

Hierarchical inference of the relationship between concentration and mass in galaxy groups and clusters

Maggie Lieu,^{1,2★} Will M. Farr,¹ Michael Betancourt,^{3,4} Graham P. Smith,¹ Mauro Sereno^{5,6} and Ian G. McCarthy⁷

¹*School of Physics and Astronomy, University of Birmingham, Birmingham B15 2TT, UK*

²*European Space Astronomy Centre (ESA/ESAC), Camino bajo del Castillo, E-28692 Villanueva de la Cañada, Madrid, Spain*

³*Applied Statistics Center, Columbia University, New York, NY 10027, USA*

⁴*Department of Statistics, University of Warwick, Coventry CV4 7AL, UK*

⁵*INAF, Osservatorio Astronomico di Bologna, via Ranzani 1, I-40127 Bologna, Italy*

⁶*Dipartimento di Fisica e Astronomia, Università di Bologna, viale Berti Pichat 6/2, I-40127 Bologna, Italy*

⁷*Astrophysics Research Institute, Liverpool John Moores University, Liverpool L3 5RF, UK*

Accepted 2017 March 16. Received 2017 March 16; in original form 2017 January 2

ABSTRACT

Mass is a fundamental property of galaxy groups and clusters. In principle, weak gravitational lensing will enable an approximately unbiased measurement of mass, but parametric methods for extracting cluster masses from data require the additional knowledge of halo concentration. Measurements of both mass and concentration are limited by the degeneracy between the two parameters, particularly in low-mass, high-redshift systems where the signal to noise is low. In this paper, we develop a hierarchical model of mass and concentration for mass inference, we test our method on toy data and then apply it to a sample of galaxy groups and poor clusters down to masses of $\sim 10^{13} M_{\odot}$. Our fit and model gives a relationship among masses, concentrations and redshift that allow prediction of these parameters from incomplete and noisy future measurements. Additionally, the underlying population can be used to infer an observationally based concentration–mass relation. Our method is equivalent to a quasi-stacking approach with the degree of stacking set by the data. We also demonstrate that mass and concentration derived from pure stacking can be offset from the population mean with differing values depending on the method of stacking.

Key words: gravitational lensing: weak – methods: statistical – galaxies: clusters: general.

1 INTRODUCTION

Galaxy groups and clusters are some of the largest structures in the observable Universe. They give insight to the growth and evolution of structure through the multiwavelength study of their properties. Knowledge of the abundance and mass of these systems can be used in combination to probe cosmological parameters through the mass function (Voit 2005; Allen, Evrard & Mantz 2011). Although mass is not a direct observable, it can be estimated in a number of ways including hydrostatic mass from the X-ray emission of the hot intracluster medium and the dynamical mass from the velocity dispersions of galaxies. These estimators of mass rely on assumptions that may be biased from the true halo mass, for example X-ray masses could incur a bias of 10–30 per cent (Piffaretti & Valdarnini 2008; Le Brun et al. 2014) from the assumption of

hydrostatic equilibrium. What’s more, mass is generally observationally expensive.

If gravity is the main contributor to the formation of clusters, then we would expect them to follow self-similarity (Kaiser 1986) and have simple power-law relationships between mass and other observable properties known as mass proxies (temperature, luminosity, etc.). These scaling relations are a useful alternative to obtain mass measurements and are observationally cheaper. Nevertheless, scaling relations provide a less accurate estimate of mass and are influenced by the calibration cluster sample (Sun 2012; Giodini et al. 2013).

Weak-lensing mass is a measure of the influence of the cluster gravitational potential on the light path of background galaxies (see e.g. Hoekstra et al. 2013, for a review) and the arising galaxy shape distortion is known as shear. The effect is purely geometrical; it is sensitive only to line-of-sight structures and does not make as many assumptions as other methods, thus it provides a good estimator of the true halo mass. However, lensing masses can suffer from the large scatter and noise. In particular, galaxy groups are

* E-mail: maggie.lieu@sciops.esa.int

$<10^{14} M_{\odot}$ making weak-lensing measurements particularly challenging due to the low shear signal-to-noise ratio (SNR) and individual mass measurements in this context can be strongly biased (Corless & King 2007; Becker & Kravtsov 2011; Bahé, McCarthy & King 2012).

The NFW model (Navarro, Frenk & White 1997) provides a reasonable description of the density profile of clusters, it is given by

$$\rho_{\text{NFW}}(r) = \frac{\rho_s}{(r/r_s)(1 + r/r_s)^2}, \quad (1)$$

where ρ_s is the central density and r_s is a characteristic scale radius at which the slope of the log density profile is -2 . The NFW model can be characterized by two parameters: halo mass M_{Δ} ¹ determines the normalization and concentration $c_{\Delta} = r_{\Delta}/r_s$ determines the radial curvature of the profile. Whilst M is both a physical quantity and a model parameter, c is less well defined; c is a parameter in the NFW profile but may not be equivalent in other density profiles (e.g. Einasto profile; Klypin et al. 2016). Concentration is difficult to constrain due its inherent covariance with mass (Hoekstra et al. 2011; Auger et al. 2013; Sereno et al. 2015) and the degeneracy is particularly high for individual weak-lensing measurements of high-redshift, low-mass systems. Depending on the number of background galaxies, even massive clusters with reasonable shear SNR require the stacking of multiple clusters in order to constrain concentration (Okabe et al. 2013; Umetsu et al. 2014). The radial averaging when stacking helps to smooth out substructures; however, it can be hard to decide which clusters to stack and how to stack them, especially if they span a wide range of redshift and mass. Stacking results in a loss of information, and whilst stacking systems of similar masses would minimize the information loss, we note that mass is known a priori. Alternatively, SNR may seem an appealing property to stack on; however, SNR does not necessarily correlate as expected with mass (see e.g. fig. 3 of Lieu et al. (2016) where redshift can be seen as a proxy for mass, since high-mass clusters tend to be at higher redshifts due to a combination of their formation and selection). One reason being that whilst massive clusters have a higher weak-lensing signal, they are also subject to larger noise because at high redshifts there are less background galaxies and larger photometric redshift uncertainties.

It is therefore common to use a fixed concentration value (Foëx et al. 2012; Oguri et al. 2012; Applegate et al. 2014), or a c – M scaling relation based on numerical simulations to aid constraints on mass (e.g. Duffy et al. 2008; Zhao et al. 2009; Bahé et al. 2012; Dutton & Macciò 2014). The choice of c – M relation is again non-trivial, as dark-matter-only simulations tend to produce high normalization relations compared to those that include baryonic physics and feedback (e.g. Duffy et al. 2010; Velliscig et al. 2014). It is also sensitive to σ_8 and Ω_M , where Duffy et al. (2008)’s c – M relation (which assumes a *WMAP5* cosmology) has 20 per cent lower concentrations than Dutton & Macciò (2014)’s relation (which assumes the Planck 2013 cosmology). These issues will affect both mass and concentration due to parameter degeneracies.

Accurate mass measurements are important for cluster cosmology; however, traditionally, methods to obtain cosmological constraints from the data are divided into separate analyses and work from the bottom-up. For example, observations are made and are processed into data catalogues, the catalogues are used to obtain in-

dividual masses of some clusters where the data quality is adequate to do so, a scaling relation fit is obtained for some mass proxy to allow further mass estimates of clusters where the data quality for mass is poor, and finally the cosmology can be obtained by fitting a mass function. Not only is this inefficient, it is also suboptimal due to the loss of information, possible introduction of biases and the difficulty in consistent propagation of uncertainties at each step.

Here, we instead consider a Bayesian inference model that embeds the global problem into a forward modelling approach and subsequently avoids these many issues. Hierarchical modelling is a unified statistical analysis of the source population and individual systems. The prior distribution on the individual cluster parameters can be seen as a common population distribution and the data can collectively be used to infer aspects of the population distribution that is otherwise not observed. In traditional non-hierarchical methods, introducing too few model parameters produces inaccurate fits to large data sets and too many parameters runs the risk of overfitting the data. By treating the problem as a hierarchical model (see e.g. Schneider et al. 2015; Alsing et al. 2016; Sereno & Ettori 2016), we have enough parameters to fit the data well when possible; the population distribution accounts for a full statistical dependence of all parameters when not otherwise constrained by data. This ‘quasi-stacking’ approach enables improved estimates on weakly constrained parameters such as concentration and masses of low SNR clusters by incorporating information from the population in a principled way.

In this paper, we propose a method to exploit the underlying cluster population properties in order to improve constraints on weak-lensing masses of individual groups and poor clusters. The data are fitted with the assumption that the parameters originate from the same underlying population. In the case of mass and concentration, the distribution of the population mass and concentration is a prior on the corresponding individual cluster parameters. This method is therefore fully self-consistent with the data and makes it possible to constrain concentration of each cluster without the need of full stacking. It works well even with low signal-to-noise data that will be important for future weak-lensing surveys, where the observations may be shallow such as DES² and KIDS.³

This paper is structured as follows: in Section 2, we describe in detail the hierarchical model and outline how it can be used for parameter prediction (Section 3). This is tested on toy data and simulations in Section 4 and applied to a typical shallow data sample in Section 5. Discussions are presented in Section 6 and finally we conclude in Section 7. Throughout, the *WMAP9* (Hinshaw et al. 2013) cosmology of $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_M = 0.28$ and $\Omega_{\Lambda} = 0.72$ is assumed. All statistical errors are reported to 68 per cent credibility and all mass values are reported in units $h_{70}^{-1} M_{\odot}$, unless otherwise stated.

2 METHOD

Our model assumes each cluster can be described by n parameters. We assume that the distribution of the parameters for a population of clusters is described by a multivariate Gaussian with a global mean n -vector μ and a $n \times n$ covariance matrix Σ that describes the intrinsic scatter of each property and the covariances between

¹ M_{Δ} is the mass within which the mean density is Δ times the critical density at the cluster redshift.

² <http://www.darkenergysurvey.org>

³ <http://kids.strw.leidenuniv.nl>

them. For now, we focus on the cluster mass M_{200} , concentration c_{200} and redshift z . Therefore $n = 3$,

$$\mu = \begin{pmatrix} \frac{\ln(M_{200})}{\ln(c_{200})} \\ \ln(1+z) \end{pmatrix},$$

and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}.$$

Here, the subscripts 1, 2, 3 on σ represent $\ln M_{200}$, $\ln c_{200}$ and $\ln(1+z)$, respectively, and ρ is the correlation coefficient.

The true distribution for the population mass should be the cluster mass function that describes the number density of clusters of a given mass and redshift (e.g. Tinker et al. 2008). Massive clusters form from rare, dense peaks in the initial mass density fluctuations of early Universe so are less abundant than poor clusters and low-mass groups that form from smaller more common fluctuations. However, the least massive systems are also the least luminous and are therefore less likely to be detected than more luminous massive clusters. This selection function causes a decrease in the number of clusters observed at low mass due to survey sensitivity limits. Although in detail, the cluster selection function will not be lognormal; here, we justify the use of a lognormal distribution as an approximation to the cluster mass function and selection function (see also Sereno et al. 2015). This is also motivated for conjugacy since simulations of the cluster concentration mass distribution shows lognormal scatter (Jing 2000; Bullock et al. 2001; Duffy et al. 2008; De Boni et al. 2013) and the results of Lieu et al. (2016) show redshift and mass distributions that are close to Gaussian.

2.1 Hyperparameters

It is common to call the parameters that describe the population (μ and Σ) hyperparameters, and the priors on them, hyperpriors. The covariance matrix Σ is a difficult parameter to sample since by definition it must be both symmetric and positive definite. Therefore for its prior, we take the Stan Development Team (2016a) recommended approach, which decomposes Σ into a correlation matrix Ω and a scale vector τ (Barnard, McCulloch & Meng 2000):

$$\Sigma = \text{diag}(\tau)\Omega\text{diag}(\tau), \quad (2)$$

where τ is a vector of the standard deviations of the hyperparameter μ that describe the population mean. The prior on τ is taken to be a Gamma distribution with shape $\alpha_\tau = 2$ and rate $\beta_\tau = 3$,

$$\Pr(\tau|\alpha_\tau, \beta_\tau) = \frac{\beta_\tau^{\alpha_\tau}}{\Gamma(\alpha_\tau, 1)} \tau^{\alpha_\tau-1} \exp(-\beta_\tau \tau). \quad (3)$$

This prior is chosen so as to prevent divergences (see Section 2.3) in the sampling whilst allowing large values of variance. An LKJ distribution prior (Lewandowski, Kurowicka & Joe 2009) is used on the correlation,

$$\Pr(\Omega|\nu) \propto \det(\Omega)^{\nu-1}, \quad (4)$$

where the shape parameter $\nu > 0$. This distribution converges towards the identity matrix as ν increases, allowing the control of the correlation strength between the multiple parameters and consequently the variance and covariance of parameters in the population. A flat prior can be imposed by setting $\nu = 1$ and for $0 < \nu < 1$, the density has a trough at the identity matrix. However, to optimize

our code, we decompose the correlation matrix Ω into its Cholesky factor L_Ω and its transpose L_Ω^\top ,

$$\Omega = L_\Omega L_\Omega^\top \quad (5)$$

$$\Pr(\Omega|\nu) = \prod_{k=2}^K L_{kk}^{K-k+2\nu-2}, \quad (6)$$

and implement on L_Ω an LKJ prior parametrized in terms of the Cholesky decomposition setting $\nu=10$, i.e. weakly preferring identity. For the global mean vector, we use a weakly informative prior

$$\Pr(\mu|\mu_0, \Sigma_0) = \frac{1}{\sqrt{2\pi\Sigma_0}} \exp\left[-\frac{(\mu - \mu_0)^2}{2\Sigma_0}\right], \quad (7)$$

where $\mu_0=(32,1,0.3)$ and $\Sigma_0 = (1, 1, 1)$. The priors and hyperpriors chosen in our model are consistent with the knowledge of these systems, since we expect masses to lie between $10^{13-16} M_\odot$ and concentration between 0 and 10. Using the prior information helps to regularize the inference and avoids numerical divergences since the projected NFW profile is numerically unstable (in particular at the *Einstein* radius). We test the sensitivity of our results to these choices of hyperpriors in Section 6.1.

Rather than using the Gamma prior on the scale and the LKJ prior on the correlation, it is more common in these sorts of hierarchical analyses to set the prior on Σ to be the scaled inverse Wishart distribution (Gelman & Hill 2006). This choice is usually made for its conjugacy on Gaussian likelihoods and simplicity within Gibbs sampling. However, this distribution undesirably assumes a prior relationship between the variances and correlations (see Alvarez, Niemi & Simpson 2014, for a review on priors for covariance matrices). In our sampling method, which we discuss in Section 2.3, conjugate priors are not necessary and, in fact, the combined scale-LKJ prior is more efficiently sampled and gives us control over the diagonal elements of Σ .

2.2 Sample parameters

The parameters that describe the properties of the i th cluster x_i are assumed to be drawn from the population distribution. We chose a centred parametrization to draw cluster parameters from the population as the non-centred parametrization (Betancourt & Girolami 2015) suffered from biases and subpar performance as indicated by sampler diagnostics:

$$x \sim \mathcal{N}(\mu, LL^\top), \quad (8)$$

where L is the Cholesky decomposition of Σ .

This re-parametrization is equivalent to drawing from a multivariate Gaussian but is less computationally expensive since the covariance matrix is only decomposed once. It makes for more efficient sampling of the deformed regions of the parameter space commonly found in hierarchical inference problems. The probability of the parameters conditional on the global population takes the form of a multivariate Gaussian distribution:

$$\Pr(x|\mu, \Sigma) = \prod_i \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left[-\frac{1}{2}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu)\right], \quad (9)$$

where $n = 3$ and

$$x_i = \begin{pmatrix} \ln(M_{200}^{(i)}) \\ \ln(c_{200}^{(i)}) \\ \ln(1+z^{(i)}) \end{pmatrix}.$$

2.3 Model fitting

The full posterior can be written as

$$\Pr(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{x} | \mathbf{d}) = \frac{\Pr(\mathbf{d} | \mathbf{x}) \Pr(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Pr(\boldsymbol{\mu}) \Pr(\boldsymbol{\Sigma})}{\Pr(\mathbf{d})}, \quad (10)$$

where \mathbf{x} are the individual cluster parameters and \mathbf{d} are the data (shear profiles and spectroscopic redshifts). The likelihood is

$$\Pr(\mathbf{d} | \mathbf{x}) = \prod_i \frac{1}{\sqrt{(2\pi)\sigma_{i,z}}} \exp \left[-\frac{(d_{i,z} - z(\mathbf{x}_i))^2}{2\sigma_{i,z}^2} \right] \times \prod_j \frac{1}{\sqrt{(2\pi)\sigma_{i,j}}} \exp \left[-\frac{(d_{i,j} - g(r_{i,j}, \mathbf{x}_i))^2}{2\sigma_{i,j}^2} \right], \quad (11)$$

where $d_{i,z}$ and $\sigma_{i,z}$ are the observed redshift and associated uncertainty of the i th cluster, $d_{i,j}$ and $\sigma_{i,j}$ are the observed shear and associated uncertainty of the i th cluster in the j th radial bin, z is the redshift associated with parameters \mathbf{x} , and g is the model shear at the radius $r_{i,j}$ from the cluster centre. The model shear is a function of the mass, concentration and redshift as computed according to an NFW (Navarro et al. 1997) density profile (see the Appendix A). Regardless of the shear SNR, we do not fix the concentration to values from a mass–concentration relation; instead information on the relationship between c–M flows through the population distribution that is simultaneously fit to our data set. We treat the quoted shear measurements as the fundamental data product, and assume above in equation (11) that the sampling distribution for the shear is Gaussian with width equal to the quoted shear uncertainties. In reality, the fundamental data product of a weak-lensing measurement is pixel-level images of background galaxies, and the summary of this data by shear measurements induces a distribution that is not Gaussian, but equation (11) is a reasonable and computationally efficient approximation. See Schneider et al. (2015) for a discussion of hierarchically modelled pixel-level likelihood functions for shear maps; such models provide a more accurate representation of the data but through a much more complicated and expensive likelihood function. A graphical model of our posterior appears in Fig. 1.

The Stan probabilistic coding language is used to implement inference on our problem with the R interface Stan Development Team (2016b). Stan samples from posterior distributions using a Hamiltonian Monte Carlo (HMC) algorithm (Neal 2011; Betancourt et al. 2016). HMC is a Markov Chain Monte Carlo (MCMC) sampling method where proposed states are determined by a Hamiltonian dynamics model. This enables more efficient exploration of the parameter space and hence faster convergence in typical problems that is crucial for problems working in high dimensions.

We run four chains with 1000 warm-up samples followed by 1000 monitored samples. Convergence is checked using trace plots, histograms of the tree depth and calculation of the Gelman–Rubin convergence criterion ($\hat{R} < 1.1$; Gelman & Rubin 1992). Sample bias is also checked by monitoring the number of divergences in a given sample. This diagnostic is specific to HMC, it indicates the number of numerical divergences occurred whilst sampling and is typical for regions of the parameter space that are hard to explore. Any number of divergences could suggest a bias in the posterior samples; however, it can be reduced by increasing the acceptance probability, or by re-parametrizing the model.

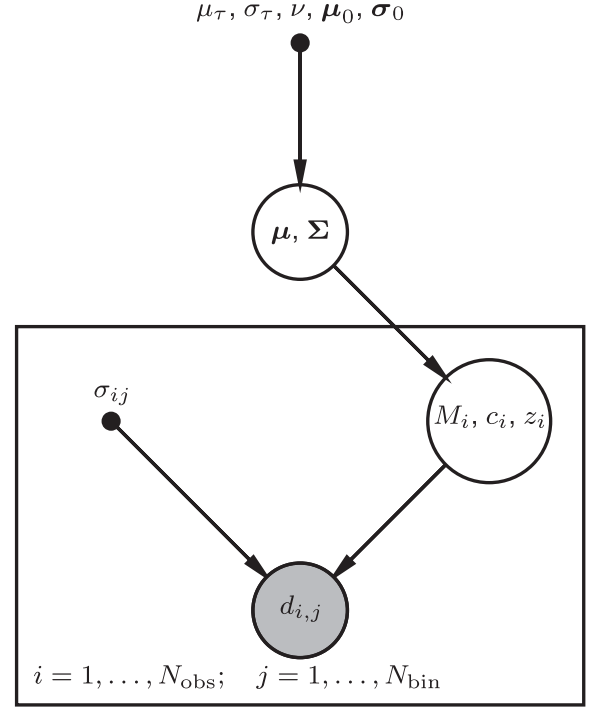


Figure 1. A graphical model of the relationships between terms in our posterior. Filled ellipses indicate quantities that are observed and therefore conditioned-on in the analysis, whilst open ellipses contain parameters that are fit, and dots indicate fixed quantities that are not probabilistically modelled. At the top level, the (fixed) parameters controlling the hyperpriors influence the distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ control the distribution of masses, concentrations and redshifts of the individual clusters. The mass, concentration and redshift of each cluster combine with the (fixed) observational uncertainties to control the distribution of the shear data.

3 PREDICTING FUTURE DATA AND SCALING RELATIONS

In order to use the results from the hierarchical model to predict parameters of future data, consider the following situation. We observe or produce noisy estimates of (some of the) parameters of a previously unobserved system that we assume comes from the same population, and we now wish to use the population-level fitting to produce better estimates and/or predictions of parameters that we did not measure. Let us assume that the observational uncertainties are Gaussian, described by a mean $\boldsymbol{\mu}_o$ and a covariance matrix $\boldsymbol{\Sigma}_o$. (If a parameter is unobserved, then we can set the corresponding diagonal element of the covariance matrix to ∞ , indicating infinite uncertainty about its value.) The true parameters of the system of interest are

$$\mathbf{x}_T = \{M_T, c_T, z_T\}, \quad (12)$$

where we use M_T , c_T , z_T as short-hand for the true underlying parameters $\ln M_{200}$, $\ln c_{200}$, $\ln(1+z)$. Combining the results of the new observations with our population model results in a Gaussian distribution for the true parameters of the system with via,

$$\mathbf{x}_T \sim \mathcal{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T), \quad (13)$$

where

$$\begin{aligned} \boldsymbol{\mu}_T &= \boldsymbol{\Sigma}_T (\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\Sigma}_o^{-1} \boldsymbol{\mu}_o) \\ \boldsymbol{\Sigma}_T &= (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_o^{-1})^{-1}. \end{aligned} \quad (14)$$

If we specialize to uncorrelated measurements, then

$$\Sigma_o^{-1} = \text{diag}(\sigma_M^2, \sigma_c^2, \sigma_z^2)^{-1} \quad (15)$$

is a diagonal matrix with the uncertainties associated with each measurement.

The posterior on the true parameters of an individual system is a normal distribution about the weighted mean of the population μ and the observable values μ_o . The uncertainties are similarly dependent both on the population width Σ and the observational uncertainty Σ_o . A small observable uncertainty will cause the parameter to be dominated by the observed value, whereas a large observable uncertainty will pull the parameter closer to the population estimate. This effect is particularly useful for measurements of low signal-to-noise data. Where observables are missing, for example a measurement of a mass and redshift but no measurement of concentration, the hierarchical model can still be used as described above by setting $\sigma_c = \infty$. In this particular case, the estimate of μ_T^c would be weighted entirely by the population distribution at the appropriate values of M and z . We now proceed to derive equation (13).

Using Bayes theorem, the conditional distribution of the true parameters can be written as

$$\begin{aligned} \Pr(\mathbf{x}_T | \mathbf{x}_o, \Sigma_o, \mu, \Sigma) &\propto \Pr(\mathbf{x}_o | \mathbf{x}_T, \Sigma_o, \mu, \Sigma) \Pr(\mathbf{x}_T | \Sigma_o, \mu, \Sigma) \\ &\propto \exp \left[-\frac{1}{2} (\mathbf{x}_o - \mathbf{x}_T)^\top \Sigma_o^{-1} (\mathbf{x}_o - \mathbf{x}_T) \right] \\ &\times \exp \left[-\frac{1}{2} (\mathbf{x}_T - \mu)^\top \Sigma^{-1} (\mathbf{x}_T - \mu) \right] \\ &\propto \exp \left[-\frac{1}{2} ((\mathbf{x}_o - \mathbf{x}_T)^\top \Sigma_o^{-1} (\mathbf{x}_o - \mathbf{x}_T) \right. \\ &\quad \left. + (\mathbf{x}_T - \mu)^\top \Sigma^{-1} (\mathbf{x}_T - \mu)) \right]. \end{aligned}$$

The log posterior is thus proportional to a Gaussian distribution:

$$\mathcal{L} = -\frac{1}{2} ((\mathbf{x}_o - \mathbf{x}_T)^\top \Sigma_o^{-1} (\mathbf{x}_o - \mathbf{x}_T) + (\mathbf{x}_T - \mu)^\top \Sigma^{-1} (\mathbf{x}_T - \mu)).$$

The posterior mean of \mathbf{x}_T occurs at the maxima of the likelihood, where the derivative of \mathcal{L} is 0. The posterior variance is the inverse of the negative second derivative of the \mathcal{L} . The first and second derivatives of the loglikelihood are

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}_T} = \Sigma_o^{-1} (\mathbf{x}_o - \mathbf{x}_T) + \Sigma^{-1} (\mu - \mathbf{x}_T), \quad (16)$$

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{x}_T^2} = -\Sigma_o^{-1} - \Sigma^{-1}. \quad (17)$$

Setting $\partial \mathcal{L} / \partial \mathbf{x}_T = 0$ and solving for $\mathbf{x}_T \equiv \mu_T$ yields

$$\mu_T = \Sigma_T (\Sigma_o^{-1} \mathbf{x}_o + \Sigma^{-1} \mu). \quad (18)$$

The variance Σ_T is

$$\Sigma_T = - \left(\frac{\partial^2 \mathcal{L}}{\partial \mathbf{x}_T^2} \right)^{-1} = (\Sigma_o^{-1} + \Sigma^{-1})^{-1}, \quad (19)$$

recovering the equations defined earlier.

3.1 Scaling relations

We can use the formalism above to derive scaling relations between the parameters in our population model. A scaling relation is obtained when two parameters are measured with zero uncertainty

and a third is unmeasured (i.e. with infinity uncertainty). For example, to compare with existing c-M relations in the literature, we can assume that we measure mass and redshift perfectly and with no uncertainty, i.e. $\sigma_M = \sigma_z = 0$, $x_o^m = x_T^m$, $x_o^z = x_T^z$ and measure concentration with infinite uncertainty, i.e. $\sigma_{o,c} \rightarrow \infty$, implying $\Sigma_{o,22}^{-1} \rightarrow 0$

$$\mu_T^c = \frac{\Sigma_{12}^{-1}}{\Sigma_{22}^{-1}} (\mu^m - x_T^m) + \frac{\Sigma_{23}^{-1}}{\Sigma_{22}^{-1}} (\mu^z - x_T^z) + \mu^c. \quad (20)$$

If we replace μ_T^c , x_T^m and x_T^z by $\ln(c_{200})$, $\ln(M_{200})$ and $\ln(1+z)$, respectively,

$$\ln(c) = \frac{\Sigma_{12}^{-1}}{\Sigma_{22}^{-1}} (\mu^m - \ln(M)) + \frac{\Sigma_{23}^{-1}}{\Sigma_{22}^{-1}} (\mu^z - \ln(1+z)) + \mu^c, \quad (21)$$

then we can rearrange into the familiar multiple regression form

$$\ln(c) = \alpha + \beta \ln(M) + \gamma \ln(1+z), \quad (22)$$

where

$$\begin{aligned} \alpha &= \frac{\Sigma_{12}^{-1}}{\Sigma_{22}^{-1}} \mu^m + \frac{\Sigma_{23}^{-1}}{\Sigma_{22}^{-1}} \mu^z + \mu^c \\ \beta &= -\frac{\Sigma_{12}^{-1}}{\Sigma_{22}^{-1}} \\ \gamma &= -\frac{\Sigma_{23}^{-1}}{\Sigma_{22}^{-1}}. \end{aligned}$$

If we instead assume that concentration and redshift are measured perfectly, we derive a different scaling relation that is *algebraically inequivalent* to the above relation because of the different assumptions ($\sigma_M = 0$ versus $\sigma_c = 0$). In the event that any of these quantities are actually measured, with associated non-zero uncertainty, it is better to use the full formalism from Section 3 (equation 13) that takes into account measurement uncertainty than to substitute into a scaling relation that ignores it.

4 TESTS ON SIMULATED DATA

We test our model on toy data by generating shear profiles for 38 clusters, each with eight radial bins spaced equally in log. The masses, concentrations and redshifts are drawn from an arbitrary multivariate distribution of mean $\mu = \{\ln(2 \times 10^{14}), \ln(3), \ln(1+0.3)\}$ and covariance $\Sigma = \{(1.1, -0.1, 0.05), (-0.1, 0.4, 0.05), (0.05, 0.05, 0.01)\}$. We note that the definition of the model in this form is not ideal since the $\ln(1+z)$ component in μ implies z can take negative values. It would be more natural to be expressed in the form $\ln(z)$; however, this would then not allow the direct inference of the c-M relation in its commonly expressed form. The $(1+z)$ factor comes from the expansion factor of the Universe so is therefore physically motivated. For this reason, we make sure that the cluster redshifts are positive. Trials of various levels of uncertainty on the shear measurements are made and we are able to recover parameters to within 2 per cent uncertainty with the exception of the mean population concentration that is biased increasingly low as the uncertainty on the shear increases (Fig. 2). None the less, the fitted values agree within the uncertainties and the bias is reduced significantly when increasing the sample size from 38 to 200 clusters (Fig. 3), which is promising for the application of this work on upcoming large cluster surveys.

We also test our model on the cosmo-OWLS cosmological hydrodynamical simulations (Le Brun et al. 2014). The large volume runs consists of 400 Mpc h^{-1} on a side periodic boxes with 2×1024^3 particles, ideal for the study of cluster populations. We use the

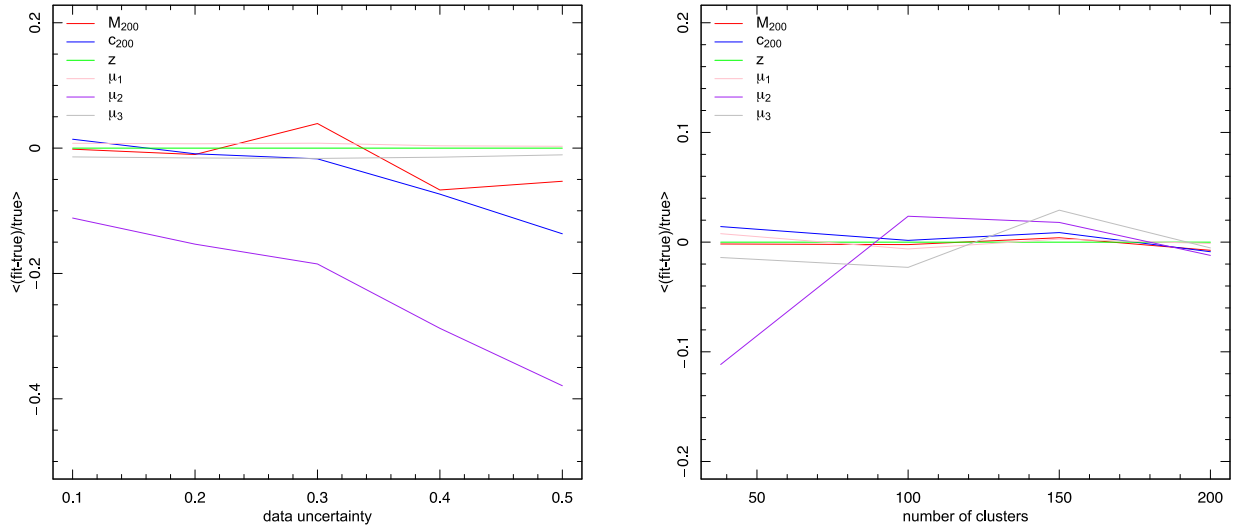


Figure 2. The bias in measured parameters for toy simulations. The subscripts 1, 2, 3 on μ represent the $\ln(M_{200}[h_{70}^{-1} M_{\odot}])$, $\ln(c_{200})$ and $\ln(1+z)$ population components, respectively. Decreasing shear uncertainty (left) and increasing cluster sample (right) improves the ability to reproduce the truth.

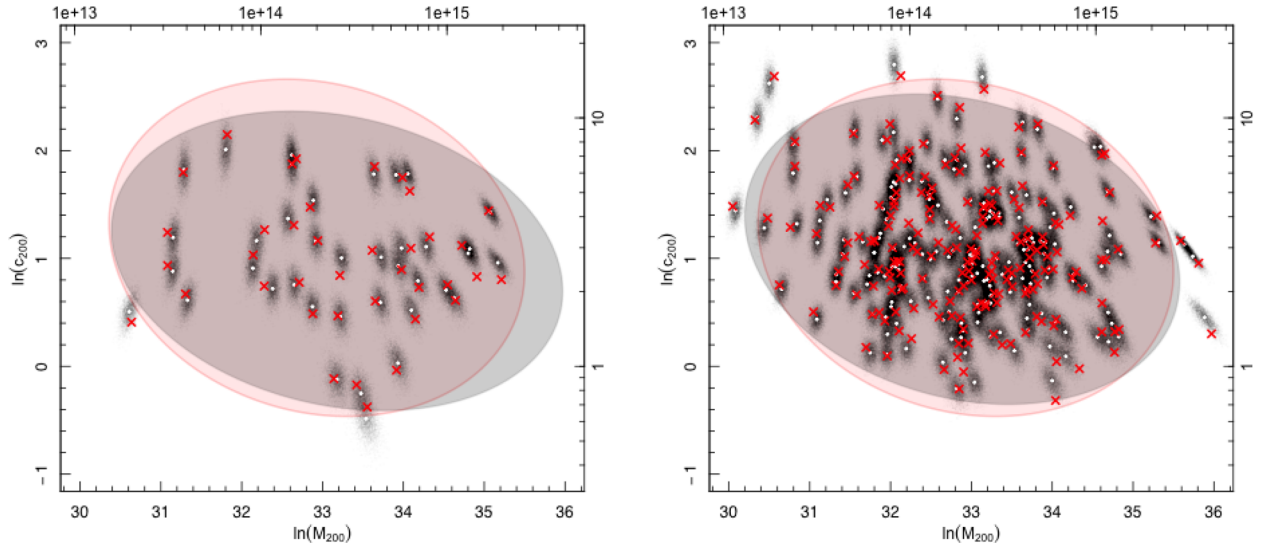


Figure 3. The results of the toy simulations of 38 and 200 clusters with 10 per cent error on shear. The true population is shown by the red ellipse (95 per cent region) and the black ellipse is based upon the fitted values of mean population parameters (note that it does not take into account the posteriors on the population, only the point estimates $\bar{\mu}$ and $\bar{\Sigma}$). The red crosses indicate the input concentration and mass values and the white points show the mean of the fitted values.

dark-matter-only run with *WMAP7* cosmology, five source galaxies arcmin^{-2} and 28 per cent shape noise. Clusters are drawn randomly from the redshift slice $z = 0.25$ such that in the range of $13 \leq \log_{10}(M_{500}) \leq 15$, mass bins of 0.25 dex width contain 100 objects when possible. The total number of clusters in the sample is 632. Omitting the redshift component from our model, we recover the cluster parameters and population estimates reasonably well (see Fig. 4a). As expected, the constraints on mass are better than those on concentration. At high mass, the data have high signal to noise, so individually measured mass values give a good estimate of the true number of clusters. However, at lower mass this is not true, and consequently, we observe an overestimation of the clusters at the mean mass of the sample. Meanwhile, the number of clusters predicted from the population shows good agreement at all mass scales (Fig. 4b). Implementation of the combined selection function and halo mass function as a hyperprior distribution is complex and

will be important for cosmological inferences; therefore, we leave it to a future paper.

The observed effect of parameter estimates influenced by the population mean is a property of hierarchical models known as shrinkage. It improves parameter estimates of both individual and hyperparameters as the sample size increases. To demonstrate this, we compare mass estimates to those obtained in a non-hierarchical manner (Fig. 5a). Naively, one may believe that the shrinkage nature is caused by the high-mass (and consequently high SNR) systems dominating the fit and flattening the low-mass end; however, we demonstrate that this is not the case. Whilst the high-mass systems tend to have a higher signal to noise, quantitatively there are many more low-mass clusters. For this reason, we see that the high-mass bin moves down a lot more than low-mass bins when comparing to masses determined in a non-hierarchical manner (note that the axes are in log scale). More importantly, we again stress that

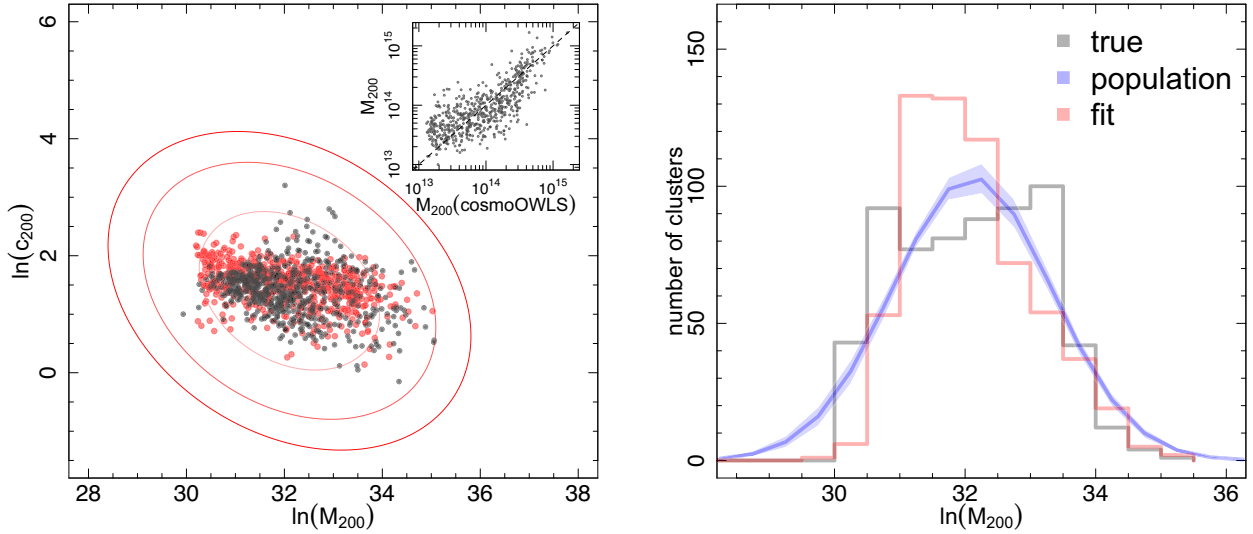


Figure 4. Left: true mass and concentration (red) of 632 cosmo-OWLS clusters at $z = 0.25$ against fitted parameters in this work (black), the contour is derived from the point estimates of the population parameters μ , Σ and inset is a comparison of masses. Right: a histogram of the true cluster masses from simulations (black) and the comparison with the individual fitted masses in this work (red). The shrinkage can be seen clearly from the overestimation at $\ln(M_{200}) = 32$ in the fitted values and the underestimation at tails. From the point estimates of the fitted population parameters (blue), we recover a much better fit to the truth but this can be further improved if assume a population with the form of the cluster mass function rather than that of a Gaussian.

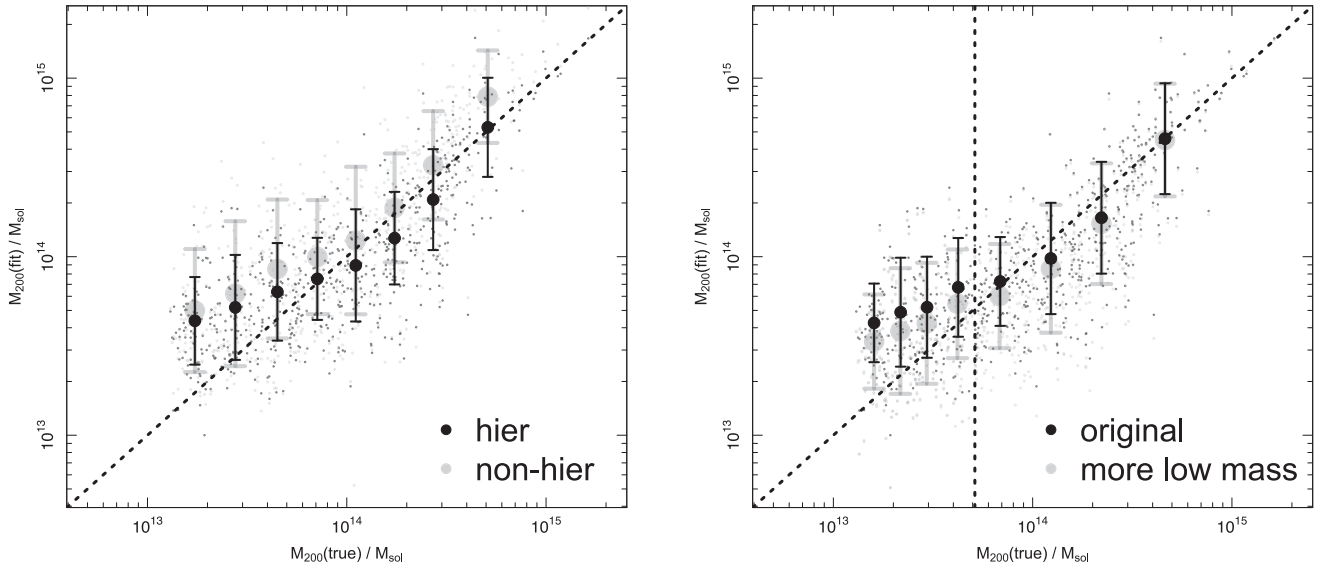


Figure 5. Left: comparison between hierarchical and non-hierarchical methods applied to cosmo-OWLS simulations. For clarity, objects have been grouped into eight bins of 79 clusters ordered in mass. Right: comparison of masses inferred from the hierarchical method when the number of low-mass systems are increased. All clusters (200) to the left of the vertical dotted line are included twice in the fit. Again for clarity, objects are binned to have equal number of clusters in the ‘more low-mass’ run.

combining the masses of determined individually is subject to improper propagation of errors and biases when used in future applications such as cluster mass function because of the information loss in relying on point estimates. Using the population estimate on the other hand correctly utilizes the full posterior of each cluster system and therefore is not prone to this effect.

On the contrary, to ensure that the high-mass clusters are not biased low by the shrinkage model, we rerun the hierarchical model on a sample where the 200 lowest mass clusters are duplicated (Fig. 5b). As expected, the highest mass clusters are not affected by the increase in low-mass systems since their signal alone is able to constrain the mass however at lower masses, the SNR

is lower and therefore more vulnerable to the increase in lower mass systems.

5 APPLICATION TO DATA

5.1 Data

We further apply our method to observational data from Lieu et al. (2016). Here, we provide a brief summary.

The sample consists 38 spectroscopically confirmed groups and poor clusters that lie at $0.05 < z < 0.6$ and span the low temperature range of $T_{300\text{ kpc}} \simeq 1\text{--}5\text{ keV}$ (Giles et al. 2016). They are selected in

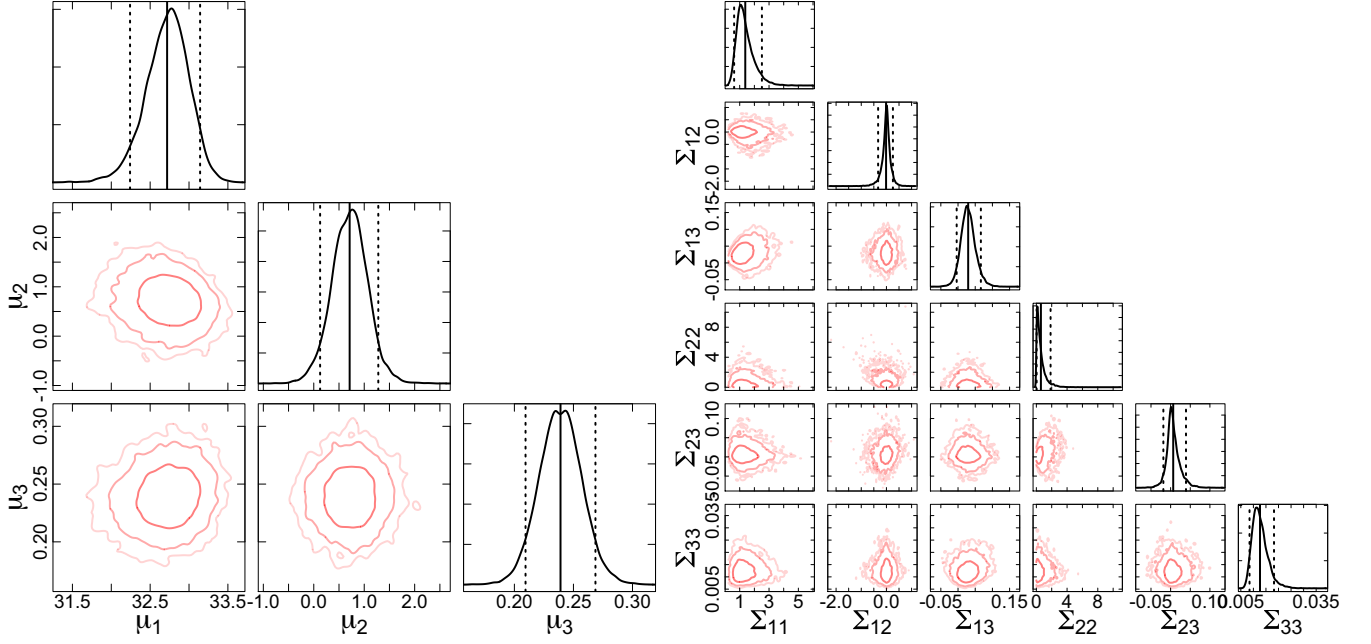


Figure 6. Posterior distributions of the nine hyperparameters, where the subscripts 1, 2, 3 represent $\ln(M_{200}[h_{70}^{-1} M_{\odot}])$, $\ln(c_{200})$ and $\ln(1+z)$, respectively. The red contours show 68, 95 and 99 percent confidence intervals, the histograms show the marginalized parameters with dashed vertical lines at 2σ . Left: global mean vector parameters. Right: covariance matrix elements.

X-ray to be the 100 brightest systems⁴ and collectively lie within both the Northern field of the XXL survey (Pierre et al. 2016) and the CFHTLenS survey⁵ (Heymans et al. 2012; Erben et al. 2013). The clusters are confined to $z < 0.6$ due to limited depth of the CFHTLenS survey, this corresponds to a background galaxy cut of $\sim 4 \text{ arcmin}^{-2}$. The sample is not simply flux-limited, the systems are selected based upon both count rate and extension (see Pacaud et al. 2016, for details).

We use shear profiles as computed in Lieu et al. (2016) that are distributed into eight radial bins equally spaced on the log scale. They use a minimum threshold of 50 galaxies per radial bin which if not met is combined with the subsequent radial bin. In the future, we intend to extend the method to the full shear catalogue without binning. The errors on the shear are computed using bootstrap resampling with 10^3 samples and incorporate large-scale structure covariance. All 38 clusters have spectroscopic redshifts, therefore we are able to use this information as data within the model. Our model applied to the XXL data set is 123-dimensional (3×38 cluster parameters and nine population-level parameters).

5.2 Global estimates

The posteriors of the hyperparameters approximately follow Gaussian distributions (Fig. 6). This justifies the use of the posterior mean and standard deviation as the estimator of the

fits. For the global mean vector and covariance matrix, these are

$$\mu = \begin{pmatrix} 32.718 \pm 0.278 \\ 0.711 \pm 0.357 \\ 0.239 \pm 0.018 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 1.379 \pm 0.609 & -0.014 \pm 0.190 & 0.030 \pm 0.022 \\ -0.014 \pm 0.190 & 0.593 \pm 0.644 & 0.007 \pm 0.018 \\ 0.030 \pm 0.022 & 0.007 \pm 0.018 & 0.013 \pm 0.003 \end{pmatrix}.$$

A comparison between the population z distribution and the distribution of spectroscopic redshifts of the sample acts as a reassurance that the model is indeed working. We also compare the posteriors of μ_M and μ_c to the posteriors of M_{200} and c_{200} of the individual clusters (Fig. 7). The individual concentration values are weakly constrained resulting in posteriors that are dominated by the population mean, whereas the individual masses are able to suppress the influence of the mean mass. This demonstrates that independently the individual clusters could not have constrained a concentration value.

5.3 Mass estimates

We find smaller masses to those computed independently from the individual shear profiles in Lieu et al. (2016, Fig. A1).

We calculate the weighted geometric mean between two mass estimates of n clusters as

$$\langle M_1/M_2 \rangle = \exp \left(\frac{\sum_{i=1}^n w_i \ln \left(\frac{M_{1,i}}{M_{2,i}} \right)}{\sum_{i=1}^n w_i} \right). \quad (23)$$

⁴ XXL-100-GC data are available in computer readable form via the XXL Master Catalogue browser <http://cosmosdb.iasf-milano.inaf.it/XXL> and via the XMM XXL Database <http://xmm-lss.in2p3.fr>.

⁵ www.cfhtlens.org

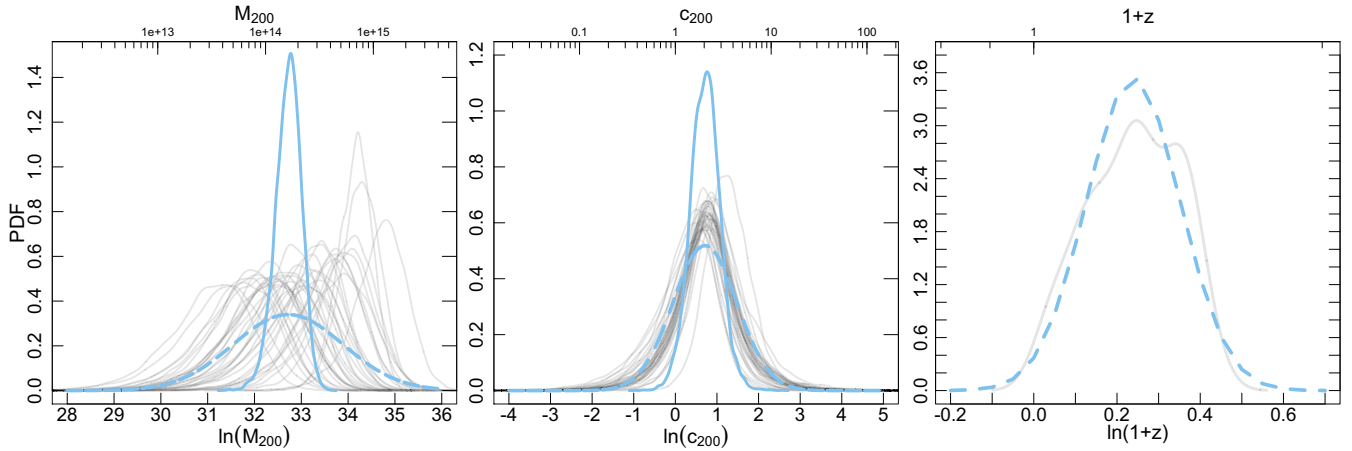


Figure 7. Left & centre: comparison of the posteriors for the population mean value (solid blue) and the posteriors for the individual clusters (solid grey) for mass and concentration, respectively. A 1σ deviation marginal centred on the mean of the population mean distribution (dashed blue). Right: the z distribution of the population plotted as a Gaussian centred on a mean and standard deviation obtained from the global mean vector and covariance matrix (dashed blue). It agrees well with the spectroscopic redshift distribution of the sample (shown as a Gaussian kernel density estimate in solid grey). From this, we can conclude that the data is able to constrain the individual cluster masses reasonably well, as the individual mass posteriors appear independent of the population mass posterior. On the contrary, the individual concentrations are completely dominated by the posterior of the population concentration, which implies that without the hierarchical model, individual cluster concentrations would not be possible.

The weight is expressed as a function of the error on the individual mass measurements ($\sigma_{M_1}, \sigma_{M_2}$)

$$w_i = \frac{1}{\sigma_{\ln(M_{2,i})}^2} = \left[\left(\frac{\sigma_{M_{1,i}}}{M_{1,i}} \right)^2 + \left(\frac{\sigma_{M_{2,i}}}{M_{2,i}} \right)^2 \right]^{-1}, \quad (24)$$

and the error we present on the mean is calculated from the standard deviation of 1000 bootstrap resamples. For an unbiased comparison, we look at only non-upper limit measurements. Lieu et al. (2016) find that the data quality limits their ability to constrain reliable concentration estimates so opt to fix concentration following the Duffy et al. (2008) c – M scaling relation. In comparison to their masses, we find a bias of $\langle M_{\text{hierarchical}}/M_{\text{Duffy}} \rangle = 0.72 \pm 0.02$. The Duffy relation is based on dark-matter-only N -body simulations and predicts concentrations that are much lower than those inferred from observations. Also, c – M relations in general tend to suffer from large scatter and therefore the choice of c – M relation will affect the mass obtained. For an unbiased comparison, we compare to their masses where the concentration is a fitted parameter and find that $\langle M_{\text{hierarchical}}/M_{\text{free}} \rangle = 0.86 \pm 0.05$. However, it is clear that it is not very informative to express the comparison in terms of a single number. The offset in mass is mass dependent, the hierarchical method measures significantly larger masses for the upper limit/low-mass systems, as they are pulled towards the population mean. The comparison of the marginalized posterior distribution functions of the masses derived here and those derived independently with concentration as a free parameter show reasonable agreement (Fig. A2 and Table A1). The obvious outliers are the low SNR objects that when treated individually show truncated posteriors at $1 \times 10^{13} M_{\odot}$. This truncation arises from the implantation of a harsh prior boundary that is well motivated from the X-ray temperatures. For the same clusters, our masses all lie above $10^{13} M_{\odot}$ but with very different values of mass, implying that even with a well-motivated prior, the effect on mass can be significant. For the high-mass clusters, the hierarchical method is systematically lower than those of Lieu et al. (2016); however, we do not expect this to be due to a bias since our tests on the cosmo-OWLS simulations (Section 4) have demon-

strated that an increase in the number of low-mass systems had no effect on the inferred values of the high-mass clusters. The mass estimates lie within 1σ of each other and difference may be a result of the small sample size used and the lack of high-mass systems in our sample. We expect our method to improve with sample size.

Smith et al. (2016) discussed alternative weight functions for comparison of two sets of cluster mass measurements via a weighted geometric mean calculation. They defined the weights for weak-lensing masses in terms of σ_M , in contrast to our choice of σ_M/M , arguing that for their sample the former definition was more closely related to data quality than the latter, which tended to up-weight more massive clusters. The Smith et al. (2016) sample spans a smaller redshift and mass range than the sample that we consider here. Therefore, issues relating to a mass-dependence of the weight function are much less clear cut for our study than for Smith et al. (2016). For simplicity in this proof of concept study, we therefore adopt the more conventional weight function given in equation (24).

5.4 Shrinkage

In the hierarchical model, mass is shrunk towards to the global population mean (Fig. 8). In comparison to the individually fitted masses measured in Lieu et al. (2016), equivalent to a population with mass variance $\sigma_{\ln M}^2 = \infty$, the hierarchical method is able to obtain better constraints on weakly constrained masses. Further, shrinkage estimates can be obtained by reducing the value of the relevant diagonal element of the global covariance matrix. As $\sigma_{\ln M}^2 \rightarrow 0$, the mass estimates shrinks towards the global mean, which is equivalent to the mass obtained by stacking all clusters.

Assuming all clusters have a single mass value, whilst allowing concentration to be free we obtain a stacked mass estimate of $\exp(\ln M_{200}) = 2.07 \pm 0.79 \times 10^{14} M_{\odot}$ with a global concentration value of $\exp(\mu_{\ln c}) = 1.78 \pm 1.72$.

We can perform the same analysis for concentration, whilst allowing mass to be free we obtain a stacked concentration estimate of $\exp(\ln c_{200}) = 2.21 \pm 1.44$ with a global concentration value of $\exp(\mu_{\ln M}) = 1.62 \pm 1.01 \times 10^{14} M_{\odot}$.

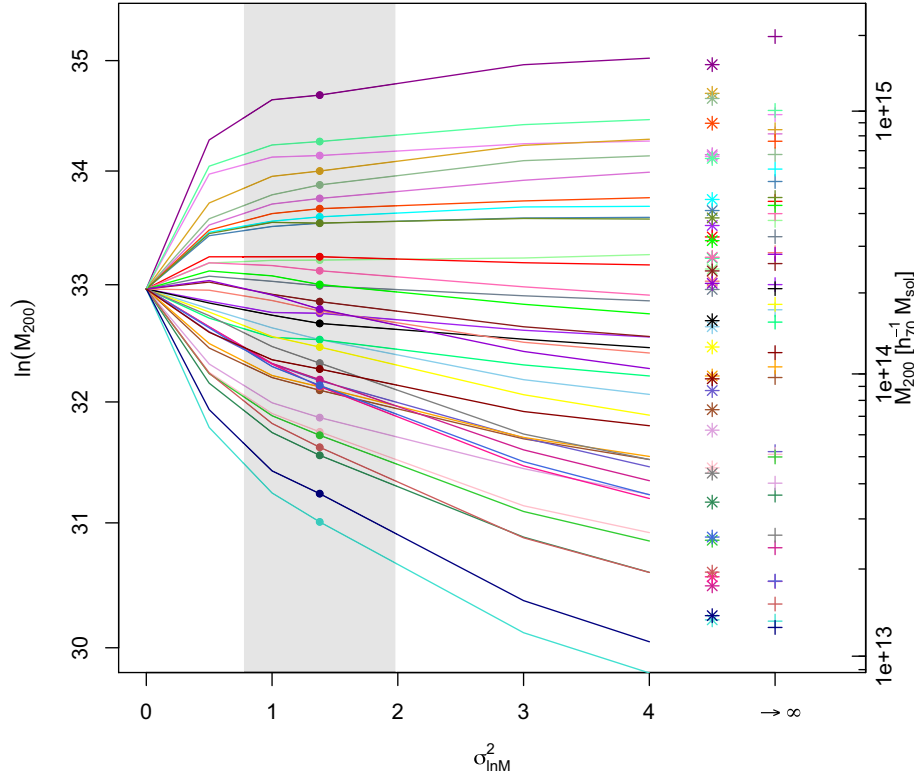


Figure 8. Individual galaxy cluster mass shrinkage estimates show the individual mass estimates shrink towards the population mean as $\sigma_{\ln M}^2$ decreases. Each cluster is represented as a different colour. The points show the fitted individual masses of clusters using the hierarchical method where $\sigma_{\ln M}^2$ is 1.38 and the shaded region is the 1σ error. The stars and crosses are the individual masses following a non-hierarchical method (Lieu et al. 2016) where concentration is a free parameter and where concentration is fixed to the Duffy et al. (2008) c–M relation, respectively.

A simultaneous fit for a single stacked mass and concentration (i.e. both $\sigma_{\ln M}^2 \rightarrow 0$ and $\sigma_{\ln c}^2 \rightarrow 0$) results in $1.91 \pm 0.70 \times 10^{14} M_{\odot}$ and 1.60 ± 1.16 , respectively. Hence, both parameters are in agreement within the errors either based on stacking only on either one of those parameters or both. The constraints on mass are stronger than concentration as expected due to the difficulty in measuring the latter.

The global means for the hierarchical fit were $\exp(\mu_{\ln M}) = 1.62 \pm 1.04 \times 10^{14} M_{\odot}$ and $\exp(\mu_{\ln c}) = 2.04 \pm 1.68$. Although within the errors these results are consistent with the shrinkage estimates, the mean mass is slightly smaller and the mean concentration is slightly larger. Simple stacking is a more severe constraint on M–c; blindly stacking clusters together can lead to incorrect mass estimates. In particular, our constraint on concentration is poor and therefore the mass estimates are not too sensitive to the concentration. More data are required to achieve a reliable estimate of the mean concentration of the population.

5.5 Mass–concentration relation

Using equation (22), we obtain mean values of $\alpha = 1.09^{+5.11}_{-2.85}$, $\beta = -0.02^{+0.11}_{-0.36}$, $\gamma = 0.59^{+1.54}_{-0.90}$ (Fig. 9). Note that the majority of the individual masses lie within 1σ since it is based not on the means of the masses but the posteriors. Here, the 1σ ellipse encompasses a third of the combined individual posteriors.

We find concentrations that are typically smaller than Duffy et al. (2008) and Dutton & Macciò (2014) though the slope of our relation is compatible. We note that with the quality of the data, we are unable to constrain concentration leading to large

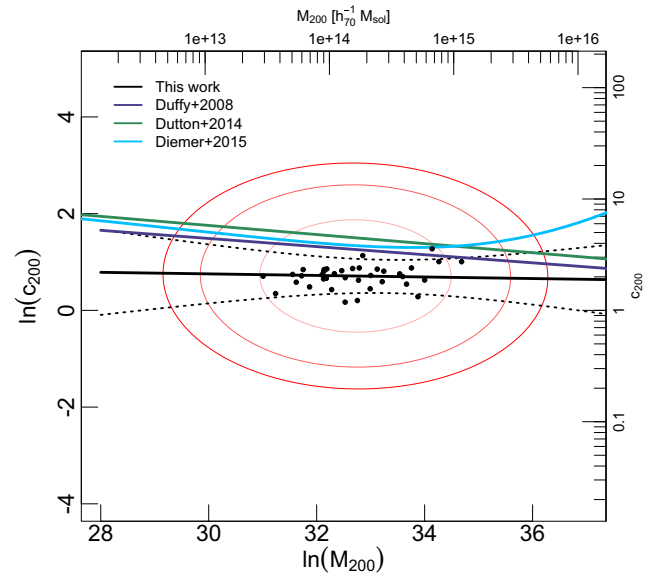


Figure 9. Concentration–mass relation. c values are computed for all M values in the range using the equation (22) for each pair of μ and Σ sampled. The mean and 1σ uncertainty is shown as the solid black and dotted lines, respectively. The fitted covariance and mean of population concentration and mass shown by red contours of 1σ , 2σ and 3σ confidence and therefore appear misaligned from the fit. For comparison, we plot the c–M relations from Duffy et al. (2008, solid purple line), Dutton & Macciò (2014, solid green line) and Diemer & Kravtsov (2015, solid blue line) are shown at our population mean redshift $z = 0.27$. The black points are the mean of the individual log parameters.

uncertainties on our regression parameters. Our data marginally agree with a weak anticorrelation between concentration and mass that is expected from simulations (e.g. Bullock et al. 2001). Low-mass groups formed in early times when the mean density of the Universe was larger, allowing concentrated cores to form. Massive clusters formed later on through the accretion of groups. In the literature, the concentration–mass relation is primarily estimated using numerical simulations where the concentration parameters are known exactly. Where a c – M relation has been measured from observations, studies have relied on high signal-to-noise clusters or stacking multiple clusters together to obtain a concentration estimate. We have already seen from the shrinkage estimates that stacking can cause overestimation of concentration. We note that the individual measurements of the $\ln \langle c_{200} \rangle$, $\ln \langle M_{200} \rangle$ are consistent with the higher values of concentration seen in the literature; however, our assumption that these parameters are lognormally distributed means that the correct values should be taken as $\langle \ln c_{200} \rangle$ and $\langle \ln M_{200} \rangle$ where the latter gives a result that is closer to the posterior peak of both $\text{Pr}(x)$ and $\text{Pr}(\ln(x))$, where x is c_{200} and M_{200} . Due to the large uncertainties, our results are not able to rule out higher concentrations (Fig. 7b). The c – M relations taken from the literature lie comfortably within the contours of the population mean and covariance.

6 DISCUSSION

6.1 Tests on priors

We test the influence of the priors on the toy data set. Recall that $\mu \sim \mathcal{N}(\mu_0, \sigma^2 = 1)$ where $\mu_0 = (32, 1, 0.3)$. For the 38 clusters and 0.1 shear data uncertainties, the estimated population mean was $\bar{\mu} = (33.18 \pm 0.19, 0.98 \pm 0.09, 0.26 \pm 0.01)$. We vary the values of μ_0 and find that the weakly informative prior does not affect the estimated population mean μ (see Table A1). For $\mu_0[1] = (30, 31, 32, 33, 34)$, the mean of the posterior samples is $\langle \mu[1] \rangle = 33.18 \pm 0.19$ and for $\mu_0[2] = (-1, 0, 1, 2)$ we obtain $\langle \mu[2] \rangle = 0.97 \pm 0.09$.

Testing the prior influence on the observational data, we again find that it does not affect the mean population mass and only weakly influences the estimation of the mean population concentration. Recall the measured population mean was $\bar{\mu} = (32.72 \pm 0.28, 0.71 \pm 0.36, 0.24 \pm 0.01)$. For the same varying values of $\mu_0[1]$ as above, we obtain $\langle \mu[1] \rangle = 32.71 \pm 0.13$ and for $\mu_0[2]$, we obtain $\langle \mu[2] \rangle = 0.63 \pm 0.16$. For the observational data at the extreme hyperprior variants, we get handful of divergences ($<10/4000$) even with a very high acceptance (0.999), as the prior extends further into the unstable regions of the parameter space. Given that the divergences remain sparse, we expect that the results to not be strongly affected.

6.2 Comparison to literature

The concentration–mass relation is still a topic of interest since the regression parameters throughout literature vary significantly and observationally, the uncertainties are large (Sereno et al. 2015). Observation based c – M relations tend to rely on stacking analyses or samples of high signal-to-noise systems. We neither stack our lensing signals nor limit our sample to a signal-to-noise threshold, since this may lead to a bias. Here, we discuss and compare our results to the literature.

Our data on average show lower concentration values compared to the Duffy et al. (2008) c – M relation that is known to be lower than many other simulation based c – M relations (e.g. Okabe et al. 2013;

Dutton & Macciò 2014). However, their relation assumes $WMA P5$ cosmology (we use $WMA P9$), and the inferred cosmology is known to have a non-negligible effect on concentration (Macciò, Dutton & van den Bosch 2008). Further, c – M relations based on numerical simulations tend to lower normalizations in comparison to observational samples. This could be due to selection effects or the physics included in the simulations.

Using cold dark matter simulations based on *Planck* cosmology, Dutton & Macciò (2014) find a c – M relation whose evolution is not described by others in the literature (see their fig. 11). Our data suggest a slight positive redshift evolution, however, with large uncertainties that are fully consistent with little or no evolution. Like many simulation based studies (Klypin et al. 2016), they find the Einasto density profile to be a better model for dark matter haloes in comparison to the NFW profile; however, the significance is more pronounced for massive systems. Gao et al. (2008) find that the Einasto profile improves the sensitivity of concentration estimates to the radial fitting range in particular for stacked clusters. To implement this model, however, would require the introduction of five extra hyperparameters and 38 parameters. More importantly, baryon physics is expected to play a more significant role in low-mass systems that are not included in these simulations. Feedback affects both the normalization and slope of the c – M relation by simultaneously decreasing the mass and increasing the scale radius, and massive neutrino free streaming can further lower the amplitude by reducing the mass (Mummery et al., in preparation).

Okabe et al. (2013) have shown that the NFW profile fits well to the observations of stacked weak-lensing data. Our method imposes a quasi-stacking so NFW may be appropriate. Compared to Duffy et al. (2008) and Dutton & Macciò (2014), our relation are 41 and 49 per cent systematically lower, respectively, although only with a significance of 1.04σ and 1.40σ lower. Our low concentration is consistent within the uncertainties with the literature (Sereno & Covone 2013); however, they consider higher redshift clusters ($0.8 < z < 1.5$). Such low concentrations are typical for haloes undergoing rapid mass accretion and tend to be less well fitted by the NFW profile.

Concentration is also correlated to the halo mass accretion history that in turn depends on the amplitude and shape of the initial density peak. In search for a universal halo concentration, Diemer & Kravtsov (2015) instead fit concentration to peak height v , their relation is also higher than ours; however, they find an upturn at high-mass (v) scales. This flattening and upturn of the c – M relation is also found in other studies (Bullock et al. 2001; Eke, Navarro & Steinmetz 2001; Prada et al. 2012) and is attributed to there being more unrelaxed haloes at higher mass. In our data, we too observe an upturn at higher mass scales; however, the low SNR of low-mass clusters will have larger concentration errors, so it is difficult to confirm this with the current small cluster sample.

Bahé et al. (2012) use mock weak-lensing observations based on numerical simulations to study the bias and scatter in M and c . They find that substructure and triaxiality can bias the concentration low (~ 12 per cent) with respect to the true halo concentration, with the effect of substructure being the dominant effect. It can also lead to large scatter whilst having a much smaller effect on M_{200} . We expect this effect to be small on our sample because substructure and triaxial haloes are more characteristic of massive clusters.

Recently, Du et al. (2015) use 220 redMaPPer (Rykoff et al. 2014) clusters with overlap with CFHTLenS to calibrate an observations c – M relation without stacking. They find a relation consistent with simulations but with large statistical uncertainties. Their clusters are slightly more massive than ours ($M_{200} \sim 10^{14} - 10^{15} M_{\odot}$), none

the less their results suggest that the c – M relation is highly sensitive to the assumed prior (their fig. 6.). They find that dilution by contaminating galaxies and miscentring can negatively bias the concentration values, we expect the latter to be more important in this work since we use spectroscopic redshifts and a conservative background selection, but our shear data are centred on the X-ray centroid. By including priors based on richness and centring offset in their model, their results change significantly. Consequently, we expect our c – M relation to change in the future with the inclusion of other cluster properties.

It is important to note that, like mass measurements, concentration values observed using different methods and definitions may vary. Concentrations derived from weak gravitational lensing, strong lensing and X-ray are yet to reach agreement (Comerford & Natarajan 2007).

Possible reasons that the low normalization of our c – M relation include the assumed cosmology, internal substructure, halo triaxiality or galaxy formation-related processes that expel baryons into the outer regions of the halo resulting in a shallower density profile (Sales et al. 2010; van Daalen et al. 2011). Also, as noted above, neutrinos can also lower the amplitude. Centre offset is degenerate with the normalization of the c – M relation and neglecting any miscentring could bias concentrations low (Viola et al. 2015). In our work, we centre shear profiles on the X-ray centroids that may not trace the centre of the dark matter halo as well as the BCG but this should be accounted for since the inner radius of 0.15 Mpc is excluded when fitting the NFW model.

Another important point regards the imposed multivariate Gaussian model and how well it fits the data. Fig. 7 shows that the posteriors of the individual cluster concentrations agree well with the Gaussian prior; however, the masses appear more constrained by the single Gaussian fit. A mixture model of three or more Gaussians may be a better prior for the mass; however, the additional flexibility introduced will also affect our ability to constrain concentration. For this work, there is no reason to believe that the clusters do not originate from the same underlying population since they are selected in the same way, however in the future if external samples are to be included then the addition of further Gaussians will be more important.

Sereno & Ettori (2015) and subsequent papers in the *CoMaLit* (COMparing MAsses in LITerature) series compare and apply methods to analyse cluster masses and scaling relations in a homogenous way whilst attempting to taking into account sample calibration issues that may lead to discrepancies in mass and scaling relations. Since in this paper, we only explore mass, concentration and redshift whilst the clusters are selected on X-ray count rate, we leave observation biases to a future paper that will include observables such as X-ray luminosity.

7 CONCLUSION

We have developed a hierarchical model to infer the population properties of galaxy groups and clusters, and present a method for its correct usage to estimate of unknown parameters of additional clusters that is superior to the ad hoc scaling relation. Nevertheless familiar scaling relations can also be extracted. We apply the method on toy data, hydrodynamical simulations and observational data. Using this model, we are able to obtain weak-lensing mass estimates of individual clusters down to $1 \times 10^{13} M_{\odot}$ without the need for harsh prior boundaries and assumptions about the concentration, even when the signal to noise is low. Below is a summary:

(i) We test the model on simulated toy data and find that the agreement with the true cluster mass and concentration is good and can further improve with increasing sample size and/or data uncertainties. Our tests on realistic weak-lensing measurements from hydrodynamical simulations similarly show promising agreement.

(ii) We then apply the method on a small sample of 38 low-mass groups and clusters from the Lieu et al. (2016). Using this hierarchical method, we are able to achieve better constraints on both mass and concentration without the compromise of upper limit measurements or the use of an external concentration–mass relation. This eliminates the bias introduced from calibrating with information derived from a sample that may not be representative of our systems. What’s more the concentrations used in Lieu et al. (2016) are derived from dark-matter-only simulations, the missing physics could invoke differences from observations. The tests on the simulations is promising for the extension of this work to the full XXL sample ($\gtrsim 600$ galaxy groups and clusters).

(iii) Eckert et al. (2016)’s study on the XXL-100-GC galaxy clusters in the XXL survey find a very low gas fraction that requires a hydrodynamical mass bias of $M_X/M_{wl} = 0.72^{+0.08}_{-0.07}$ to reconcile the difference. We measure masses on average 28 per cent smaller compared to the mass estimates from Lieu et al. (2016) that may be able to resolve the issue. We note, however, that at the low-mass end, we measure higher masses compared to Lieu et al. (2016), which would drive those gas fractions even lower.

(iv) The mean population cluster mass and concentration are measured to be $\mu_M = 1.62 \pm 1.04 \times 10^{14} M_{\odot}$ and $\mu_c = 2.04 \pm 1.68$. The shrinkage of individual masses towards the population mean suggests that hierarchical modelling has a larger effect on the low-mass systems where the SNR is low. Tests with shrinkage of parameters suggest that blindly stacking clusters for mass and concentration can bias the estimated value of the population mean. Parametrizing a single concentration whilst allowing mass to be free results in a concentration that is biased high compared to the population mean by 8 per cent. Stacking both concentration and mass to a single value on the contrary results in a positive mass bias of 18 per cent and negative concentration bias of 22 per cent. This is worrisome for studies that rely on single concentrations for mass estimation those that blindly stack large samples of clusters.

(v) We estimate the concentration–mass relation from the underlying population obtaining a result that within the uncertainties is consistent with the literature. We are able to recover the weak anti-correlation between concentration and mass; however, we find that the data suggest much lower concentrations than those previously measured in observations and simulations. We attribute this to the fact that observation based c – M relations rely on stacking analyses that we do not use and as stated previously stacked concentration estimates tend to be biased high. Our c – M relation suggests an evolutionary dependence, however, within the errors is not able to rule out no evolution.

Our method can be easily modified to incorporate more population parameters such as X-ray temperature, luminosity, gas mass, etc. The additional cluster information will help to improve the constraints on mass predictions. In the future, we hope to extend to cosmological inference by implementing a more accurate function to describe the population of clusters, namely convolving the true selection function with the cluster mass function. When the weak-lensing data for XXL-South clusters becomes available, we will be able to incorporate the additional systems to improve constraints on our model as well as other cluster samples in the literature (e.g. see Sereno et al. 2015). This work will be important for current

wide field surveys (such as DES, KiDS, etc.) where the data may be limited by the shallow survey depth, and for future big data surveys (e.g. *Euclid*, LSST, *e-ROSITA*) who will need more efficient ways to deal with processing the predicted quantities of data whilst extracting the maximum amount of information from them.

ACKNOWLEDGEMENTS

We thank Catherine Heymans, Gus Evrard and Jim Barrett for helpful discussions. We thank David van Dyk for a colloquium on shrinkage that inspired this work. We are grateful to the CFHTLenS team for making their shear catalogue publicly available. ML acknowledges a Postgraduate Studentship from the Science and Technology Facilities Council and an ESA Research Fellowship at the European Space Astronomy Centre (ESAC) in Madrid, Spain. WMF & GPS acknowledge support from STFC. MS acknowledges the financial contribution from contracts ASI-INAF I/009/10/0, PRIN-INAF 2012 ‘A unique data set to address the most compelling open questions about X-Ray Galaxy Clusters’, and PRIN-INAF 2014 1.05.01.94.02 ‘Glittering Kaleidoscopes in the sky: the multifaceted nature and role of galaxy clusters’.

REFERENCES

- Allen S. W., Evrard A. E., Mantz A. B., 2011, *ARA&A*, 49, 409
- Alsing J., Heavens A., Jaffe A. H., Kiessling A., Wandelt B., Hoffmann T., 2016, *MNRAS*, 455, 4452
- Alvarez I., Niemi J., Simpson M., 2014, preprint ([arXiv:1408.4050](https://arxiv.org/abs/1408.4050))
- Applegate D. E. et al., 2014, *MNRAS*, 439, 48
- Auger M. W., Budzynski J. M., Belokurov V., Koposov S. E., McCarthy I. G., 2013, *MNRAS*, 436, 503
- Bahé Y. M., McCarthy I. G., King L. J., 2012, *MNRAS*, 421, 1073
- Barnard J., McCulloch R., Meng X.-L., 2000, *Stat. Sin.*, 10, 1281
- Becker M. R., Kravtsov A. V., 2011, *ApJ*, 740, 25
- Betancourt M. J., Girolami M., 2015, *Hamiltonian Monte Carlo for Hierarchical Models* Vol. 79, Chapman and Hall/CRC, p. 30
- Betancourt M., Byrne S., Livingstone S., Girolami M., 2016, *Bernoulli*. Available at: <http://www.bernoulli-society.org/index.php/publications/bernoulli-journal/bernoulli-journal-papers>
- Bullock J. S., Kolatt T. S., Sigad Y., Somerville R. S., Kravtsov A. V., Klypin A. A., Primack J. R., Dekel A., 2001, *MNRAS*, 321, 559
- Comerford J. M., Natarajan P., 2007, *MNRAS*, 379, 190
- Corless V. L., King L. J., 2007, *MNRAS*, 380, 149
- De Boni C., Ettori S., Dolag K., Moscardini L., 2013, *MNRAS*, 428, 2921
- Diemer B., Kravtsov A. V., 2015, *ApJ*, 799, 108
- Du W., Fan Z., Shan H., Zhao G.-B., Covone G., Fu L., Kneib J.-P., 2015, *ApJ*, 814, 120
- Duffy A. R., Schaye J., Kay S. T., Dalla Vecchia C., 2008, *MNRAS*, 390, L64
- Duffy A. R., Schaye J., Kay S. T., Dalla Vecchia C., Battye R. A., Booth C. M., 2010, *MNRAS*, 405, 2161
- Dutton A. A., Macciò A. V., 2014, *MNRAS*, 441, 3359
- Eckert D. et al., 2016, *A&A*, 592, A12
- Eke V. R., Navarro J. F., Steinmetz M., 2001, *ApJ*, 554, 114
- Erben T. et al., 2013, *MNRAS*, 433, 2545
- Foëx G., Soucaill G., Pointecouteau E., Arnaud M., Limousin M., Pratt G. W., 2012, *A&A*, 546, A106
- Gao L., Navarro J. F., Cole S., Frenk C. S., White S. D. M., Springel V., Jenkins A., Neto A. F., 2008, *MNRAS*, 387, 536
- Gelman A., Rubin D. B., 1992, *Stat. Sci.*, 7, 457
- Gelman A., Hill J., 2006, *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge Univ. Press, Cambridge
- Giles P. A. et al., 2016, *A&A*, 592, A3
- Giodini S., Lovisari L., Pointecouteau E., Ettori S., Reiprich T. H., Hoekstra H., 2013, *Space Sci. Rev.*, 177, 247
- Heymans C. et al., 2012, *MNRAS*, 427, 146
- Hinshaw G. et al., 2013, *ApJS*, 208, 19
- Hoekstra H., Hartlap J., Hilbert S., van Uitert E., 2011, *MNRAS*, 412, 2095
- Hoekstra H., Bartelmann M., Dahle H., Israel H., Limousin M., Meneghetti M., 2013, *Space Sci. Rev.*, 177, 75
- Jing Y. P., 2000, *ApJ*, 535, 30
- Kaiser N., 1986, *MNRAS*, 222, 323
- Klypin A., Yepes G., Gottlöber S., Prada F., Heß S., 2016, *MNRAS*, 457, 4340
- Le Brun A. M. C., McCarthy I. G., Schaye J., Ponman T. J., 2014, *MNRAS*, 441, 1270
- Lewandowski D., Kurowicka D., Joe H., 2009, *J. Multivariate Anal.*, 100, 1989
- Lieu M. et al., 2016, *A&A*, 592, A4
- Macciò A. V., Dutton A. A., van den Bosch F. C., 2008, *MNRAS*, 391, 1940
- Navarro J. F., Frenk C. S., White S. D. M., 1997, *ApJ*, 490, 493
- Neal R. M., 2011, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall, London
- Oguri M., Bayliss M. B., Dahle H., Sharon K., Gladders M. D., Natarajan P., Hennawi J. F., Koester B. P., 2012, *MNRAS*, 420, 3213
- Okabe N., Smith G. P., Umetsu K., Takada M., Futamase T., 2013, *ApJ*, 769, L35
- Pierre M. et al., 2016, *A&A*, 592, A1
- Pacaud F. et al., 2016, *A&A*, 592, A2
- Piffaretti R., Valdarnini R., 2008, *A&A*, 491, 71
- Prada F., Klypin A. A., Cuesta A. J., Betancort-Rijo J. E., Primack J., 2012, *MNRAS*, 423, 3018
- Rykoff E. S. et al., 2014, *ApJ*, 785, 104
- Sales L. V., Navarro J. F., Schaye J., Dalla Vecchia C., Springel V., Booth C. M., 2010, *MNRAS*, 409, 1541
- Schneider M. D., Hogg D. W., Marshall P. J., Dawson W. A., Meyers J., Bard D. J., Lang D., 2015, *ApJ*, 807, 87
- Sereno M., Covone G., 2013, *MNRAS*, 434, 878
- Sereno M., Ettori S., 2015, *MNRAS*, 450, 3633
- Sereno M., Ettori S., 2016, preprint ([arXiv:1603.06581](https://arxiv.org/abs/1603.06581))
- Sereno M., Giocoli C., Ettori S., Moscardini L., 2015, *MNRAS*, 449, 2024
- Smith G. P. et al., 2016, *MNRAS*, 456, L74
- Stan Development Team, 2016a, *Stan Modeling Language Users Guide and Reference Manual*, Version 2.14.0. Available at: <http://mc-stan.org>
- Stan Development Team, 2016b, *RStan: the R interface to Stan*. R package version 2.14.1. Available at: <http://mc-stan.org>
- Sun M., 2012, *New J. Phys.*, 14, 045004
- Tinker J., Kravtsov A. V., Klypin A., Abazajian K., Warren M., Yepes G., Gottlöber S., Holz D. E., 2008, *ApJ*, 688, 709
- Umetsu K. et al., 2014, *ApJ*, 795, 163
- van Daalen M. P., Schaye J., Booth C. M., Dalla Vecchia C., 2011, *MNRAS*, 415, 3649
- Velliscig M., van Daalen M. P., Schaye J., McCarthy I. G., Cacciato M., Le Brun A. M. C., Dalla Vecchia C., 2014, *MNRAS*, 442, 2641
- Viola M. et al., 2015, *MNRAS*, 452, 3529
- Voit G. M., 2005, *Rev. Mod. Phys.*, 77, 207
- Zhao D. H., Jing Y. P., Mo H. J., Börner G., 2009, *ApJ*, 707, 354

APPENDIX A: NFW DENSITY PROFILE MODEL

The 3D NFW density profile is defined as

$$\rho_{\text{NFW}}(r) = \frac{\rho_s}{(r/r_s)(1+r/r_s)^2}, \quad (\text{A1})$$

where the central density is

$$\rho_s = \frac{200}{3} \frac{\rho_{\text{cr}} c^3}{\ln(1+c) - c/(1+c)}. \quad (\text{A2})$$

Here, $\rho_{\text{cr}} = (3H^2(z))/(8\pi G)$ is the critical density of the Universe at redshift z , where $H(z)$ is the Hubble parameter and G is Newton's

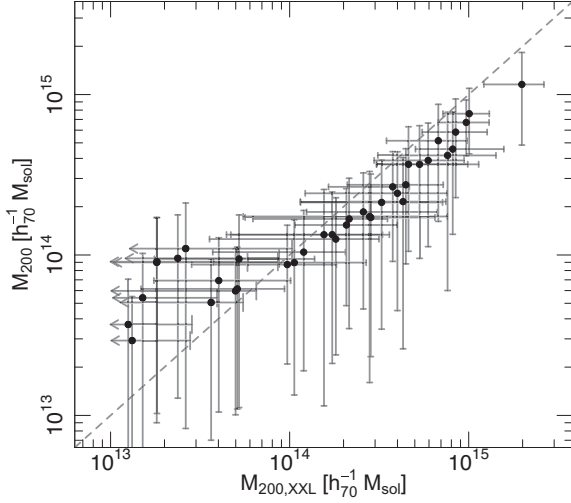


Figure A1. Comparison between our masses and those measured within Lieu et al. (2016), where they assume a fixed c – M relation from Duffy et al. (2008). Their upper limit measurements are shown in grey, where the estimate is confined by the lower prior boundary of $1 \times 10^{13} M_\odot$. The dashed line is equality. Our mass estimates show a systematic difference that is expected from the shrinking nature of the hierarchical model in that for high-mass clusters we predict lower masses and low-mass groups we predict higher mass values. The influence of the population distribution is more pronounced for the low-mass systems, where the uncertainties on the data are larger.

gravitational constant. We fit our data to the reduced gravitational shear

$$g_{\text{NFW}} = \frac{\gamma_{\text{NFW}}}{1 - \kappa}, \quad (\text{A3})$$

where the convergence can be expressed as the ratio of the surface mass density and the critical surface mass density $\kappa = \Sigma / \Sigma_{\text{cr}}$ and

$$\Sigma_{\text{cr}} = \frac{c^2}{4\pi G} \frac{D_S}{D_L D_{LS}}, \quad (\text{A4})$$

where c is the speed of light and D_S , D_L and D_{LS} are the angular diameter distances between the observer–source, observer–lens and lens–source, respectively. The shear is the difference between the mean surface mass density and the surface mass density

$$\gamma_{\text{NFW}} = \frac{\bar{\Sigma} - \Sigma}{\Sigma_{\text{cr}}}. \quad (\text{A5})$$

To obtain Σ and $\bar{\Sigma}$, we integrate the 3D density profile along the line of sight l ,

$$\begin{aligned} \Sigma(x) &= 2 \int_0^\infty \rho_{\text{NFW}} dl \\ &= \frac{2r_s \rho_s}{x^2 - 1} (1 - \xi), \end{aligned} \quad (\text{A6})$$

$$\begin{aligned} \bar{\Sigma}(< x) &= \frac{2}{x^2} \int_0^x x' \Sigma(x') dx' \\ &= \frac{4r_s \rho_s}{x^2} \left[\xi + \ln\left(\frac{x}{2}\right) \right], \end{aligned} \quad (\text{A7})$$

where $x = R/r_s$, R is the projected radial distance from the lens centre and ξ is a fourth-order power series expansion as x approaches 1,

$$\xi = \begin{cases} \frac{2}{\sqrt{1-x^2}} \operatorname{arctanh} \sqrt{\frac{1-x}{x+1}} & \text{if } x < 0.98, \\ \frac{2}{\sqrt{x^2-1}} \operatorname{arctan} \sqrt{\frac{x-1}{x+1}} & \text{if } x > 1.02, \\ 1 - \frac{2}{3}(x-1) + \frac{7}{15}(x-1)^2 - \frac{12}{35}(x-1)^3 \\ + \frac{166}{630}(x-1)^4 & \text{otherwise.} \end{cases} \quad (\text{A8})$$

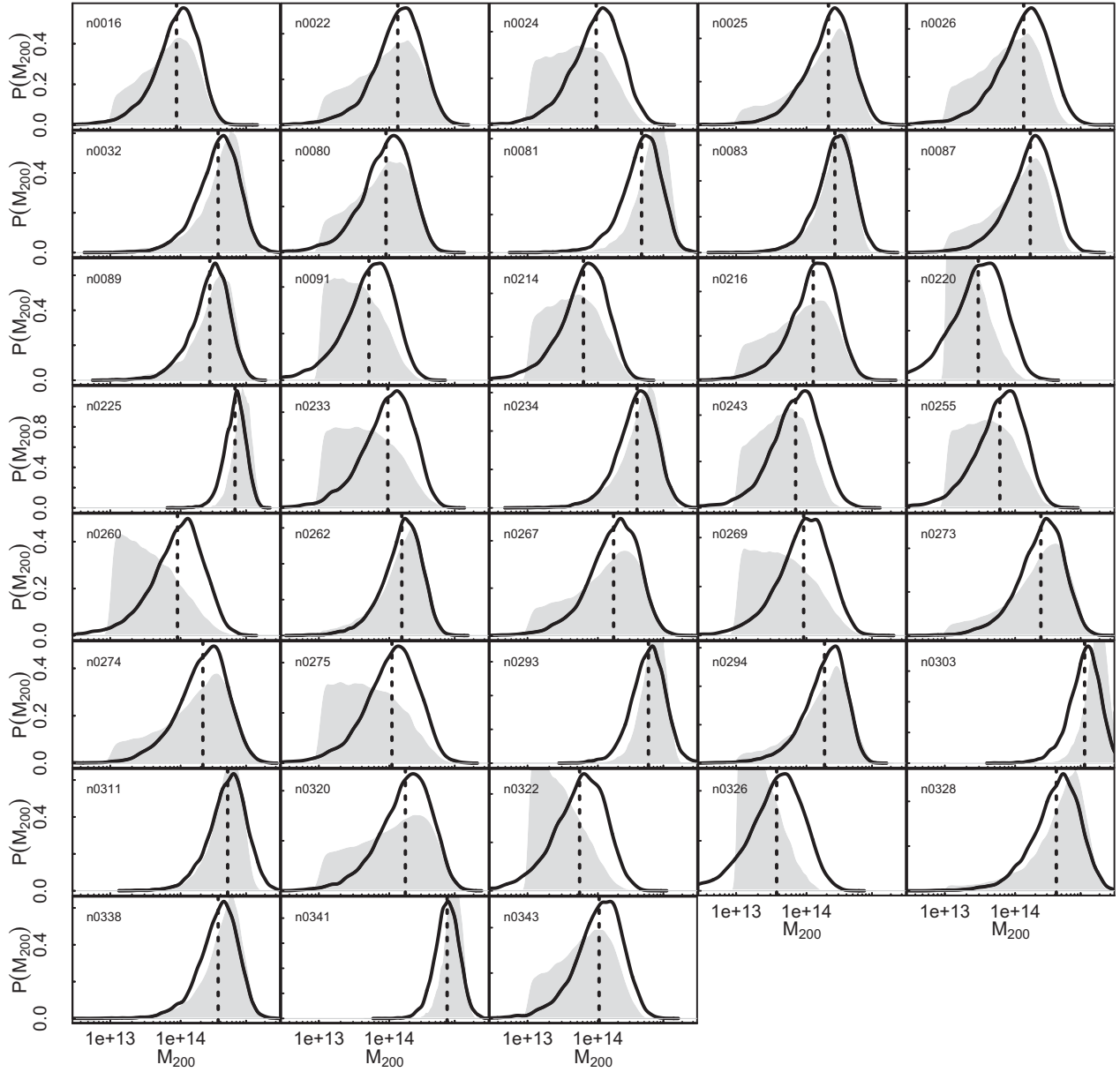


Figure A2. The posterior distribution functions of the individual mass measurements (solid black line) and the fit statistic taken as the posterior mean (dotted black line). The grey shaded regions show the posteriors of the individual masses from Lieu et al. (2016), assuming a free concentration parameter for comparison. The posteriors of both methods are in reasonable agreement. The truncated prior used in Lieu et al. (2016) can be seen at $10^{13} M_{\odot}$ for their clusters where only an upper limit on mass is measured, whereas our posteriors do not incur a sharp prior boundary yet are still able to constrain a posterior peak.

Table A1. The results of the tests on the assumed priors. The table shows fitted values of the population means for various prior central values.

Toy data				Observational data			
	Prior		Fit value		Prior		Fit value
$\mu_0[1] =$	30	$\overline{\mu_1} =$	33.11 ± 0.18	$\mu_0[1] =$	30	$\overline{\mu_1} =$	32.53 ± 0.33
	31		33.14 ± 0.18		31		32.64 ± 0.28
	32		33.18 ± 0.19		32		32.72 ± 0.28
	33		33.21 ± 0.18		33		32.80 ± 0.26
	34		33.24 ± 0.18		34		32.86 ± 0.25
$\mu_0[2] =$	-1	$\overline{\mu_2} =$	0.96 ± 0.09	$\mu_0[2] =$	-1	$\overline{\mu_1} =$	0.44 ± 0.42
	0		0.97 ± 0.09		0		0.57 ± 0.36
	1		0.98 ± 0.09		1		0.71 ± 0.36
	2		0.99 ± 0.09		2		0.79 ± 0.34