When SZ masses are calibrated on simulation, the calibration is dependent on correct modeling of the gas physics (e.g., Nagai 2006; Battaglia et al. 2012), which may also introduce a bias.

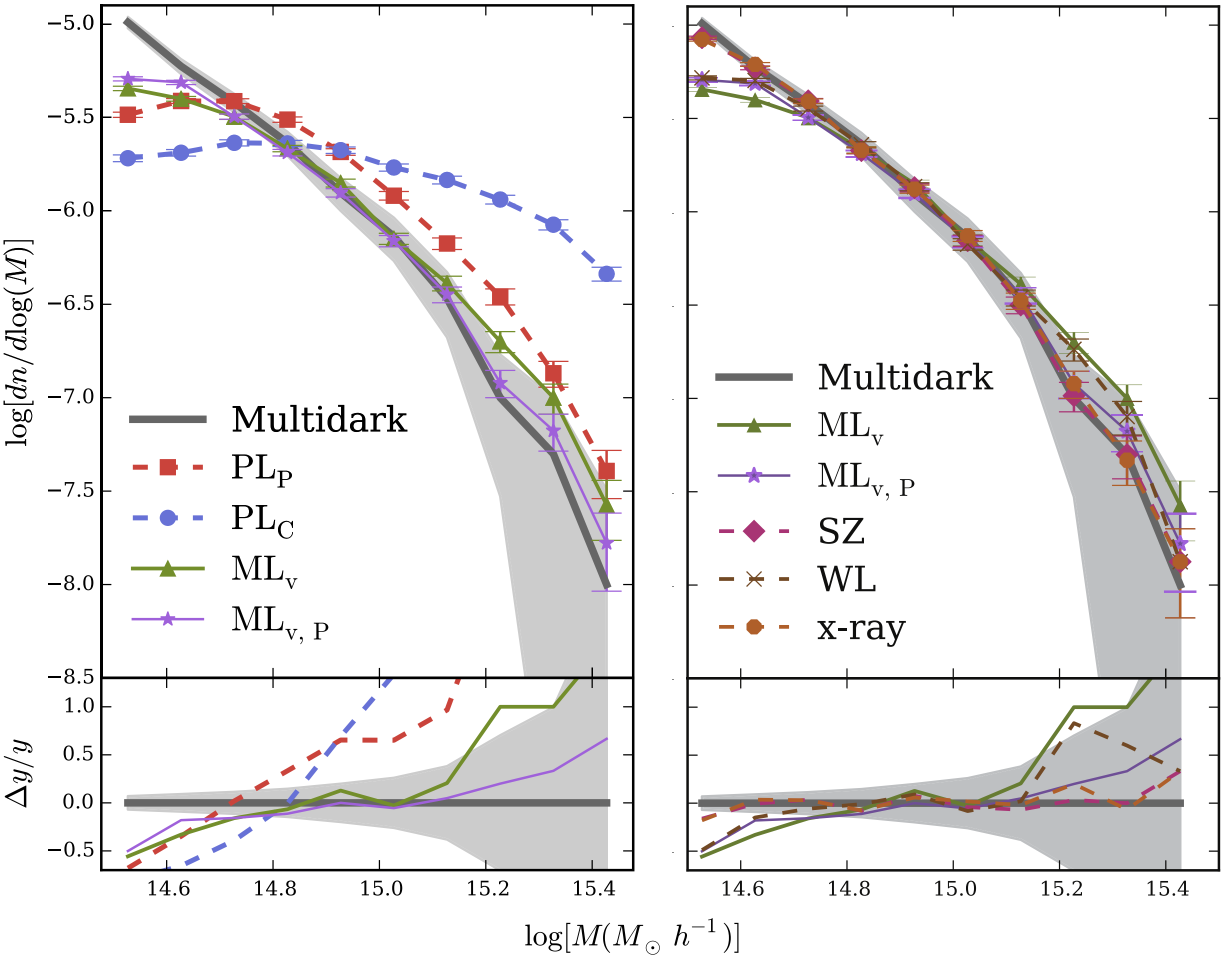

Figure 9. Halo mass functions of dynamical cluster mass estimates with intrinsic scatter only (pure catalog) and intrinsic scatter plus observational selection effects (contaminated catalog). Any scatter in the mass-observable relationship, if uncorrected, will affect the observed halo mass function. The large scatter associated with the power-law scaling relation (PL$_P$, red squares, and PL$_C$, blue circles) causes an upscatter at high masses, while ML methods (ML$_{v,P}$, purple stars, and ML$_v$, green triangles) have a smaller intrinsic scatter and more accurately reproduce the true Multidark cluster abundance (dark gray solid curve). While 6834 (7449) clusters contribute to the HMF for the pure (contaminated) catalog, a more moderate observation of 500 clusters yields larger Poisson error bars (light gray band). Right: HMF of ML methods compared to mock HMF with the typical intrinsic scatter of Sunyaev–Zel'dovich (pink diamond), weak lensing (brown x), and X-ray (orange octagon) cluster masses. The biases and the observational effects associated with SZ, WL, and X-ray masses may introduce additional scatter, causing the HMF to deviate further from the simulation HMF.

Dynamical and ML masses, however, can be directly compared as they are produced from the same data from the same mock catalog and are affected by the same observational selection effects. From their direct comparison, it can be concluded that the ML method presented in this work is more competitive than a power-law scaling relation for decreasing errors in cluster mass measurements. While ML$_v$ over predicts the abundance of high-mass clusters, the upscatter is smaller than PL$_P$'s. ML$_v$ provides a much better match to the simulation's true mass function across a larger mass range, comparable to those of SZ, WL, and X-ray for the large mock observation of ≈(1 Gpc h^{-1})^3. This agreement with the true HMF is primarily due to the small spread in errors associated with these methods; abundant, low-mass clusters tend not to be catastrophically overpredicted by methods with small intrinsic scatter. The smaller errors produced in SDM's mass prediction results in a more accurate representation of the halo mass function, particularly at the high-mass end. SDM's ability to more accurately recreate the true halo mass function makes it a valuable tool for producing cluster mass functions to evaluate cosmological models. The predictive power of SDM to reproduce the true halo mass function and its implications for constraining cosmological parameters σ_8 and Ω_M will be explored in detail in an upcoming work.

The Appendix explores how the aperture and, less directly, the purity and completeness of the cluster sample, affect the scatter in both power law and ML dynamical masses. We find that the power-law fit changes as a function of aperture, shallowing with smaller aperture. When a large aperture is used, the distribution of errors at low masses is not lognormal, but is better described by a double Gaussian (see Figure 11).

With the simple cylindrical cut and 2σ paring used in this work, mock cluster observations performed with a large aperture will tend to be more complete (compared to a mock observation made with a smaller aperture), with cluster members near the edges of the cluster being included in the sample. Mock observations with a smaller aperture will tend to be more pure, with fewer interlopers contaminating the observation. As we will show in the Appendix, SDM performs slightly better with a large aperture, showing a preference for completeness over one for purity.

One may consider improving SDM mass predictions further by training and testing features beyond simply R and $v_{\rm los}$, applying a more accurate cluster interloper removal technique, or limiting the training sample to a particular subpopulation of galaxies. Because elliptical galaxies preferentially reside in galaxy clusters (Dressler 1980), limiting the training sample to this population may provide a straightforward and natural approach to excluding many interlopers while still providing limited information about infalling matter. However, before such a training set can be explored and applied to observational data, there remains a need for a reliable training N-body simulation that is large, high resolution, and realistically populated with galaxies.

5. CONCLUSIONS

We compare cluster mass predictions from a standard $M(\sigma_v)$ power-law scaling relation to those generated by SDMs, an ML class of algorithms that learns from a distribution of data to predict a scalar.

We focus on mass predictions for a mock catalog of impure and incomplete clusters. This catalog is created from the publicly available Multidark MDPL1 simulation, with an intentionally simplistic cylindrical cut imposed around the known centers of clusters with true mass $\geqslant 1 \times 10^{14}\,M_\odot\,h^{-1}$. The aperture ($R_{\rm aperture} = 1.6$ Mpc h^{-1}) and initial velocity cut ($v_{\rm cut} = 2500$ km s^{-1}) correspond to a typical radius and $2 \times \sigma_v$ of a halo with a mass of $1 \times 10^{15}\,M_\odot\,h^{-1}$. Velocity outliers beyond $2\sigma_v$ are iteratively pared until convergence, and only clusters with at least 20 cluster members are kept in the sample. This creates a catalog of clusters that are both impure (interlopers contaminate the clusters) as well as incomplete (some true cluster members are excluded from the sample). A second catalog, both pure and complete, is also prepared for comparison.

Cluster masses are predicted in two ways: in the PL approach, a standard $M(\sigma_v)$ power law is used to train and test, while in the ML approach, SDM is utilized. Four feature sets are considered with SDM: ML$_v$ (absolute value of the LOS velocity, $|v_{\rm los}|$), ML$_R$ (galaxy projected distance from the cluster center, R), ML$_{v,R}$ ($|v_{\rm los}|$ and R), and ML$_{v,\sigma,R}$ ($|v_{\rm los}|$, $|v_{\rm los}|/\sigma_v$, and R). Results for halos with true mass $M \geqslant 3 \times 10^{14}\,M_\odot\,h^{-1}$ are reported.

Our main conclusions can be summarized as follows.

1. ML$_v$ and ML$_R$ (SDM with $|v_{\rm los}|$ feature only and SDM with R feature only, respectively) reduce errors by 69% compared to a power law applied to the same contaminated catalog.
2. Furthermore, though a simple cylindrical cut causes significant scatter in the $M(\sigma_v)$ power law compared to when the cluster membership is perfectly known, both SDM methods each outperform PL$_{\rm P}$, a power law applied to a catalog with pure and complete clusters. Compared to this ideal power law, ML$_v$ and ML$_R$ each reduce error by 23%.
3. Though it reduces error width, ML$_{v,R}$ (SDM with $|v_{\rm los}|$ and R) systematically underpredicts the highest-mass clusters. It is identified as a disfavored method.
4. ML$_{v,\sigma,R}$ (SDM with $|v_{\rm los}|$, $|v_{\rm los}|/\sigma_v$, and R) minimizes the bias for the high-mass clusters ($M \geqslant 7 \times 10^{14}\,M_\odot\,h^{-1}$). It reduces error by 64% and 12% compared to PL$_{\rm C}$ and PL$_{\rm P}$, respectively.
5. In some instances, a higher-than-expected σ_v causes a catastrophic overprediction by method PL$_{\rm C}$. The ML methods, however, predict reasonable masses for even these outliers.

The SDM approach to determining cluster mass from galaxy observables reduces errors by more than a factor of two compared to a standard power-law scaling approach applied to a cluster catalog with impure, incomplete cluster membership information. SDM predicts cluster masses well even when a traditional $M(\sigma_v)$ approach fails. Additionally, this technique works well even with catalogs of impure and incomplete clusters created with a simplistic cylindrical cut about the cluster center. Ultimately, high-resolution, large-volume simulations are needed for training before SDM can be applied to observation. With such a simulation for training, the reduced errors and more accurate predictions for impure, incomplete, nonvirialized systems makes SDM a valuable tool for constraining cosmological models.

We thank our referee, and also Nicholas Battaglia, Arthur Kosowsky, Rachel Mandelbaum, and Cristóbal Sifón for their valuable feedback on this manuscript. 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.

APPENDIX

Here, we explore how our choices of $R_{\rm aperture}$ and $v_{\rm cut}$ affect the PL and ML predictions and results. Two new catalogs are prepared to correspond to a $3 \times 10^{14}\,M_\odot\,h^{-1}$ cluster ($R_{\rm aperture} = 1.1$ Mpc h^{-1} and $v_{\rm cut} = 1570$ km s^{-1}, denoted "Small Aperture") and $3 \times 10^{15}\,M_\odot\,h^{-1}$ cluster ($R_{\rm aperture} = 2.3$ Mpc h^{-1} and $v_{\rm cut} = 3785$ km s^{-1}, denoted "Large Aperture"). The contaminated catalog used in the main body of this work has been renamed "medium aperture" for clarity. As before, a 2σ iterative paring scheme is applied to the initial cylindrical cut. With the exception of the $R_{\rm aperture}$ and $v_{\rm cut}$ values, the methods described in Section 2 are followed. These catalogs, along with the pure catalog, are summarized in Table 4.

Figure 10 shows how the choice of $R_{\rm aperture}$ and $v_{\rm cut}$ affect the power-law fits. This two-dimensional histogram of σ_v versus M shows that the best-fit α and σ_v, as well as the scatter about the best-fit line, changes as a function of initial cylinder size. Overlaid on the two-dimensional histogram is a best fit with 1σ and 2σ lognormal errors, calculated for clusters with masses above $3 \times 10^{14}\,M_\odot\,h^{-1}$ and extrapolated down to lower masses. Additionally overlaid is the best-fit power law for PL$_{\rm P}$.

When the small aperture cuts are applied, this overly small cylinder clips the σ_v values at the high mass. This leads to a shallow slope ($\alpha = 0.209$) and small velocity dispersion

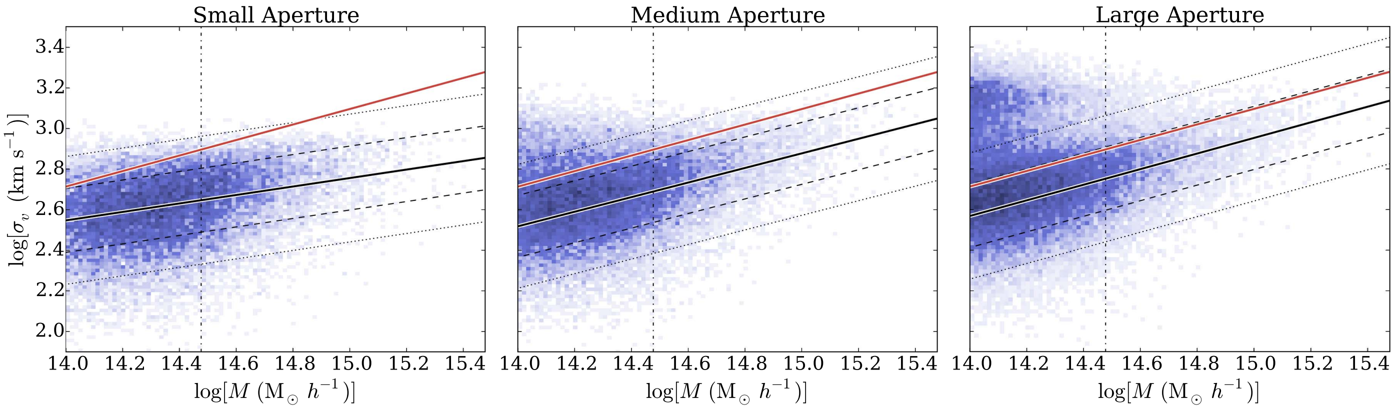

Figure 10. Left: small aperture catalog's LOS velocity dispersion of galaxies, σ_v, vs. cluster mass, M, shown as a 2D histogram. Only clusters above $3 \times 10^{14}\,M_\odot\,h^{-1}$ (black dash dotted) are used to determine the best-fit power law (black solid); the small aperture and $v_{\rm cut}$ lead to smaller-than-expected σ_v's for the high-mass halos and result in a shallow fit. The $M\,(\sigma_v)$ fit for pure and complete clusters (PL$_{\rm P}$, red) is overlaid for reference. Center: medium aperture catalog. If the lognormal scatter in σ_v was consistent across the entire mass range, the 1σ and 2σ errors (black dashed and dotted, respectively) calculated at the high-mass end would describe the scatter in σ_v even at low masses. However, a clear trend emerges, with increased scatter in σ_v at lower masses. Right: large aperture catalog. The slope of the power law has steepened. This is due to the larger $R_{\rm aperture}$ and $v_{\rm cut}$ used for this catalog, which capture more true members of the high-mass clusters, allowing these objects to be more accurately described. Though the high-mass clusters are now well-represented by their measured σ_v, a clear second population emerges at low-mass and high σ_v, with 20% of halos with $M < 3 \times 10^{14}\,M_\odot\,h^{-1}$ lying above the 2σ dotted line.

associated with a $10^{15}\,M_\odot\,h^{-1}$ cluster ($\sigma_{15} = 569$ km s^{-1}). In contrast, a large cylindrical fit increases scatter at the low-mass end. The resulting fit for the large aperture catalog is steep ($\alpha = 0.384$) and has a higher normalization ($\sigma_{15} = 895$ km s^{-1}) caused by the many high-σ_v objects and the substantial fraction of outliers above the 2σ line. These catalogs and fits are summarized in Table 4 for reference.

As the large aperture catalog's cuts are used to probe lower masses, a bimodal distribution emerges with a second population of clusters residing far above the best fit; this second population is visible in Figure 2. These high-σ_v, low-mass objects increase scatter at the low-mass end. More worrisome, they have a velocity dispersion typically associated with clusters of roughly an order of magnitude larger in mass. Of halos with $M \geqslant 3 \times 10^{14}\,M_\odot\,h^{-1}$, 3% reside above the $+2\sigma$ dotted line and 3% reside below the -2σ dotted line. However, of halos with $1 \times 10^{14}\,M_\odot\,h^{-1} \leqslant M < 3\times 10^{14}\,M_\odot\,h^{-1}$, 20% reside above $+2\sigma$ and 3% below -2σ. The best fit and lognormal scatter found for the higher-mass clusters in the large aperture catalog is clearly not descriptive of the lower-mass clusters.

To further explore this outlier population, we will consider the residual, δ (Equation (2)). Figure 11 shows that the large aperture catalog has a residual PDF is adequately described by a single Gaussian, parameterized by

$$\mathrm{PDF} \propto \exp\left[\frac{-(\delta - \mu)^2}{2\,\sigma_{\rm gauss}^2}\right], \tag{4}$$

with best-fit width $\sigma_{\rm gauss} = 0.13$ and a nearly zero offset, $\mu = 0.01$.

However, when the lower-mass limit of this large aperture catalog is decreased to $1 \times 10^{14}\,M_\odot\,h^{-1}$, the δ PDF is better described by the sum of two Gaussians, as shown Figure 11. The relative amplitude and width of high-δ Gaussian is dependent on the minimum mass cut applied to the catalog, and our choice of $1 \times 10^{14}\,M_\odot\,h^{-1}$ is for illustrative purposes only. Note, however, that the zero-centered Gaussian has $\sigma_{\rm gauss} = 0.16$ and $\mu = 0.03$, comparable to the single Gaussian fit found previously. This is suggestive that a single lognormal scatter describes the population that is well-characterized by the $M\,(\sigma_v)$ power law, while a second population of high-σ_v outliers emerges at low masses. Exploring observational methods for describing and identifying members of this outlier population will be considered in future work.

Table 4
Catalog Summary

Catalog Name	Type	$R_{\rm aperture}$ (Mpc h^{-1})	$v_{\rm cut}$ (km s^{-1})	σ_{15} (km s^{-1})	α
Small Aperture	PL Train	1.1	1570	569	0.209
Medium Aperture	PL Train	1.6	2500	895	0.384
Large Aperture	PL Train	2.3	3785	900	0.400
Pure	Train	⋯	⋯	1244	0.382

Figure 12 shows that the resulting large scatter produces PL$_{\rm C}$ error PDF that is wide and flat as before, with the shape of the PL$_{\rm C}$ PDF dependent on the cylindrical cut parameters. For the small aperture catalog, the shallow fit coupled with the large number of clusters with large negative δ contribute to the substantial population of clusters being underestimated by an order of magnitude or more ($\epsilon \lesssim -0.1$). SDM produces a slightly wider error distribution for this small initial cylinder compared to the Medium Aperture cuts, though still reducing $\Delta\epsilon$ compared to both PL$_{\rm C}$ and PL$_{\rm P}$. Distributions of error as a function of mass are comparable to those seen in Figure 5, regardless of the training catalog, though $\bar{\epsilon}$ tends to decrease and $\Delta\epsilon$ tends to widen for small initial cylinders.

As before, there are also a number of catastrophically overpredicted clusters by applying the PL$_{\rm C}$ scaling relation to the small aperture catalog. These overpredicted objects are identified by their residual relative to the lognormal scatter: $\delta \geqslant 1.5 \times \sigma_{\rm gauss}$. The shallow slope leads to the overprediction being much more pronounced. However, Figure 12 shows that, even in this case, SDM predicts reasonably accurate masses for these objects.

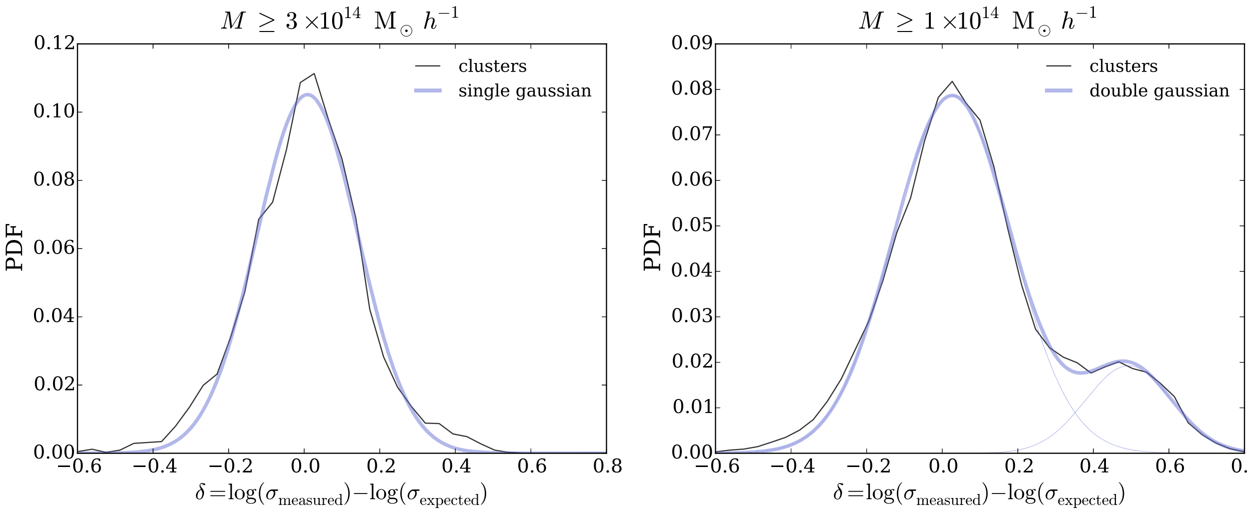

Figure 11. Left: PDF of residual, δ, for the large aperture catalog. With a lower-mass cut of $M = 3 \times 10^{14}\, M_\odot\, h^{-1}$, the PDF of clusters' δ (thin black) is well-described by a single Gaussian (thick blue). Right: when the mass limit of the large aperture catalog is lowered to $M = 1 \times 10^{14}\, M_\odot\, h^{-1}$, the PDF is better described by a double Gaussian. Observational methods for identifying members of this outlier population will be explored in a later work.

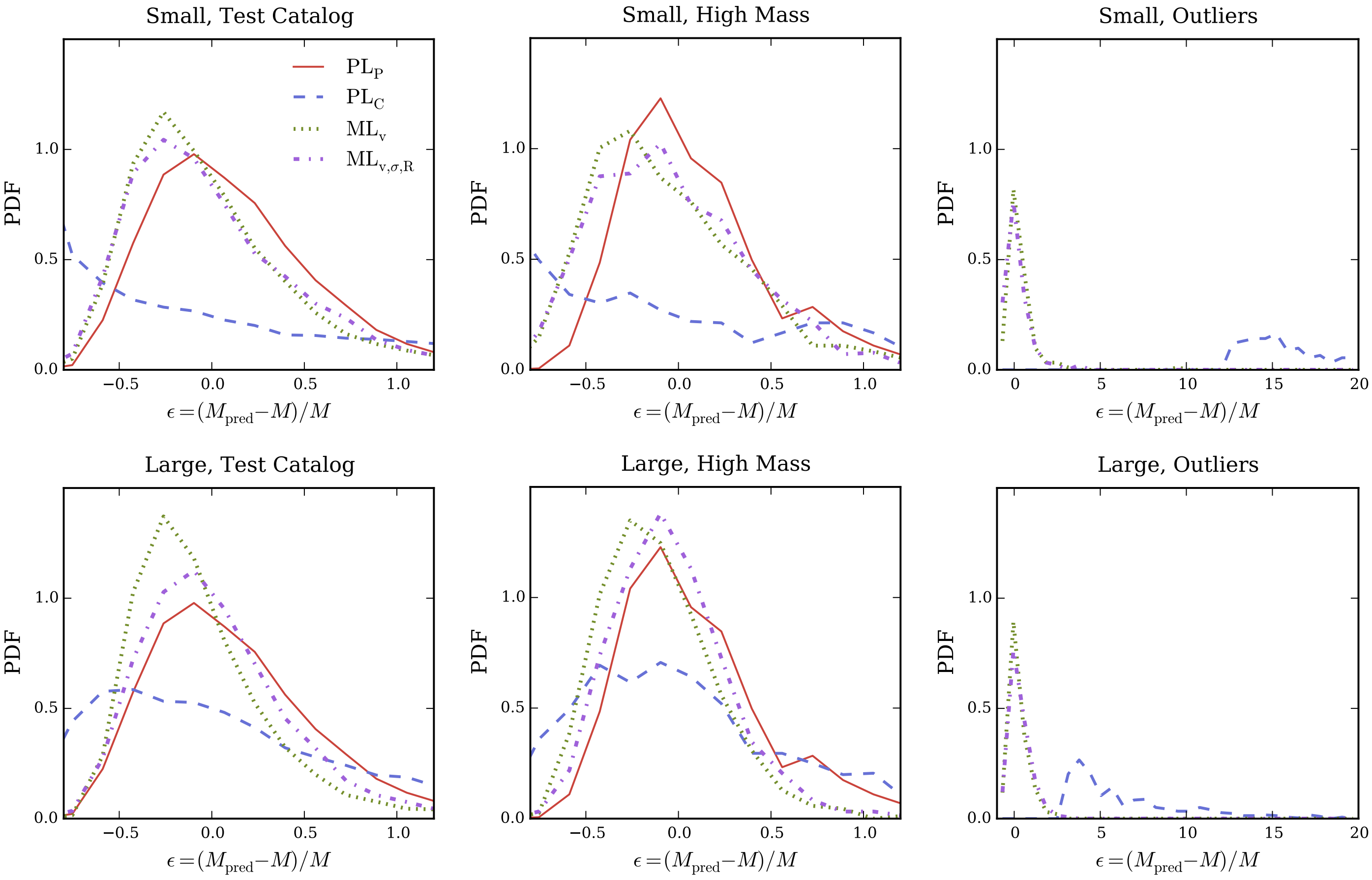

Figure 12. Top left: PDF of errors for the small aperture catalog. When this small cut is imposed on the mock observation, the shallow slope of the fit causes large-negative-δ population to be underpredicted in mass by an order of magnitude or more, creating the abundance of clusters with $\epsilon \lesssim 0.1$. Top center: small aperture catalog, high-mass halos only ($M \geqslant 7 \times 10^{14}\, M_\odot\, h^{-1}$), has a similar abundance of underpredicted halo masses. Top right: PDF of errors for the high-error objects. The shallow small aperture fit also results in a number of catastrophically overpredicted clusters. SDM, however, predicts reasonable masses for even these outliers. Bottom left: PDF of errors for the large aperture catalog. The large cut leads to more interlopers, but SDM predicts better than a scaling relation applied to a pure and complete catalog. Bottom center: large aperture catalog, high-mass halos only. Bottom right: PDF of high-error objects for the large aperture catalog. SDM predicts reasonably accurate masses here, though a power-law scaling relation fails catastrophically.

The population of high-σ_v, low-mass, high-δ objects in the large aperture catalog similarly produces a substantial number of catastrophically overpredicted clusters. These large-ϵ objects shown in Figure 12 are also well-predicted by SDM. While the $\rm PL_C$ gives a large range of errors, SDM can more accurately predict these cluster masses despite overly large or small cylindrical cuts that contribute to significant impurity or incompleteness in the mock clusters.

ML_v and $\mathrm{ML}_{v,\sigma,R}$ produce the smallest $\Delta\epsilon$ when the initial cylinders are large, with $\Delta\epsilon = 0.670$ and 0.763, respectively, for the medium aperture catalog and $\Delta\epsilon = 0.660$ and 0.752 for the large aperture catalog. The small aperture catalog error distribution is wider: $\Delta\epsilon = 0.809$ and 0.898. However, in all cases except $\mathrm{ML}_{v,\sigma,R}$ applied to the small aperture cylinder, the width of error distribution is narrower than the pure catalog power law, which has $\Delta\epsilon = 0.871$. SDM performs better with impurity over incompleteness, with larger cylinders producing slightly more accurate mass predictions.

Errors produced by a power-law scaling relation are clearly dependent on the choices of R_{aperture} and v_{cut}, sometimes catastrophically overpredicting cluster masses. Though standard power-law scaling fits and error distributions are sensitive to choices in cuts, SDM can predict accurately under a wide range of scenarios, provided the training and test data have the same imposed cuts.

REFERENCES

Allen, S. W., Evrard, A. E., & Mantz, A. B. 2011, ARA&A, 49, 409
Battaglia, N., Bond, J. R., Pfrommer, C., & Sievers, J. L. 2012, ApJ, 758, 74
Becker, M. R., & Kravtsov, A. V. 2011, ApJ, 740, 25
Bhattacharya, S., Habib, S., Heitmann, K., & Vikhlinin, A. 2013, ApJ, 766, 32
Biviano, A., & Girardi, M. 2003, ApJ, 585, 205
Bocquet, S., Saro, A., Mohr, J. J., et al. 2015, ApJ, 799, 214
Brodwin, M., Ruel, J., Ade, P. A. R., et al. 2010, ApJ, 721, 90
Cohn, J. D. 2012, MNRAS, 419, 1017
de Haan, T., Benson, B. A., Bleem, L. E., et al. 2016, arXiv:1603.06522
Dressler, A. 1980, ApJ, 236, 351
Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. 1997, in Advances in Neural Information Processing Systems 9 (Cambridge, MA: MIT Press), 155
Eddington, A. S. 1913, MNRAS, 73, 359
Evrard, A. E. 1990, ApJ, 363, 349
Evrard, A. E., Arnault, P., Huterer, D., & Farahi, A. 2014, MNRAS, 441, 3562
Evrard, A. E., Bialek, J., Busha, M., et al. 2008, ApJ, 672, 122
Fabjan, D., Borgani, S., Rasia, E., et al. 2011, MNRAS, 416, 801
Fadda, D., Girardi, M., Giuricin, G., Mardirossian, F., & Mezzetti, M. 1996, ApJ, 473, 670
Falco, M., Hansen, S. H., Wojtak, R., et al. 2014, MNRAS, 442, 1887
Geller, M. J., Diaferio, A., Rines, K. J., & Serra, A. L. 2013, ApJ, 764, 58
Gifford, D., & Miller, C. J. 2013, ApJL, 768, L32
Hansen, S. M., McKay, T. A., Wechsler, R. H., et al. 2005, ApJ, 633, 122
Henry, J. P., Evrard, A. E., Hoekstra, H., Babul, A., & Mahdavi, A. 2009, ApJ, 691, 1307
Klypin, A., & Holtzman, J. 1997, arXiv:astro-ph/9712217
Klypin, A., Yepes, G., Gottlober, S., Prada, F., & Hess, S. 2014, arXiv:1411.4001
Lau, E. T., Kravtsov, A. V., & Nagai, D. 2009, ApJ, 705, 1129
Mamon, G. A., Biviano, A., & Boué, G. 2013, MNRAS, 429, 3079
Mamon, G. A., Biviano, A., & Murante, G. 2010, A&A, 520, A30
Mantz, A., Allen, S. W., Ebeling, H., Rapetti, D., & Drlica-Wagner, A. 2010, MNRAS, 406, 1773
Mantz, A. B., von der Linden, A., Allen, S. W., et al. 2015, MNRAS, 446, 2205
Mortonson, M. J., Hu, W., & Huterer, D. 2011, PhRvD, 83, 023015
Munari, E., Biviano, A., Borgani, S., Murante, G., & Fabjan, D. 2013, MNRAS, 430, 2638
Nagai, D. 2006, ApJ, 650, 538
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563
Noh, Y., & Cohn, J. D. 2012, MNRAS, 426, 1829
Ntampaka, M., Trac, H., Cisewski, J., & Price, L. C. 2016, arXiv:1602.01837
Ntampaka, M., Trac, H., Sutherland, D. J., et al. 2015, ApJ, 803, 50
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
Old, L., Wojtak, R., Mamon, G. A., et al. 2015, MNRAS, 449, 1897
Pearson, R. J., Ponman, T. J., Norberg, P., Robotham, A. S. G., & Farr, W. M. 2015, MNRAS, 449, 3082
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014a, A&A, 571, A16
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014b, A&A, 571, A20
Rasia, E., Tormen, G., & Moscardini, L. 2004, MNRAS, 351, 237
Ribeiro, A. L. B., Lopes, P. A. A., & Trevisan, M. 2011, MNRAS, 413, L81
Riebe, K., Partl, A. M., Enke, H., et al. 2013, AN, 334, 691
Rines, K., & Diaferio, A. 2006, AJ, 132, 1275
Rines, K., Geller, M. J., & Diaferio, A. 2010, ApJL, 715, L180
Rines, K., Geller, M. J., Diaferio, A., & Kurtz, M. J. 2013, ApJ, 767, 15
Rines, K., Geller, M. J., Kurtz, M. J., & Diaferio, A. 2003, AJ, 126, 2152
Rozo, E., Wechsler, R. H., Rykoff, E. S., et al. 2010, ApJ, 708, 645
Ruel, J., Bazin, G., Bayliss, M., et al. 2014, ApJ, 792, 45
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 (Cambridge, MA: MIT Press)
Schuecker, P., Böhringer, H., Collins, C. A., & Guzzo, L. 2003, A&A, 398, 867
Sehgal, N., Trac, H., Acquaviva, V., et al. 2011, ApJ, 732, 44
Serra, A. L., Diaferio, A., Murante, G., & Borgani, S. 2011, MNRAS, 412, 800
Sifón, C., Menanteau, F., Hasselfield, M., et al. 2013, ApJ, 772, 25
Sunyaev, R. A., & Zeldovich, Y. B. 1972, CoASP, 4, 173
Sutherland, D. J., Xiong, L., Póczos, B., & Schneider, J. 2012, arXiv:1202.0302
Svensmark, J., Wojtak, R., & Hansen, S. H. 2014, arXiv:1405.0284
Vanderlinde, K., Crawford, T. M., de Haan, T., et al. 2010, ApJ, 722, 1180
Vikhlinin, A., Kravtsov, A. V., Burenin, R. A., et al. 2009, ApJ, 692, 1060
Voit, G. M. 2005, RvMP, 77, 207
von der Linden, A., Best, P. N., Kauffmann, G., & White, S. D. M. 2007, MNRAS, 379, 867
Wang, Q., Kulkarni, S., & Verdu, S. 2009, ITIT, 55, 2392
White, M., Cohn, J., & Smit, R. 2010, MNRAS, 408, 1818
Wojtak, R. 2013, A&A, 559, A89
Wojtak, R., Łokas, E. L., Mamon, G. A., et al. 2007, A&A, 466, 437
Wu, H.-Y., Hahn, O., Evrard, A. E., Wechsler, R. H., & Dolag, K. 2013, MNRAS, 436, 460
Zu, Y., & Weinberg, D. H. 2013, MNRAS, 431, 3319
Zwicky, F. 1933, AcHPh, 6, 110