Cluster Strong Lensing with Hierarchical Inference

Formalism, functional tests, and public code release

P. Bergamini1, A. Agnello2, and G. B. Caminha3

1 INAF - OAS, Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, via Gobetti 93/3, I-40129 Bologna, Italy
2 DARK, Niels-Bohr Institute, Lyngbyvejen 2, 2100 Copenhagen
e-mail: adriano.agnello@nbi.ku.dk
3 Kapteyn Astronomical Institute, University of Groningen, Postbus 800, 9700 AV Groningen, The Netherlands

ABSTRACT

Context. Lensing by galaxy clusters is a versatile probe of cosmology and extragalactic astrophysics, but the accuracy of some of its predictions is limited by the simplified models adopted to reduce the (otherwise untractable) number of degrees of freedom. Aims. We aim at cluster lensing models where the parameters of all cluster-member galaxies are free to vary around some common scaling relations with non-zero scatter, and deviate significantly from them if and only if the data require it. Methods. We have devised a Bayesian hierarchical inference framework, which enables the determination of all lensing parameters and of the scaling-relation hyperparameters, including intrinsic scatter, from lensing constraints and (if given) stellar kinematic measurements. We achieve this through BayeLens, a purpose-built wrapper around common, parametric lensing codes for the lensing likelihood and samples the posterior on parameters and hyperparameters, which we release with this paper. Results. We have run functional tests of our code against simple mock cluster lensing datasets with realistic uncertainties. The parameters and hyperparameters are recovered within their 68% credibility ranges, and the positions of all the ‘observed’ multiple images are accurately reproduced by the BayeLens best-fit model, without overfitting. Conclusions. We have shown that an accurate description of cluster member galaxies is attainable, despite the large number of degrees of freedom, through fast and tractable inference. This extends beyond the state-of-the-art of current cluster lensing models. Key words. Gravitational lensing: strong – Methods: numerical – Galaxies: clusters: general – Cosmology: observations – dark matter

1. Introduction

Galaxy clusters (at $z \approx 0.3 - 0.9$) are ideal cosmic telescopes to study background galaxies out to $z \approx 7$, and they boost the lensing signal of their own galaxies. When galaxies reside in clusters, their lensing cross section is enhanced, allowing the study of galaxy populations at $z \approx 0.2 - 0.4$, over a wide mass range (down to $M_{\star} \approx 10^{9.5} M_\odot$, e.g. Keeton 2003; Grillo et al. 2014; Parry et al. 2016; Niemiec et al. 2017; Bergamini et al. 2019). If the total mass density distributions of lensing clusters are accurately reconstructed, cosmological parameters can also be inferred (e.g. Golse et al. 2002; Gilmore & Natarajan 2009; Caminha et al. 2016; Magaña et al. 2018; Grillo et al. 2018). Dedicated Hubble Space Telescope (HST) surveys, such as the Cluster Lensing And Supernova survey with Hubble (CLASH, Postman et al. 2012), the Hubble/Spitzer Frontier Fields mission (HFF, Lotz et al. 2017), and the Reionization Lensing Cluster Survey (RELICS, Coe et al. 2019), have enabled the identification and multi-band characterization of tens, sometimes hundreds, of multiple images of background sources, and hundreds of cluster member galaxies per cluster so as to constrain cluster lens models (Priewe et al. 2017). Subsequent multi-slit spectroscopic follow-up observations, such as the CLASH-VLT survey (Rosati et al. 2014) has gathered a wealth of multi-object spectroscopic data on cluster-member galaxies and background sources, sampling lightcones around 13 clusters ($\approx 20 \times 20$ arcmin$^2$). Due to their extent and magnifying power, galaxy clusters have also been used to study faint and high-redshift sources (e.g. Ishigaki et al. 2015; Livermore et al. 2015; Kawamata et al. 2016; Bouwens et al. 2017; Atek et al. 2018; Cerny et al. 2018; Hoag et al. 2018; Vanzella et al. 2017; Hashimoto et al. 2019; Vanzella et al. 2019). Since these studies may have implications on our understanding of star-formation at high redshift and reionisation, accurate magnification maps are required.

The advent of Multi-Unit Spectroscopic Explorer (MUSE, Bacon et al. 2012) has enabled integral-field spectroscopy of galaxy clusters. In particular, in recent years numerous MUSE observations have targeted the cores of several massive clusters (e.g. Richard et al. 2014; Jauzac et al. 2015; Limousin et al. 2016; Kawamata et al. 2016b; Lagattuta et al. 2017; Caminha et al. 2016, 2017a,b; Vanzella et al. 2019). Every MUSE pointing provides a data-cube with a field of view of 1 arcmin$^2$ and a spatial sampling of 0.2′′. The spectra cover $4750 < \lambda < 9350$ Å with a resolution of $\approx 2.6$ Å, almost constant along the whole wavelength range, and a spectral sampling of 1.25 Å/pix (the spectral resolution is robustly characterised, see Gérout et al. 2017). Integral-field spectroscopy of cluster cores enabled a se-
cure identification of tens of multiple images and their redshifts, and complete and pure sets of member galaxies.

Besides the identification of cluster members and multiple images, spectroscopy provides further constraints on mass models through kinematics. The homogeneous spectroscopic coverage of CLASH-VLT means that velocities of individual cluster members can be used to aid the cluster profile reconstruction, through dynamical modeling (Biviano et al. 2013; Sartoris et al. 2016). Similarly, integral-field follow-up with MUSE yields the internal stellar kinematics of cluster members, and allows to independently constrain their mass profiles.

Lensing clusters have undergone an extensive modeling effort by multiple independent teams. Up to now, the constraints have consisted in the positions of tens of multiple images per cluster (e.g. Richard et al. 2014; Grillo et al. 2013; Jauzac et al. 2015; Limousin et al. 2016; Kawamata et al. 2016; Caminha et al. 2016; Lagattuta et al. 2017; Caminha et al. 2017a,b; Bonamigo et al. 2018; Caminha et al. 2019). Mostly, lens models describe these clusters as superpositions of extended dark matter halos and more localized overdensities corresponding to cluster-member galaxies. So-called "non-parametric", grid-based models (e.g. Dye & Taylor 1998; Saha et al. 2001; Bradač et al. 2015; Lagattuta et al. 2017; Caminha et al. 2016). Mostly, lens models decompose the cluster potential into different (smooth) components that are expected to follow a physical hypothesis on how mass should be distributed in galaxies and clusters (see e.g. Jullo et al. 2007, for an overview). In particular, Natarajan & Kneib (1997) have argued that including cluster-member galaxies explicitly, as individual components of cluster mass models, is crucial. This has also emerged from the HFF model comparison project (Meneghetti et al. 2017), where parametric models routinely outperformed current versions of grid-based models.

The bottle-neck in cluster lens models consists in: ≈ 0.3″-0.6″ errors in image-position reconstruction (e.g. Grillo et al. 2016), significantly higher than current positional uncertainties from HST data; and the appreciable discrepancies in magnification maps produced by different models on the same clusters (Meneghetti et al. 2017). One cause of such discrepancies may be the rigidity of the models used to reduce the degrees of freedom associated to cluster members, since galaxies are currently modeled as belonging to zero-scatter scaling relations, typically with hyperparameters that are imposed externally (e.g. Limousin et al. 2007). The freedom and accuracy in lensing models have gained importance in the era of high-accuracy cosmography (see e.g. Treu et al. 2016), and on the reliability of magnification maps for the study of high-redshift galaxy populations (Bouwens et al. 2017). Up to now, departures of individual galaxies from prescribed scaling relations were determined heuristically, on a galaxy-by-galaxy basis (Jauzac et al. 2018). Ideally, all galaxies should be let free to vary around some finite-scatter scaling relations, whose parameters (including scatter) are to be determined directly from the data on each given cluster.

The solution is a Hierarchical Bayesian inference formalism, where each galaxy has its own associated parameters, and the parameters of all galaxies are posited to be drawn from common relations with hyperparameters to be determined through lens modeling and (if given) auxiliary kinematic information. In this paper, we illustrate this formalism and its application to lensing models of three simple mock clusters of increasing complexity with HFF/CLASH-VLT data quality, accounting for all observational constraints in a self-consistent manner. This ensures that cluster lens models are as flexible as possible, given the data, and that higher accuracy is reached in predicted image positions. Since only one of the contributions to the inference is from lensing, and in order to ensure a fair comparison with state-of-the-art technology, we build our inference as a modular wrapper of common lens modeling codes. This also ensures that further constraints, e.g. from time delays, flux-ratios and shapes of background sources, can be easily included in the inference.

This paper is structured as follows. Our hierarchical inference is detailed in Section 2. Section 3 covers some technicalities inherent to our code implementation of the models. In Section 4, we perform functional tests on three simple but realistic mock cluster. Results are discussed in Section 5, and we conclude in Section 6.

While a cosmological model is adopted to generate mock clusters and fit them, the main results of this work are general and separate from the choice of cosmology. In fact, if needed, cosmological parameters may be sampled as additional hyperparameters in our inference scheme.

2. A Hierarchical Cluster Lensing Model

Here we generalize the formalism of parametric cluster-lensing models to allow for hierarchical inference, in particular scaling relations with non-zero scatter. Given the abundance of symbols, in the following (Sections 2.2 and 2.3) all model parameters and hyperparameters are marked by a "hat" symbol.

2.1. Cluster Lensing Components

In a parametric cluster lensing model, the total mass distribution is typically divided into a few component classes (Jullo et al. 2007). Here, we indicate each class through its contribution, \( \rho^{\text{class}} \), to the lensing potential. A first, smooth, cluster-scale component (\( \rho^{\text{halo}} \)) accounts for both the dark matter (DM) content of the cluster and the baryonic intra-cluster gas and light contributions. A second, ‘clumpy’ component describes the mass in cluster member halos (\( \rho^{\text{halo}} \)), in DM and baryons. A third component accounts for the presence of massive structure in the outer cluster regions, and possibly additional massive halos along the line-of-sight (\( \rho^{\text{halo}+\text{halo}} \)). Within this model class, the total cluster potential is then:

\[
\phi_{\text{tot}} = \sum_{i=1}^{N_h} \rho^{\text{halo}} + \sum_{j=1}^{N_{\text{halo}}} \rho^{\text{halo}} + \sum_{k=1}^{N_{\text{los}}} \phi^{\text{shear+los}},
\]

with \( N_h, N_{\text{halo}}, N_{\text{los}} \), the number of cluster scale halos, cluster members, and shear plus line-of-sight contributions respectively. The number of constraints, given by the observed multiple image positions, is usually not sufficient to fit for the parameters of mass profiles used to describe each individual galaxy in a given cluster lens model. For this reason, cluster members are usually parameterized as circular dual pseudo-isothermal mass distributions (circular dPIEs, Limousin et al. 2005, Eladsdöttir et al. 2007), with negligible core radii, whose parameters are related to galaxy luminosities according to fixed scaling relations.

The circular dPIE is defined by the 3D mass density (Limousin et al. 2005):

\[
\rho(r) = \rho_0 \left(1 + \frac{r}{r_{\text{core}}} \right)^{-2} \left(1 + \frac{r^2}{r_{\text{cut}}} \right)^{-1},
\]
\[ \rho_0 = \frac{\sigma_0^2}{2\pi G} \left( \frac{r_{\text{cut}}}{r_{\text{core}}} + r_{\text{core}} \right). \] 

(3)

For a vanishing core radius, \( r_{\text{cut}} \) encircles about half of the total 3D dPIE mass and 60% of the projected mass. While, in the limit \( r_{\text{cut}} \to \infty \) and \( r_{\text{core}} > 0 \) the dPIE coincides with the pseudo isothermal mass distribution defined in Kassiola & Kovner (1993) (PIEMD).

In lensing models of galaxy clusters, the following scaling relations are commonly adopted for the central velocity dispersions, core radii and truncation radii of the cluster members:

\[ \sigma_{0,i}^{\text{gal}} = \sigma_{0,i}^{\text{ref}} \left( \frac{L_i}{L_0} \right)^{\beta_{\text{core}}} = \sigma_{0,i}^{\text{ref}} \left( \frac{L_i}{L_0} \right)^{\beta_{\text{cut}}} = \sigma_{0,i}^{\text{ref}} \left( \frac{L_i}{L_0} \right)^{\beta_{\text{cut}}}, \]

(4)

without any intrinsic scatter. Under the hypothesis of a power-law scaling \( M_{\text{total}}/L_0 \propto L_i^\gamma \) of the total mass of a cluster member, and since the total dPIE mass is \( M_{\text{gal}} = (\sigma_0^2 r_{\text{cut}})/G \), the exponents are therefore related through:

\[ \beta_{\text{cut}} = \gamma - 2\alpha + 1. \]

(5)

### 2.2. Measured Kinematics

Spectroscopic information, when incorporated inside the lensing models, produces robust and accurate reconstructions of the projected cluster masses. However, strong lensing models still suffer from internal degeneracies between the parameters of their mass components (Eq. 1). An obvious degeneracy exists, for example, between the two normalizations, \( \sigma_0^{\text{ref}} \) and \( r_{\text{cut}}^{\text{ref}} \) in Eq. (4). This degeneracy is easily understandable if we consider that the total mass of a dPIE, with negligible \( r_{\text{core}} \), inside an aperture of radius \( R \) is (Jullo et al. 2007):

\[ M(R) = \frac{\pi \sigma_0^2}{G} \left( R + r_{\text{cut}} - \sqrt{r_{\text{cut}}^2 + R^2} \right), \]

(6)

and that multiple images constrain only the total mass within their (projected) distance from the center of a galaxy. Other degeneracies exist also between the cluster scale DM halo parameters and the clumpy sub-halo components, since the overall mass may be redistributed differently between the main halo and the sub-halos, except where the multiple-image positions are strongly constraining.

Recently, velocity dispersions of cluster members (from fits to spectra with \( S/N > 10 \)) were used to break or reduce these internal degeneracies (Verdugo et al. 2007; Monna et al. 2015, 2017 and more extensively by Bergamini et al. 2019). Monna et al. (2015) and Monna et al. (2017) fixed the velocity dispersions of several cluster galaxies to their measured values, thereby breaking their model degeneracies by imposing a strong hypothesis on their mass content. In more flexible approach, Bergamini et al. (2019) used high-quality MUSE spectra of cluster-member galaxies to build priors on the parameters of the two scaling relations in Eq. (4).

In this work, we combine both of the above to fully characterize the sub-halo components of clusters. Our purpose-built BayesLens wrapper uses the available measured velocity dispersions of the cluster galaxies, \( \sigma_{m}^{\text{gal}} \pm \delta\sigma_{m}^{\text{gal}} \), to infer the hyperparameters \((\hat{\Delta}^{\text{ref}}, \hat{\alpha})\) of the \( \sigma \)-mag scaling relation. A third hyperparameter, \( \Delta\sigma^{\text{ref}} \), quantifies the scatter of the measured galaxies around the \( \sigma \)-mag scaling relation. Gaussian priors, centered on the measured \( \sigma_{m}^{\text{gal}} \) and with standard deviations equal to five times the measured errors \( \delta\sigma_{m}^{\text{gal}} \), are adopted for the measured kinematics of galaxies inside the lens model. For galaxies without a measured velocity dispersion, we assumed Gaussian priors centered on the inferred \( \sigma \)-mag scaling relation and with a standard deviation equal to \( \Delta\sigma^{\text{ref}} \). We adopt a uniform prior for the reference cut radius of the \( r_{\text{cut}} \)-mag scaling relation, and \( \beta_{\text{cut}} \) is obtained from Eq. (5). Unless otherwise stated, all model hyperparameters are left free to vary, so as to fully explore the model posterior probability as described below. All “hat” symbols are introduced more formally in the following subsection (2.3).

### 2.3. Fitting it all together: the Posterior

In our models, we use the measured velocity dispersions of \( N_{\text{gal}}^{\text{mag}} \) cluster galaxies, \( \sigma_{m}^{\text{gal}} \pm \delta\sigma_{m}^{\text{gal}} \), together with the positions \( \mathbf{x}_{\text{gal}} \) of \( N_{\text{im}}^{\text{mag}} \) multiple images, from \( N_{\text{im}}^{\text{mag}} \) different sources with positions \( \mathbf{x}_{\text{obs}} \), as observational constraints to the lens model free parameters. Hereafter, these free parameters will be marked with a hat symbol. In order to explore the lens models, we sample the total posterior probability function \( p_{\text{tot}} \), expressed as the product of: the odds \( p_{\text{odds}} \) of drawing each galaxy (independently) from a given scaling relation (with intrinsic scatter); the likelihood \( p_{\text{odds}} \) of the scaling relation hyperparameters, given the measured velocity dispersions and luminosities; the likelihood \( p_{\text{odds}} \) of reproducing the measured image positions, given the cluster model (including the parameters of each cluster-member galaxy); a prior \( p_{h} \) on the parameters of the main cluster-scale halo(s). We also include a term \( p_{\text{mag}} \) that links the dPIE lensing velocity dispersions of the galaxies with measured kinematics to their stellar velocity dispersions, with large uncertainties. Although not formally needed, this term is used only as a loose regularization to ensure convergence.

In synthesis,

\[ \begin{align*}
& p_{\text{tot}} \left( \hat{\sigma}^{\text{ref}}, \hat{\alpha}, \hat{\Delta}^{\text{ref}}, \hat{\sigma}^{\text{mag}}_{m} | \mathbf{x}_{\text{gal}}^{\text{mag}} \right) \propto
& \quad \times p_{\text{odds}} \left( \hat{\sigma}^{\text{ref}}, \hat{\alpha}, \hat{\Delta}^{\text{ref}}, \hat{\sigma}^{\text{mag}}_{m} | \mathbf{x}_{\text{gal}}^{\text{mag}} \right)
& \quad \times p_{\text{im}} \left( \hat{\sigma}^{\text{mag}}_{m}, \mathbf{x}_{\text{im}} \right) \times p_{h} \left( \hat{\sigma}^{\text{mag}}_{m}, \mathbf{x}_{\text{obs}} \right).
\end{align*} \]

(7)

Each of the five factors at the right-hand-side is discussed below. In the following, the quantities referring to cluster galaxies with measured velocity dispersions are marked with the subscript "mag".

Some of the factors (e.g. the one on hyperparameters) can be interpreted as a posterior on some parameters given some observations, which (by the ‘Bayes chain rule’) can further be used as a prior for the full hierarchical posterior. To avoid confusion between those posteriors and the final posterior, we alternatively refer to them as ‘term’ or ‘odds’ in the following.
2.3.1. Odds on the scaling relation hyperparameters, \( p_{sr} \)

This factor is responsible for the \( \sigma_{\text{mag}} \) and \( r_{\text{cut}} \)-mag scaling relation hyperparameters, given the set of \( N_m^\text{gal} \) measured cluster galaxies. For the galaxy velocity dispersions we consider here a scaling relation of the same form of Eq. (4) parameterized by the reference measured velocity \( \hat{\sigma}_{\text{ref}} \) and slope \( \hat{\alpha} \) plus an intrinsic scatter \( \hat{\Delta}_{\text{ref}} \) in measured velocity dispersions. Regarding the \( r_{\text{cut}} \)-mag scaling relation, in the current version of our models we optimize only the reference value \( r_{\text{cut}}^{\text{ref}} \) while the slope \( \beta_{\text{cut}} \) is determined using Eq. (5) from the inferred \( \hat{\sigma}\text{ref} \) \( \hat{\alpha} \) and assuming a fixed mass-to-light scaling for the cluster galaxies. No scatter around this relation considered.

The term \( p_{sr} \) can be expressed as:

\[
p_{sr} \left( \hat{\sigma}_{\text{ref}}, \hat{\alpha}, \hat{\Delta}_{\text{ref}}, r_{\text{cut}}^{\text{ref}} \mid \text{mag}^{\text{gal}}, \sigma_{m}^{\text{gal}}, \delta\sigma_{m}^{\text{gal}} \right) \propto p_{0,\text{sr}} \left( \sigma_{m}^{\text{gal}} \mid \text{mag}^{\text{gal}}, \delta\sigma_{m}^{\text{gal}}, \hat{\sigma}_{\text{ref}}, \hat{\alpha}, \hat{\Delta}_{\text{ref}}, r_{\text{cut}}^{\text{ref}} \right) \times \sigma_{sr} \left( \hat{\sigma}_{\text{ref}}, \hat{\alpha}, \hat{\Delta}_{\text{ref}}, r_{\text{cut}}^{\text{ref}} \right), \tag{8}
\]

where the uninformative prior \( \sigma_{sr} \) is defined by:

\[
\ln \left( \sigma_{sr} \left( \hat{\sigma}_{\text{ref}}, \hat{\alpha}, \hat{\Delta}_{\text{ref}} \right) \right) = \begin{cases} -\ln(\hat{\Delta}_{\text{ref}}), & \text{if } \hat{\sigma}_{\text{ref}} \leq \hat{\sigma}_{\text{ref}}^{\text{min}}, \\
\sum_{\text{i}=1}^{N_{\text{gal}}} \left[ \frac{\left( \sigma_{m,i}^{\text{gal}} - \sigma_{m,i}^{\text{ref}} \right)^2}{\left( \hat{\sigma}_{\text{ref}} \right)^2 + \left( \hat{\Delta}_{\text{ref}} \right)^2} \right] + 2\pi \left[ \left( \delta\sigma_{m,i}^{\text{ref}} \right)^2 \right], & \text{otherwise.} \end{cases}
\tag{9}
\]

and limits the \( \sigma_{\text{mag}} \) plus scatter scaling relation hyperparameters and the reference truncation radius, \( r_{\text{cut}}^{\text{ref}} \), to lie within suitably chosen boundaries \( \sigma_{\text{ref}}^{\text{min(max)}}, \sigma_{\text{ref}}^{\text{min(max)}}, \Delta\sigma_{\text{ref}}^{\text{min(max)}} \) and \( r_{\text{cut}}^{\text{min(max)}} \)

The log-likelihood term

\[
\ln \left( p_{0,\text{sr}} \left( \sigma_{m}^{\text{gal}} \mid \text{mag}^{\text{gal}}, \delta\sigma_{m}^{\text{gal}}, \hat{\sigma}_{\text{ref}}, \hat{\alpha}, \hat{\Delta}_{\text{ref}} \right) \right) =
-\frac{1}{2} \sum_{i=1}^{N_{\text{gal}}} \left[ \frac{\left( \sigma_{m,i}^{\text{gal}} - \sigma_{m,i}^{\text{ref}} \right)^2}{\left( \hat{\sigma}_{\text{ref}} \right)^2 + \left( \hat{\Delta}_{\text{ref}} \right)^2} \right] + 2\pi \left[ \left( \delta\sigma_{m,i}^{\text{ref}} \right)^2 \right], \tag{10}
\]

quantifies the goodness-of-fit of scaling relation hyperparameters to the measured velocity dispersions. In Eq. (10) we define \( \hat{\sigma}_{\text{ref}}^{\text{ref}} \) as:

\[
\dot{\sigma}_{\text{ref}}^{\text{ref}} = \sigma_{\text{ref}}^{\text{ref}} \cdot 10^{0.4 \left( \text{mag}^{\text{ref}} \right)}, \tag{11}
\]

where \( \text{mag}^{\text{ref}} \) corresponds to the reference luminosity \( L_0 \) in Eq. (4).

2.3.2. Term on measured galaxies, \( p_{\text{mg}} \)

This term applies only to the \( N_{\text{mgal}}^{\text{gal}} \) galaxies with measured velocity dispersions. It allows for residual uncertainties (from mass-anisotropy degeneracy, asphericity) in converting measured velocity dispersions into dPIE \( \sigma_{m}^{\text{gal}} \) which in turn propagate on the scaling relations that all galaxies are drawn from. This is attained by linking the kinematic sigma to the lensing sigma with some tolerance. If this were not done, one may obtain biased results if the kinematic sigma-mag relation has some intrinsic scatter but the lensing sigma-mag does not. This term is also needed in order to account for systematic uncertainties in the "measured" dispersions, which in the high S/N regime can dominate over the statistical uncertainties. Erring on the conservative side, we choose the conversion tolerance to five times the uncertainties in the measured velocity dispersion.

Therefore, \( p_{\text{mg}} \) alone consists in Gaussian priors centering the cluster-member aperture-average velocity dispersions \( \sigma_{m}^{\text{gal}} \) on their kinematic values \( \sigma_{m}^{\text{gal}} \). We choose the standard deviation of Gaussian priors equal to five times the error on the kinematic measurements \( \delta\sigma_{m}^{\text{gal}} \). This term is given by:

\[
\ln \left( p_{\text{mg}} \left( \sigma_{m}^{\text{gal}} \mid \sigma_{m}^{\text{ref}}, 5 \delta\sigma_{m}^{\text{gal}} \right) \right) =
-\frac{1}{2} \sum_{i=1}^{N_{\text{mgal}}} \left[ \frac{\left( \sigma_{m,i}^{\text{ref}} - \sigma_{m,i}^{\text{gal}} \right)^2}{\left( 5\delta\sigma_{m,i}^{\text{gal}} \right)^2} \right] + 2\pi \left[ \left( 5\delta\sigma_{m,i}^{\text{gal}} \right)^2 \right]. \tag{12}
\]

In other words, this term attributes to galaxies their measured velocity dispersions, with very wide tolerance, unless a deviation from these values produces a significant improvement of the lensing model. We emphasize that this term is, strictly speaking, not necessary and formally it may prevent the finite-scatter models to reduce to zero-scatter models (if this is needed by the data). However, given the typical \( \delta\sigma_{m}^{\text{gal}} \gtrsim 15 \text{ km/s} \) uncertainties on measured kinematics, the \( 5 \delta\sigma_{m}^{\text{gal}} \) term ensures that this factor only acts as a loose regularisation, preventing pathological solutions and aiding the convergence of the models.

2.3.3. Prior on unmeasured galaxies, \( p_{\text{ug}} \)

This term is a collection of Gaussian priors on velocity dispersion values for the \( N_{\text{ugal}} - N_{\text{mgal}}^{\text{gal}} \) galaxies without kinematics measurements. Its form is such that the final posterior prefers lens models in which the unmeasured galaxies lie on the \( \sigma_{\text{mag}} \) scaling relation inferred by \( p_{sr} \), unless otherwise required by the lensing data:

\[
\ln \left( p_{\text{ug}} \left( \sigma_{m}^{\text{ugal}} \mid \sigma_{m}^{\text{ref}}, \sigma_{m}^{\text{ref}}, \sigma_{m}^{\text{ref}} \right) \right) =
-\frac{1}{2} \sum_{i=1}^{N_{\text{ugal}}} \left[ \frac{\left( \sigma_{m,i}^{\text{ref}} - \sigma_{m,i}^{\text{ugal}} \right)^2}{\left( \sigma_{m}^{\text{ref}} \right)^2} \right] + 2\pi \left[ \left( \sigma_{m}^{\text{ref}} \right)^2 \right]. \tag{13}
\]

The \( \sigma^{\text{ref}} \) are computed through Eq. (11), but now considering galaxies without measured velocity dispersions.
2.3.4. Prior on halo parameters, \( p_h \)

This term consists in flat priors on the smooth cluster-scale halo parameters collectively indicated as \( \phi^h \):

\[
\ln \left\{ p_h(\phi^h) \right\} = \begin{cases} 
0, & \text{if } \phi_{\text{min}}^h < \phi^h < \phi_{\text{max}}^h \\
-\infty, & \text{otherwise}
\end{cases} \quad (14)
\]

If (as follows) these halos are parameterized as PIEMD profiles, \( p_h \) is a prior on the sky coordinates \( x^h \) and \( y^h \), on the ellipticity \( \epsilon^h \), position angle \( \theta^h \), core radius \( r_{\text{core}}^h \) and central velocity dispersion \( \sigma_{\text{v}}^h \).

2.3.5. Multiple-image likelihood, \( p_{\text{im}} \)

This final term is a likelihood function, which quantifies the agreement between observed and lens model predicted multiple-image positions.

Given \( N_{\text{im}} \) sources with \( N_{\text{im}}^i \) multiple images associated to the same \( i^{\text{th}} \) source (usually called a family), we follow Jullo et al. (2007) and express the likelihood as:

\[
p_{\text{im}} \left( \hat{x}_{\text{im}}, \hat{y}_{\text{im}} \mid \hat{\theta}^i_{\text{ref}}, \hat{\alpha}^i_{\text{ref}}, \Delta \sigma^i_{\text{ref}}, \hat{r}_{\text{cut}}^i, \hat{\sigma}^i_{\text{gal}}, \hat{\sigma}^i_{\text{v}}, \hat{x}_{\text{pre}}^i \right) = \\
\frac{1}{N_{\text{im}}^i} \prod_{i=1}^{N_{\text{im}}^i} \frac{e^{-\chi_i^2/2}}{\sqrt{2\pi} \Delta x_i \Delta y_i}, \quad (15)
\]

where the \( \chi_i^2 \) associated to the \( i^{\text{th}} \)-family is:

\[
\chi_i^2 = \sum_{j=1}^{N_{\text{im}}^i} \left[ \frac{(x_{\text{obs}}^i - x_{\text{pre}}^i)^2 + (y_{\text{obs}}^i - y_{\text{pre}}^i)^2}{\Delta x_j^i \Delta y_j^i} \right], \quad (16)
\]

with \( x_{\text{obs}}^i, y_{\text{obs}}^i \) the observed positions of the multiple images on the lens plane, \( \Delta x_j^i, \Delta y_j^i \) are the isotropic uncertainties on these positions and \( x_{\text{pre}}^i, y_{\text{pre}}^i \) are the model predicted positions, given the adopted cosmology and the inferred set of model parameters: \( \hat{\theta}^i_{\text{ref}}, \hat{\alpha}^i_{\text{ref}}, \Delta \sigma^i_{\text{ref}}, \hat{r}_{\text{cut}}^i, \hat{\sigma}^i_{\text{gal}}, \hat{\sigma}^i_{\text{v}}, \hat{x}_{\text{pre}}^i \).

3. Technicalities

To sample the complete posterior in Eq. (7), we use the Affine-Invariant sampling as originally introduced by Goodman & Weare (2010), which is especially suited to our highly-dimensional (and possibly degenerate) parameter space. In particular, to enable full portability and reproducibility, we use the latest python release\(^1\) of emcee (Foreman-Mackey et al. 2013).

The first four terms in equation Eq. (7) are directly implemented in our code, while to compute the multiple-images likelihood \( p_{\text{im}} \) we exploit the publicly available software LensTool (Kneib et al. 1996, Jullo et al. 2007, Julio & Kneib 2009).

The synergy between our code and LensTool requires some technicalities, as described in the following subsections. In any case, our code is fully modular, so that LensTool can be replaced by any other parametric lensing software by changing only a few python lines.

\(^1\) https://github.com/dfm/emcee

3.1. Calls to LensTool: Input and output files

For each parameters combination, corresponding to a given walker position inside the parameters space, BayesLens silently calls LensTool to compute the \( p_{\text{im}} \) term of the total posterior. Every LensTool call needs a different input file generated in our code by a specific python-function. Another function reads the resulting likelihood, computed by LensTool, from an output file. Since all these output files are saved on the disk using the same name, we create folders with unique (random) names to differentiate each LensTool call. These folders are deleted at the end of every \( p_{\text{im}} \) computation.

To sample the posterior \( p_{\text{im}} \), millions of walker positions are required. Thus millions of input and output LensTool files are quickly created and deleted on the disk during this process. To avoid the disk wear and the bottle neck represented by the process of writing/reading files, part of our computer RAM is reserved to create a RAM-disk where the input and output files are temporarily saved and deleted.

3.2. From measured \( \sigma_{\text{gal}} \) to LensTool fiducial \( \sigma_{\text{LT}} \)

The dPIE is implemented in LensTool through a fiducial velocity dispersion \( \sigma_{\text{LT}} \) related to the 1D central velocity dispersion \( \sigma_0 \) (in Eq. 3) by: \( \sigma_0 = \sqrt{3/2} \sigma_{\text{LT}} \).

To convert the model predicted aperture-average line-of-sight velocity dispersions \( \hat{\sigma}_{\text{gal}} \) and \( \hat{\sigma}_{\text{v}} \) to their fiducial LensTool values \( \hat{\sigma}_{\text{gal}}(\rho_{\text{LT}}) \), we relate them through:

\[
\hat{\sigma}_{\text{gal}}(\rho_{\text{LT}}) = \frac{\hat{\sigma}_{\text{gal}}(\rho_{\text{0}})}{c_p(R)}, \quad (17)
\]

where \( R \) is the aperture radius chosen for the cluster member spectral extraction. Adopting orbital isotropy and a (spherical) galaxy surface brightness profile proportional to the dPIE matter density, the projection coefficient \( c_p(R) \) is given by:

\[
c_p(R) = \frac{6 r_c + r_t}{\pi} \left( \sqrt{r^2 + R^2} - r_c - \sqrt{r^2 + R^2 + r_t^2} \right)^{-1} \times \\
\int_0^R \int_0^{\pi/2} \frac{r_c \arctan \left( \frac{r_c}{R} \right)}{\arctan \left( \frac{r_c}{R} \right)} \sqrt{r^2 - R^2} \, r \, dr \, dR, \quad (18)
\]

with \( r_c = r_{\text{core}}, r_t = r_{\text{cut}} \), see Bergamini et al. 2019.

4. Functional Tests

To test the ability of BayesLens in recovering the correct halos and sub-halos mass parameters and in predicting the correct multiple-image positions, we develop three mock galaxy clusters, i.e. Cluster-1, Cluster-2 and Cluster-3, with an increasing degree of complexity. Cluster-1 is a simple and well controlled cluster mock, while Cluster-2 and 3 have the main scope to analyse the role of low-mass cluster sub-structures and accuracy in the multiple-image positions. Despite none of the models claim to reproduce the complexity of real galaxy clusters, they represent ideal preliminary tests for our code, leaving to future works its application to more complex simulations (see Meneghetti et al. 2017) and to real cluster observations. In particular, the mocks allow us to verify the flexibility of our hierarchical approach, their dependence on the lensing or kinematics constraints, and their robustness against over-fitting.

Article number, page 5 of 16
Since our simple mocks are developed using LensTool, the results that we obtain are directly comparable to model inputs, thus to avoid the possibility of parameter misidentifications in the discussion of the results.

In this section, we describe the main characteristics of the simulated clusters and the settings adopted in the BayesLens optimization.

We emphasize that these are functional tests, i.e. on the behaviour of the method itself in recovering the input parameters of simple but realistic mocks. These are the very first tests that any
new method undergoes, before it is tested on less-well-controlled mocks (e.g. from simulations), to ensure that it can properly reproduce its own input parameters.

### 4.1. Cluster-1: The simplest toy cluster model

Cluster-1 is based on the best-fit LensTool lens model developed by Caminha et al. (2017b) (hereafter C-17) for the CLASH cluster MACS J1206.2−0847 at redshift $z = 0.439$. With respect to C-17, the cluster-scale component of our simulated model is made by a single halo, parametrized through a PIEMD mass profile. The center of the halo has an offset of 0.92$''$ from the brightest cluster galaxy (BCG) reference position. It has a core radius $r_{\text{core}} = 3.00''$, while its LensTool fiducial velocity dispersion has a value of $\sigma_{\text{LT}}^h = 1000.00\text{ km s}^{-1}$. Its ellipticity $e = (a^2 - b^2)/(a^2 + b^2)$ is fixed to $e = 0.70$, with a position angle of $\theta^h = 19.14''$ counterclockwise from West.

The clumpy sub-halo component of the mock cluster is composed by 138 cluster galaxies, selected from the C-17 members catalogue to have magnitudes in the HST/F160W filter given by $m_{\text{F160W}} < 22$. All these galaxies are parametrized as circular dPIE profiles whose $r_{\text{core}}$, $r_{\text{cut}}$, and $\sigma_{\text{LT}}^r$ values are determined as follows. For $r_{\text{core}}$ and $r_{\text{cut}}$ we adopt the scaling relations in Eq. (4) with slopes $\beta_{\text{core}} = 0.5$ and $\beta_{\text{cut}} = 0.66$. The two normalizations, computed at the BCG luminosity $L_0$ ($m_{\text{F160W}} = 17.19$, $r_{\text{ref}} = 0.01''$, and $r_{\text{ref}} = 5.0''$). We assign line-of-sight stellar velocity dispersions to the cluster member galaxies, averaged within a circular aperture, assuming a 15% Gaussian scatter around the scaling relation in Eq. (4) with $\alpha = 0.27$ and normalization equal to $350\text{ km s}^{-1}$ at the BCG luminosity. To determine the LensTool fiducial velocity dispersions, we deproject the aperture-averaged velocity dispersions using Eq. (17) and assuming apertures of radius $R = 0.8''$. No shear nor foreground structures are present in our Cluster-1 mock.

Given the total mass distribution for the mock cluster, we use LensTool to ray-trace the position of 15 background sources, randomly selected from the C-17 catalogue, to their multiple images on the lens plane. The sources are within a redshift range of $0.1 \leq z \leq 6.06$ and produce a total of 85 multiple images. Images are identified using an integer number and a letter in such a way that images with identical numbers in their ID belong to the same family. 26 of the 85 multiple images are excluded in C-17 due to light contamination. We also exclude one of the 15 families of multiple images (i.e. the 10-th family adopting imaging IDs in Fig. 1) since it is constituted by a single lensed image. Therefore, our final mock multiple images catalogue consists in 58 multiple images, from 14 background sources, shown in Fig. 1. The final simulated cluster, despite a purposely simple mass distribution, contains a number of halos, sub-halos and multiple images comparable to those of some CLASH or HFF clusters.

In our tests, we consider isotropic statistical errors of 0.2$''$ on the multiple-image positions. We also suppose that the stellar velocity dispersions of 58 luminous cluster galaxies are measured, within apertures of radius $R = 0.8''$. This is comparable to the number of velocity dispersions that Bergamini et al. (2019) measured from the MUSE data-cube of MACS J1206.2−0847. To associate an error to these simulated measures we use the following empirical relation derived from the measurements by Bergamini et al. (2019):

$$\frac{\delta \sigma_{\text{LT}}^r}{\sigma_{\text{LT}}^r} = 1.6 \times 10^{-3} m_{\text{F160W}}^3 - 8.53 \times 10^{-2} m_{\text{F160W}}^2 + 1.509 m_{\text{F160W}} - 8.879 \tag{19}$$

### 4.2. Cluster-2 and Cluster-3: Harder rungs with low-mass substructures and systematics in the astrometry of cluster galaxies and multiple images

Similarly to Cluster-1, the cluster-scale component of both Cluster-2 and Cluster-3 mocks is described through a single elliptical PIEMD profile. The PIEMD has an off-set of about 1.41$''$ from cluster BCG and it is characterized by an ellipticity $e^h = 0.72$, a position angle of $\theta^h = 19.80''$, a core radius $r_{\text{core}}^h = 11.0''$, and a LensTool fiducial velocity dispersion $\sigma_{\text{LT}}^h = 1080.00\text{ km s}^{-1}$.

The galaxy-scale component of the new mocks extends over a wider magnitude range from the BCG luminosity ($m_{\text{F160W}} = 17.19$) down to $m_{\text{F160W}} = 24.00$, and amounts to 258 galaxies extracted from the cluster member catalogue by C-17. Cluster galaxies are modelled as circular dPIEs whose $r_{\text{core}}$, $r_{\text{cut}}$, and $\sigma_{\text{LT}}^r$ are determined as in Sec. 4.1 but assuming $\sigma_{\text{ref}} = 295.5\text{ km s}^{-1}$, $\beta_{\text{cut}} = 0.64$ and $r_{\text{cut}} = 3.74''$. As in Cluster-1, a 15% Gaussian scatter around the $\sigma_{\text{LT}}$ scaling relation is assumed for the line-of-sight projected, aperture average, stellar velocity dispersion of galaxies. With this choice of dPIE parameter values, the cluster member population in Cluster-2 and Cluster-3 ranges from a total mass of $1.6 \times 10^{12}\text{ M}_\odot$ for the cluster BCG, down to a total mass of $1.1 \times 10^{10}\text{ M}_\odot$, for the less massive galaxy of the cluster.

In the following, we assume to know the measured stellar velocity dispersion, and associated errors, of 58 cluster galaxies (see Sec. 4.1). These velocities are given as inputs to BayesLens for the lens model optimization.

What differentiates Cluster-2 from Cluster-3 is that in the latter model we introduce, in addition to cluster galaxies, a large population of low mass sub-halos inside the simulated cluster. Thus, Cluster-3 is mainly designed to test the impact of numerous, undetected, sub-systems on BayesLens performances. To assign a mass to the faint sub-halos we exploit the sub-halo mass function fitted by Giocoli et al. (2010) and adopted by the MOKA algorithm for simulating the gravitational lensing by galaxy clusters (Giocoli et al. 2012, see). In using this formula, we assume a virial mass of $m_{\text{vir}} = 1.59 \times 10^{15}\text{ M}_\odot$ for the host cluster halo. This value corresponds to the total mass of MACS J1206-0847 (C-17) contained within a distance from the cluster center of $r_{200} = 2.06\text{ Mpc}$ (362.79$''$ at $z = 0.439$) that is expected to be close to the virial radius of the cluster (Umetsu et al. 2014). Note that a circularly symmetric PIEMD profile with $r_{\text{core}} = 11.0''$ and $\sigma_{\text{LT}}^r = 1080.00\text{ km s}^{-1}$, has a mass, within a radius $r = 362.79''$, exactly equal to $1.59 \times 10^{15}\text{ M}_\odot$. Low mass sub-halos are spatially distributed according to the function derived by Gao et al. (2004) fitting the results of cosmological simulations. In particular, we adopt a concentration parameter $c = r_c / R_{200} = 3.5$ for the host cluster halo and we assume $a = 1.944$ in Eq. 3 of Gao et al. (2004). Recent work by Meneghetti et al. (2020) shows a detailed comparison between sub-halos in state-of-the-art N-body and hydrodynamical simulations and observed sub-halos from a sizeable sample of cluster lens models.

\[Subm. to Science, Meneghetti private comm.\]
In Cluster-3, we consider all the faint sub-structures within a distance of 89.65″ from the cluster BCG (this is the distance of the most external galaxy considered into the model) and with masses ranging between $1.56 \times 10^9 M_{\odot}$, i.e. about the mass of the least massive cluster member, down to $0.93 \times 10^8 M_{\odot}$. Under these assumptions, the low mass sub-halo population of the Cluster-3 mock amounts to 910 profiles. Each low-mass sub-halo in the mock is parameterized as a circu-
Scaling relation hyperparameters ($mag^{15}$F160W = 17.19)

|          | $\alpha$ | $\sigma^{mag}_{15}$ [km s$^{-1}$] | $\beta^{mag}$ [arcsec] | $\sigma^{mag}_{15}$ [km s$^{-1}$] | $\rho^{mag}_{core}$ [arcsec] |
|----------|----------|----------------------------------|------------------------|----------------------------------|-------------------------------|
| True (Cluster-1) | 0.27 | 350.0 | 27.6 | 5.00 |
| BL (Cluster-1) | 0.27$^{+0.02}_{-0.02}$ | 351.3$^{+17.6}_{-15.7}$ | 28.9$^{+13.0}_{-11.3}$ | 4.50$^{+11.4}_{-0.71}$ |
| True (Cluster-2,3) | 0.28 | 295.5 | 22.7 | 3.74 |
| BL (Cluster-2) | 0.28$^{+0.02}_{-0.02}$ | 289.9$^{+11.5}_{-11.2}$ | 19.3$^{+12.2}_{-2.3}$ | 3.91$^{+10.8}_{-0.61}$ |
| BL (Cluster-3) | 0.28$^{+0.02}_{-0.02}$ | 290.1$^{+10.8}_{-12.0}$ | 19.3$^{+12.6}_{-2.1}$ | 3.90$^{+10.7}_{-0.64}$ |

| Galactic parameters | $\delta^{15}$F160W (mag$^{15}$F160W = 17.19) | $\delta^{15}$F160W (mag$^{15}$F160W = 17.72) | $\delta^{15}$F160W (mag$^{15}$F160W = 17.98) | $\delta^{15}$F160W (mag$^{15}$F160W = 18.21) |
|---------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| True (Cluster-1)    | 354.0 ± 12.9 | 278.2 ± 10.4 | 344.1 ± 12.9 | 230.8 ± 8.8 |
| BL (Cluster-1)      | 350.0$^{+32.1}_{-35.7}$ | 265.5$^{+41.1}_{-32.7}$ | 351.3$^{+15.2}_{-18.2}$ | 239.6$^{+30.9}_{-41.6}$ |
| True (Cluster-2,3)  | 292.9 ± 10.6 | 261.2 ± 9.7 | 202.0 ± 7.6 | 233.8 ± 8.9 |
| BL (Cluster-2)      | 290.5$^{+15.6}_{-15.8}$ | 252.3$^{+45.5}_{-42.8}$ | 208.3$^{+16.8}_{-16.4}$ | 231.3$^{+41.8}_{-42.4}$ |
| BL (Cluster-3)      | 292.1$^{+15.2}_{-16.4}$ | 256.8$^{+46.2}_{-35.6}$ | 209.4$^{+16.8}_{-17.1}$ | 234.9$^{+41.2}_{-42.4}$ |

| Galactic parameters | $\delta^{15}$F160W (mag$^{15}$F160W = 18.01) | $\delta^{15}$F160W (mag$^{15}$F160W = 18.30) | $\delta^{15}$F160W (mag$^{15}$F160W = 18.31) | $\delta^{15}$F160W (mag$^{15}$F160W = 18.66) |
|---------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| True (Cluster-1)    | 266.6 | 274.1 | 244.9 | 285.5 |
| BL (Cluster-1)      | 282.8$^{+27.4}_{-28.1}$ | 266.8$^{+28.1}_{-31.8}$ | 263.2$^{+39.1}_{-26.3}$ | 288.3$^{+14.0}_{-15.4}$ |
| True (Cluster-2,3)  | 174.3 | 217.6 | 210.4 | 258.8 |
| BL (Cluster-2)      | 231.6$^{+20.9}_{-20.6}$ | 218.9$^{+20.3}_{-20.8}$ | 218.4$^{+19.0}_{-19.7}$ | 216.1$^{+20.4}_{-20.3}$ |
| BL (Cluster-3)      | 233.3$^{+19.4}_{-20.5}$ | 217.8$^{+20.9}_{-20.9}$ | 217.1$^{+20.8}_{-19.9}$ | 225.6$^{+18.6}_{-18.3}$ |

Table 1. Comparison between the ‘true’ input values of the Cluster-1, Cluster-2 and Cluster-3 mocks and the BayesLens output results. The 50-th, 16-th and 84-th percentiles computed from the marginalized posteriors are quoted in the second line of each table. The first table refers to scaling relations hyperparameters, including the scatter of cluster galaxies around the inferred $\sigma$-mag relation. In the second table, we show the cluster-scale halo parameters. The third table shows the results for the four brightest galaxies with a ‘measured’ velocity dispersion: BCG, Gal(1), Gal(2) and Gal(3). Finally, the last table shows the results for the galaxy Gal(7) (see text) and for the three brightest galaxies without a ‘measured’ velocity dispersion: Gal(4), Gal(5), and Gal(6). The galaxy magnitudes are shown within round brackets.

We fit $M_{200}^{gal}$ to obtain a value for their $\sigma_{LT}$ and $r_{cut}$ parameters, we adopt the following procedure. Firstly, we compute the total mass for all the simulated cluster member galaxies ($M_{200}^{gal}$). Then, we fit a $M_{200}^{gal}$-$mag_{F160W}$ relation between galaxy masses and magnitudes. This relation is adopted to associate a fictitious magnitudes to each low-mass sub-halo. Finally, we use scaling relations in Eqs. 4 to derive velocity dispersions and core and truncation radii for the simulated sub-halos from their inferred magnitudes. As for cluster galaxies, also for the low-mass sub-halos we assume 15% Gaussian scatter around the $\sigma$-mag scaling relation.

Using Cluster-2 and Cluster-3 mocks, we map the position of 22 background sources, within the redshift range $z = 1.012$ up to $z = 5.793$ and randomly selected from the C-17 catalogue, to their multiple images on the lens plane. Both Cluster-2 and Cluster-3 amount to 69 magnified multiple images, whose positions are used as constrains for the lens model optimizations. As in Cluster-1, we still assume a statistical isotropic error of 0.2" on multiple image positions but an additional systematic uncertainty on the position of multiple images and cluster member...


| Cluster properties | Cluster-1 | Cluster-2 | Cluster-3 |
|--------------------|-----------|-----------|-----------|
| $N_{gal}^{tot}$ | 58 | 58 | 58 |
| $N_{gal}^{est}(Out SR)$ | 80 (80) | 200 (222) | 200 (222) |
| $N_{PH}$ | 0 | 0 | 910 |
| $(\Delta x_{est})_{1 \sigma}''$ | 0.0 | 0.01 | 0.01 |
| $N_{im}^{tot}$ | 58 | 69 | 69 |
| $(\Delta x_{im})_{1 \sigma}''$ | 0.2 | 0.2 | 0.2 |
| $(\Delta x_{im})_{3 \sigma}''$ | 0.0 | 0.01 | 0.01 |
| $\chi^2_{im}$ | 8.66 | 12.53 | 13.75 |
| $\Delta \chi^2_{tot}$ | 0.08 | 0.09 | 0.09 |

Table 2. Main features of the Cluster-1, Cluster-2 and Cluster-3 mocks.

The number of galaxies with and without a measured velocity dispersion is given by $N_{gal}^{est}$ and $N_{gal}^{tot}$ respectively. For the latter we quote also (in brackets) the number of galaxies, in BayesLens, that in principle are free to scatter around the best-fit $\sigma$-mag scaling relation. The systematic uncertainty assumed on galaxy positions is $(\Delta x_{est})_{1 \sigma}$. $N_{gal}^{est}$ is the total number of observed multiple images, while $(\Delta x_{im})_{1 \sigma}$ and $(\Delta x_{im})_{3 \sigma}$ are the statistical and systematic uncertainties assumed on their positions. In the last two lines, we quote the total chi-square and the r.m.s. displacement between the “measured” and model-predicted image positions.

galaxies is considered into the Cluster-2 and Cluster-3 mocks. To do so, we assign random Gaussian displacements to the ‘true’ positions, with dispersion of 0.01″ in both directions. These are comparable to the residual uncertainties from HST-like quality of imaging data.

In Fig. 5, we plot the total projected mass distribution of Cluster-2 and Cluster-3. Cluster member galaxies with measured velocity dispersions are encircled in green, while we mark the position of the ‘observed’ multiple images using cyan crosses. The spatial distribution of the low-mass sub-halo population included in the Cluster-3 mock is shown using magenta data-points on the right panel of the figure. In Tab. 2, we report a summary of the main proprieties of Cluster-1,2,3 mocks, such as: the number of galaxies with and without a measured velocity dispersion ($N_{gal}^{est}$ and $N_{gal}^{tot}$), the number of low-mass sub-halos ($N_{PH}$), the statistical and systematic uncertainties on galaxy positions and multiple images ($\Delta x_{est}$), $(\Delta x_{im})_{1 \sigma}$, $(\Delta x_{im})_{3 \sigma}$, and the number of observed multiple images ($N_{im}^{tot}$).

4.3. BayesLens on the mock clusters

In this subsection, we describe the parameter ranges adopted in BayesLens to probe the mass distribution of the Cluster-1, Cluster-2, and Cluster-3 mocks.

The cluster-scale component of all three lens models is parameterized using a single PIEMD profile. The $x$, $y$ coordinates of its center can vary within flat priors, 3″ wide, centered on the BCG position, while for the ellipticity $\epsilon^{h}$, position angle $\theta^{h}$, and fiducial velocity dispersion $\sigma^{h}$, we adopt uniform priors inside the following intervals respectively: 0.0 - 0.9, 5 - 35°, and 700 - 1300 km s$^{-1}$. Finally, for the PIEMD core radius $R_{core}$, we assume a uniform prior between 1″ and 7″ into the Cluster-1 lens model and between 3″ and 15″ for Cluster-2 and Cluster-3. The sub-halo component of lens models is parameterized using the three scaling relations in Eqs. (4) normalized at the BCG luminosity. In particular, we fix $\beta_{core}^{ref} = 0.01^\prime\prime$ and $\beta_{core}^{ran} = 0.5$, while the following flat priors are assumed on the others scaling relation parameters: Cluster-1 (0.20 ≤ $\alpha$ ≤ 0.34, 250 km s$^{-1}$ ≤ $\sigma^{ref} \leq 450$ km s$^{-1}$, 1" ≤ $R_{cut}^{ref}$ ≤ 11''); Cluster-2, and Cluster-3 (0.21 ≤ $\alpha$ ≤ 0.35, 150 km s$^{-1}$ ≤ $\sigma^{ref} \leq 350$ km s$^{-1}$, 1" ≤ $R_{cut}^{ref}$ ≤ 9'). Note that the slope of the $r_{cut}$-mag scaling relation, $\beta_{cut}$, is not a lens model free parameter since its value is derived from $\alpha$ through Eq. (5) with $\gamma = 0.2$. Differently from LensTool, in our code, the $\sigma$-mag cluster member scaling relation refers to measured (line-of-sight) stellar velocity dispersions averaged within apertures of radius $R = 0.8"$.

The BayesLens model optimization is performed on the lens plane using the positions of the multiple images inside the simulated catalogue and the 58 ‘measured’ stellar velocity dispersions of the cluster galaxies. These velocities are used on the one hand to determine the best $\sigma$-mag scaling relation parameters, and on the other to derive a Gaussian prior for each measured galaxy. Thanks to these priors, the lens model tends to prefer solutions with cluster members velocities that coincide with the measured values (see Eq. 12), unless the lensing data require them to deviate. The presence of a non-zero scatter also enables the models to fairly sample the whole parameter space, thereby avoiding the underestimation of systematics.

In running BayesLens on the Cluster-1 mock, we leave all galaxies free to scatter around the fitted $\sigma$-mag scaling relation, while in Cluster-2 and Cluster-3 lens models only the 58 galaxies with a measured velocity dispersion, and the 22 most luminous unmeasured galaxies are left free to vary. The number of free-to-scatter galaxies is one the inputs of our code and, in general, a lower number of free galaxies sensibly speed-up the optimization of the lens model.

The total posterior probability distribution ($p_{tot}$) of the Cluster-1 lens model is sampled using 298 walkers, while we use 500 walkers for the other two models. The walker initial positions are randomly distributed within the flat priors of the cluster-scale parameters ($\phi$), but they are initialized inside a Gaussian hyper-sphere, around an estimated maximum of the likelihood (in $p_{tot}$, $p_{mag}$, $P_{b}$) in the sub-space of the other free parameters.

In the following, we derive the marginalized posterior distribution of model free parameters by flattening the final Monte Carlo Markov chain (MCMC) of the walkers after removing a sufficiently large burn-in phase. Moreover, to determine the model-predicted multiple image positions and the r.m.s. displacement, $\Delta \chi^2_{tot}$, we use the best-fit BayesLens model. This best-fit model is defined as the set of free-parameter values that maximize the total posterior $p_{tot}$.

5. Results and discussion

In this section, we describe the main results obtained from an application of BayesLens on the three simulated clusters, namely Cluster-1, Cluster-2, and Cluster-3.

5.1. BayesLens results for Cluster-1

In Fig. 2, we show in black the marginalized BayesLens posterior distributions for the free parameters of the Cluster-1 lens model. The ‘true’ parameter values assumed in the mock cluster are marked with solid red lines. Panels a and b show the posterior distributions of the scaling relation and cluster-scale halo parameters, respectively. In panel c, we plot the marginalized posterior distributions for the stellar velocity dispersions.
of the four brightest ‘measured’ cluster galaxies: $\delta_{\text{BCG}}$, $\delta_{\text{Gal(1)}}$, $\delta_{\text{Gal(2)}}$ and $\delta_{\text{Gal(3)}}$. Similarly, panel d shows the velocity dispersions of the three brightest cluster galaxies without measured kinematics ($\delta_{\text{Gal(4)}}$, $\delta_{\text{Gal(5)}}$ and $\delta_{\text{Gal(6)}}$) and of the lens galaxy, $\delta_{\text{Gal(7)}}$, forming the galaxy scale strong lensing system zoomed in Fig 1. In these last two panels, we plot also blue solid lines showing the value of galaxy stellar velocity dispersions, as predicted by the best-fit $\sigma_{\text{ap}}$-$m_{\text{F160W}}$ scaling relation. The ‘true’ and BayesLens optimized values of the lens model free parameters are reported in Tab. 1. The latter correspond to the medians of parameter marginalized posterior distributions, while the 16-th and 84-th percentiles are quoted as an error. Thanks to the ‘measured’ cluster members stellar velocity dispersions and to the observed multiple-image positions, BayesLens recovers, well within the 1-$\sigma$ uncertainty, all of the mock true hyperparameters of the scaling relations and the cluster-scale halo parameters. To quantify the accuracy of BayesLens in determining the correct stellar velocity dispersions of the cluster members, we plot in Fig. 3 the aperture average velocity dispersions within aperture of 0.8'', $\sigma_{\text{gal(50th)}}$, as a function of the galaxy magnitudes, $m_{\text{F160W}}$. The solid black lines correspond to the scaling relation, with the dashed lines marking the 15% scatter in the mock. The (red) errorbars denote the ‘measured’ velocity dispersions with their uncertainties, and the (green) rectangles mark the 16th-to-84th percentiles from the posterior distributions. All of the galaxies lie on their measured velocity dispersions, and the zero-scatter solution still exists within the sampled parameter space. This shows that the loose prior on ‘measured’ galaxies still allows the models to generalize the zero-scatter Ansatz. The bottom panel shows the mock and posterior dispersions for the ‘unmeasured’ galaxies. Most of them are compatible with their true values (red dots) within the uncertainties. Most notably, one galaxy-scale strong lensing system (marked as Gal(7), see Fig. 1 closeup) departs automatically form the zero-scatter best-fit scaling relation and has a dispersion that is very well constrained by the lensing data. The compatibility of mock and posterior dispersions for all galaxies, within the intrinsic scatter, is displayed in Fig. 4, showing that there is no appreciable systematic mismatch trend between the ‘true’ and inferred dispersions.

Fig. 8 shows the cumulative mass profile of the whole cluster, projected along the line of sight, for the mock and the best-fit ($maximum a posteriori$) model. Even though the total mass of the mock cluster is redistributed in a slightly different way between the cluster halo and the member galaxies, its value at large radii is accurately recovered by the lens model. At the distances where most of the multiple images form, yielding more constraints to the lens model, the model predicts a correct value for the mass profiles within 1%. This is simply because the projected mass...
$M(< R_E) = \pi \Sigma_c R_E^2$ within the Einstein radius of a lens does not depend on its density profile.

To relate the goodness of fit of the models to a commonly used benchmark, we examine the r.m.s. offset of ‘measured’ vs inferred positions of the multiple images ($\Delta \chi^2_{\text{rms}}$). Our best-fit hierarchical model for Cluster-1 can reproduce all images with $\Delta \chi^2_{\text{rms}} = 0.08''$, while the lens model total chi-square is $\chi^2_{\text{tot}} = 8.66$ (see Tab. 2). This non-zero $\Delta \chi^2_{\text{rms}}$ shows that our hierarchical model does not over-fit, while also reproducing all observables well within their uncertainties.

The robustness against overfitting can be understood in two ways. First, if we re-run the models on a mock with 0.05'' statistical uncertainties on the image positions, the maximum-a-posteriori parameters do not change and the r.m.s. offset is still $\approx 0.08''$, while only the confidence intervals in the inferred parameters are shrunk. This is because all image-positions in this first mock are kept on their true locations, and the statistical uncertainties are simply a tolerance parameter that enables a smooth exploration of parameter space. Second, from the posterior we see that only about 10 galaxies deviate by more than 1$\sigma$ from the backbone of the scaling relation; if we re-run a LensToot model with only these galaxies free, and all other objects on the zero-scatter scaling relation, the $\chi^2$ improves by $\approx 33$, i.e. three times the change in degrees of freedom.

5.2. BayesLens results for Cluster-2 and Cluster-3

A comparison between ‘true’ values of the Cluster-2 and Cluster-3 parameters with BayesLens results is in Tab. 1. Note that in this table we are showing the same set of free parameters for the three mocks, however their ‘true’ values are different into the Cluster-1 and Cluster-2.3 mocks.

Despite the increasing complexity of the last two mock clusters, BayesLens recovers well within 1$\sigma$ all cluster-scale halo and scaling relations parameters. Moreover, it finds similar values for the stellar, aperture-average, velocity dispersions of measured and unmeasured cluster galaxies in the two mocks. In Fig 7, we plot marginalized posterior distributions, obtained from the Cluster-3 MCMC chains, for the same set of model parameters reported in Tab. 1. The figure shows that the four measured galaxies BCG, Gal(1), Gal(2), Gal(3), and the two unmeasured galaxies Gal(5), Gal(6) have marginalized posterior distributions that pick very close to the ‘true’ mock stellar velocity dispersions (red vertical lines). Conversely, we observed large deviations from ‘true’ values for Gal(7) and even more for Gal(4). The reason of these deviations is at the basis of the hierarchical approach implemented in BayesLens and it is due to the interplay between the $p_T$ and $p_{\text{im}}$ terms of the total posterior in Eq. (7). In fact, the unmeasured galaxy Gal(4) has marginalized posterior distribution centered on the stellar velocity dispersion value predicted by the best-fit $\sigma^*_{\text{mag}-F160W}$ scaling relation (marked with a blue line in Fig 7). In absence of constrains from the position of the observed multiple images, the $p_T$ term encourages a zero-scatter lens model solution with all unmeasured galaxy velocity dispersions lying on the best-fit scaling relation. Similarly, the displacement of the $\delta_{\text{Gal}(7)}$ posterior distribution from the scaling relation toward the ‘true’ mock value is possible because the increase of the $p_{\text{im}}$ term (due to a lower r.m.s. displacement of the images close to Gal(7)) overcomes the decrease of $p_T$ that penalizes scattered solutions. The best-fit Cluster-2 lens model reproduces the position of multiple images with a $\Delta \chi^2_{\text{rms}} = 0.09''$, while the total chi-square is $\chi^2_{\text{tot}} = 12.53$.

As displayed in Tab. 2, the addition of the low-mass sub-halo population into the Cluster-3 mock produces a small increase, equal to 1.22, of the lens model total chi-square, however the final $\Delta \chi^2_{\text{rms}}$ of multiple images remains essentially unchanged.

6. Conclusions

We have shown that, despite the (nominally) large number of degrees of freedom in cluster lensing, the exploration of flexible models is feasible thanks to tractable, hierarchical Bayesian inference. Unlike conventional models, where galaxies are placed on a razor-thin scaling relation and some are ‘freed’ ad hoc, we populate a scaling relation with non-zero scatter, with hyperparameters that are inferred directly from the lensing constraints and (when available) stellar kinematic data. Our tests on a realistic, albeit simplified, mock cluster show that the improvement over conventional models may be significant: the velocity dispersions of all the galaxies of our mock cluster are reliably recovered, as are the scaling-relation hyperparameters, and the r.m.s. displacement of multiple images between measurements and model predictions can be as low as 0.08''.

We have also tested the robustness of the hierarchical models against two systematic effects, i.e. dark subhalos and small systematic offsets in the measured image and cluster-member positions. Within realistic regimes, compatible with current (HST) or upcoming (ELT) depth and image quality, the model performance on the mock cluster does not degrade appreciably.

A crucial consequence of our hierarchical approach is that it also resolves a false dichotomy, between well-fitting cluster models and realistic population properties of galaxies. Zero-scatter models around a kinematic prior (dubbed LT$_{\text{kin}}$ above) have realistic galaxy populations but a higher r.m.s. offset than models without any kinematic prior (dubbed LT), whose scaling relations are significantly discrepant from the ‘true’ input relation. In zero-scatter models, there is typically a tradeoff between the r.m.s. and recovering realistic galaxy populations (see e.g. Bergamini et al. 2019). Our hierarchical models have realistic galaxy populations, and produce a small r.m.s. offset (on our mock cluster). With respect to zero-scatter models, BayesLens automatically identifies the galaxy-scale systems where more freedom from the scaling-relation backbone is needed, in order to reproduce the positions of multiple images around those galaxies with high accuracy. This makes the hierarchical models robust against overfitting, since in our mocks only 10-11 galaxies deviate from the scaling-relation backbone by more than 1$\sigma$ while the chi-square improves by $\Delta \chi^2 = 33$.

We should emphasize that this is a functional test on mocks, and additional effects may play a role in real-life systems, such as massive substructure, deviations from simple geometry of the main DM halo(s) and cluster members, and additional contributions along the line of sight. However, our hierarchical models can eliminate part of the systematic source of uncertainties, and show that internal systematics can be controlled.

The freedom in the parameters of each cluster-member galaxy, within the intrinsic scatter of their parent population, has multiple implications. First, the data themselves dictate which galaxy should be ‘freed’ and deviate significantly from a baseline scaling relation. Second, intrinsic scatter is a hyperparameter that is left free in the modeling inference, and this allows for a direct determination of galaxy-population properties, whence accurate studies of galaxy formation and evolution. In particular, accurate mass functions of cluster-member galaxies can be compared to predictions from cosmological simulations. Also, the freedom in cluster-member parameters allows (in principle)
for the quantification of environmental effects on galaxies at different locations across the cluster – which up to now has been possible only on stacked weak-lensing measurements (Niemiec et al. 2017).

An important feature of our inference is that BayesLens is fully modular. Changing specific python functions within our code, it is possible, in theory, to allow for calls to any chosen parametric lensing software (besides LensTool, used here as the benchmark), as well as to implement different mass profiles. Moreover, different prescriptions to relate the stellar velocity dispersions and lensing parameters of cluster-member galaxies can be used. The currently used scaling relations may be easily replaced with fundamental-plane relations (with free hyperparameters), so as to study the evolution of the fundamental plane across redshift and environment. Its modular structure also enables the use of additional constraints from e.g. flux ratios or extended-source reconstruction. Samples of magnification maps may be produced simply through an additional module, which may enable accurate studies of high-redshift galaxy populations. This is not central to this first paper, so for the sake of brevity, in Fig. 6, we show simply a comparison of magnification maps from the best-fit models in the three mock clusters. The magnification maps are well reproduced, and as can be expected the appreciable differences happen only across the critical lines and in high magnification regions ($\mu l > 20$), where the posterior also predicts a wider variance in the magnification. We remark that these models are already an extension over the state of the art, and the availability of samples form the posterior allows one to evaluate magnification uncertainties at all points.

While our set of hyperparameters is currently limited to those of the scaling relations (intercepts, slopes, scatter widths), this can be in principle extended to cosmological parameters, enabling (fully self-consistent) cosmographic measurements. Our code is released publicly and documented at https://github.com/pietrobergamini89/BayesLens.

Acknowledgements. PB and GBC thank Piero Rosati, Claudio Grillo, Massimo Meneghetti and Amata Mercurio who have made this work possible. We thank A. Sonnenfeld and M. Lombardi for useful comments on a previous version of this paper. PB acknowledges financial support from ASI though the agreement ASI-INAF n. 2018-29-HH.0. AA was supported by a grant from VILLUM FONDEN (project number 16599). This project is partially funded by the Danish council for independent research under the project “Fundamentals of Dark Matter Structures”, DFF–6108-00470.

References

Atek, H., Richard, J., Kneib, J.-P., & Schaerer, D. 2018, MNRAS, 479, 5184
Bacon, R., Accardo, M., Adjali, L., et al. 2012, The Messenger, 147, 4
Bergamini, P., Rosati, P., Mercurio, A., et al. 2019, A&A, 631, A130
Biviano, A., Rosati, P., Balestra, I., et al. 2013, A&A, 558, A1
Bonamigo, M., Grillo, C., Ettori, S., et al. 2018, ApJ, 864, 98
Bouwens, R. J., Oesch, P. A., Illingworth, G. D., Ellis, R. S., & Stefanon, M. 2017, ApJ, 843, 129
Bradač, M., Schneider, P., Lombardi, M., & Erben, T. 2005, A&A, 437, 39
Cammina, G. B., Grillo, C., Rosati, P., et al. 2016, A&A, 587, A80
Cammina, G. B., Grillo, C., Rosati, P., et al. 2017a, A&A, 600, A90
Cammina, G. B., Grillo, C., Rosati, P., et al. 2017b, A&A, 607, A93
Cammina, G. B., Rosati, P., Grillo, C., et al. 2019, A&A, 632, A36
Cerny, C., Sharon, K., Andrade-Santos, F., et al. 2018, ApJ, 859, 159
Co, D., Salmon, B., Bradač, M., et al. 2019, ApJ, 884, 85
Diego, J. M., Protopapas, P., Sandvik, H. B., & Tegmark, M. 2005, MNRAS, 360, 477
Dye, S., & Taylor, A. 1998, MNRAS, 300, L23
Ellisdóttir, A., Limousin, M., Richard, J., et al. 2007, ArXiv e-prints [arXiv:0710.5636]

3 The github link is proprietary until acceptance.
Fig. 6. Magnification maps for the three simulated clusters computed for a source redshift $z_s = 3.0$. In the first line, we plot magnification maps for the ‘true’ mock Cluster-1, Cluster-2 and Cluster-3 mocks. Maps in the second line are obtained from BayesLens best-fit lens models of the clusters, while in the last line we show residuals of the ratio: Mock cluster/BayesLens.

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Publications of the Astronomical Society of the Pacific, 125, 306

Gao, L., White, S. D. M., Jenkins, A., Stoehr, F., & Springel, V. 2004, MNRAS, 355, 819

Gilmore, J. & Natarajan, P. 2009, MNRAS, 396, 354

Giocoli, C., Meneghetti, M., Bartelmann, M., Moscardini, L., & Boldrin, M. 2012, MNRAS, 421, 3343

Article number, page 14 of 16
Fig. 7. Same as Fig. 2 but for the Cluster-3 mock.
Fig. 8. Top: Projected cumulative total mass profiles for the mock Custer-1 mock as a function of the projected distance from the center of the BCG. The red line corresponds to the true mass profile. The black solid line is obtained from the best-fit BayesLens model. The projected distances of the multiple images from the BCG center are shown as vertical black lines. Bottom: Relative variation of the BayesLens total mass profile with respect to the true mass profile of the mock cluster. The blue dashed line in the bottom panel marks a 2% relative difference between the mock and model mass profiles. In both panels, the grey-shaded stripes delimit the 16th-84th percentiles of the mass profiles with parameters drawn from the posterior.