

SLUG IV: a novel forward-modelling method to derive the demographics of star clusters

Mark R. Krumholz ¹★, Angela Adamo,^{2,3} Michele Fumagalli ⁴ and Daniela Calzetti⁵

¹Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2612, Australia

²Centre of Excellence for Astronomy in Three Dimensions (ASTRO-3D), Australia

³Department of Astronomy, Oskar Klein Centre, Stockholm University, SE-10691 Stockholm, Sweden

⁴Institute for Computational Cosmology and Centre for Extragalactic Astronomy, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK

⁵Department of Astronomy, University of Massachusetts – Amherst, Amherst, MA 01003, USA

Accepted 2018 October 23. Received 2018 October 22; in original form 2018 March 2

ABSTRACT

We describe a novel method for determining the demographics of a population of star clusters, for example distributions of cluster mass and age, from unresolved photometry. This method has a number of desirable properties: it fully exploits all the information available in a data set without any binning, correctly accounts for both measurement error and sample incompleteness, naturally handles heterogeneous data (e.g. fields that have been imaged with different sets of filters or to different depths), marginalizes over uncertain extinctions, and returns the full posterior distributions of the parameters describing star cluster demographics. We demonstrate the method using mock star cluster catalogues and show that our method is robust and accurate, and that it can recover the demographics of star cluster populations significantly better than traditional fitting methods. For realistic sample sizes, our method is sufficiently powerful that its accuracy is ultimately limited by the accuracy of the underlying physical models for stellar evolution and interstellar dust, rather than by statistical uncertainties. Our method is implemented as part of the Stochastically Lighting Up Galaxies (SLUG) stellar populations code, and is freely available.

Key words: methods: data analysis – methods: statistical – techniques: photometric – galaxies: star clusters: general.

1 INTRODUCTION

Stars form in regions where the stellar density is vastly higher than the mean for the galactic field. Over tens to hundreds of Myr after their formation, stars disperse from these birthplaces, leaving behind a small fraction of long-lived, gravitationally bound old star clusters. This process of formation and dispersal encodes a great deal of physics regarding the formation of stars, the expulsion of gas from star-forming clouds, and the dynamical evolution of stellar systems in a galactic potential. For recent reviews, see Krumholz (2014), Krumholz et al. (2014), and Longmore et al. (2014).

Because of the physics it encodes, the distribution of star cluster ages and masses has long been an important topic of study, both observationally and theoretically. Theoretical models for cluster dispersal have emphasized processes such as gas expulsion (e.g. Baumgardt & Kroupa 2007; Parmentier et al. 2008; Krumholz & Matzner 2009; Fall, Krumholz & Matzner 2010; Murray, Quataert &

Thompson 2010), tidal disruption of clusters after gas expulsion (e.g. Lamers et al. 2005; Gieles, Lamers & Portegies Zwart 2007; Kruijssen 2009, 2012; Kruijssen et al. 2012; Elmegreen & Hunter 2010), and two-body relaxation and evaporation over long time-scales (e.g. Fall & Zhang 2001). These models predict a variety of functional forms for the joint age and mass distribution of surviving star clusters. Observational studies have attempted to measure these quantities for star clusters in the Milky Way (Williams & McKee 1997; Lada & Lada 2003; Borissova et al. 2011), the Magellanic Clouds (Hunter et al. 2003; Rafelski & Zaritsky 2005; Chandar, Fall & Whitmore 2010; Popescu, Hanson & Elmegreen 2012), and more distant systems (Zhang et al. 1999; Larsen 2002; Goddard, Bastian & Kennicutt 2010; Chandar et al. 2010, 2011; Bastian et al. 2012; Fall & Chandar 2012; Fouesneau et al. 2012, 2014; de Meulenaer et al. 2015; Krumholz et al. 2015a; Johnson et al. 2016, 2017; Adamo et al. 2017; Messa et al. 2018), with the goal of testing the predictions of these models.

Because it is not at present possible to resolve the individual stars in young star clusters beyond the Milky Way and its few nearest neighbours, observational studies that go beyond samples of a few

* E-mail: mark.krumholz@gmail.com

galaxies are generally restricted to working with unresolved light, where the raw data consist of measurements of luminosities in some set of filters for each star cluster. Consequently, there is an urgent need for robust statistical techniques to derive the physical properties of star cluster populations from such integrated light data; the method we introduce below is intended for this type of analysis.

The traditional approach for analysing these data is to assign an age and mass to each cluster by comparing their unresolved luminosities and colours to a set of evolutionary tracks for simple stellar populations, with the best-fitting mass and age determined by χ^2 minimization or a similar procedure. Once the masses and ages are determined, the clusters are placed in mass and age bins, and the distribution in the population as a whole can, in principle, be measured. However, such an approach encounters several difficulties. First, the process of binning inevitably discards some of the information present in the original data, and fitting parameters to binned distributions can introduce severe biases (e.g. Maschberger & Kroupa 2009). Second, at low masses, and for certain age and colour combinations even at higher masses, the assignment of mass and age to an individual cluster is highly uncertain, and the errors in the assignments are not well approximated by simple Gaussians. Instead, the posterior probability distribution function (PDF) of mass and age can have a complex, multi-peaked shape (Popescu & Hanson 2009, 2010a,b; Foesneau et al. 2014; de Meulenaer et al. 2015; Krumholz et al. 2015a). A single best-fitting mass and age may be a very poor representation of the PDF for a single cluster, but the process of assigning a cluster to a single bin ignores this complexity. Third and most seriously, determining the properties of the population as a whole requires considering the completeness of the observed sample. Variations in whether and how one takes completeness into account can lead to quite different inferences in the final physical distributions (e.g. Lamers 2009). Part of the reason for this sensitivity is that completeness is a function of the luminosity and surface brightness profile of the cluster, the background, and the level of crowding in the image, leading to a completeness that has a complex functional form in mass–age–extinction space.

The simplest approach to handling the problem of completeness is to be extremely conservative, and discard all data in regions of parameter space where the observations are not complete or nearly so. However, this invariably requires one to discard much of the available data. A somewhat more sophisticated approach is forward modelling: rather than deriving the mass and age distribution of the population from estimates of mass age for individual clusters, one could instead consider a proposed distribution of masses and ages, predict the resulting photometry distribution including the effects of incompleteness, and adjust parameters of the mass and age distribution until they match the observations. Approaches of this type are widely used in astronomy, for example to infer star formation histories or stellar mass distributions from observed colour–magnitude diagrams (CMDs; e.g. Dolphin 2002; Harris & Zaritsky 2009; Weisz et al. 2013; Conroy & van Dokkum 2016; see Cerviño 2013 for a review). However, methods of this type have not previously been applied to deriving the properties of populations of star clusters, at least in part due a unique challenge not present in other applications. In existing applications such as CMD fitting, the forward model is deterministic, i.e. for a given stellar mass, age, and other properties, there is a single predicted colour and magnitude. This is not the case for star clusters with masses $\lesssim 3000 M_{\odot}$, because such clusters are too small to fully sample the stellar initial mass function (IMF, e.g. Cerviño & Luridiana 2004, 2006; da Silva, Fumagalli & Krumholz 2012). As a result, two clusters of the same total mass and age can produce wildly different luminosities and

colours. This means that the forward model is not deterministic, but instead depends on an additional random variable that couples non-linearly with the deterministic variables like cluster mass and age. This situation presents computational challenges that are not addressed by existing methods.

In this paper, we introduce a new approach for determining the distribution of the properties of star clusters from unresolved photometry that allows us to consider arbitrary functional forms for distributions of mass, age, and extinction, and to exploit all the information available in heterogenous data (i.e. data where not all fields are observed with the same filters or to the same depth). Crucially, it naturally accounts for both incomplete observations and the uncertainties in the assignment of masses and ages to individual clusters that arise when the mapping between physical properties and luminosity is non-deterministic due to finite sampling. The essence of our approach is to consider a proposed distribution of physical parameters, determine the corresponding luminosity distribution in a probabilistic way so that we preserve the non-unique mapping between physical properties and photometry, apply the completeness function in observed luminosity space, and then compare to the data. We then adjust the underlying physical distribution until the best match to the observations is found. We implement this method using fast numerical algorithms that enable us to identify the parameters describing a cluster distribution on a workstation-level computer in ~ 10 h of computing time. The software is based on the Stochastically Lighting Up Galaxies (SLUG) software suite (da Silva et al. 2012; da Silva, Fumagalli & Krumholz 2014; Krumholz et al. 2015b), and is freely available from the SLUG website, <http://www.slugsp.com/cluster-population-pipeline>.

The plan for the remainder of this paper is as follows. In Section 2, we describe our new method and the computational techniques we use to implement it. In Section 3, we test the method on mock data to verify its accuracy and demonstrate its capabilities, and in Section 4, we compare the performance of our new method to more conventional approaches. We summarize our findings in Section 5.

2 METHOD

2.1 Statement of the problem

Our goal is to infer the mass and age distribution of a population of star clusters in a galaxy for which we have a sample of star clusters observed in some set of photometric bands; we must include extinction as an additional nuisance parameter, which we will marginalize over. The method we develop to achieve this goal generalizes and extends the one proposed by Weisz et al. (2013) for inferring the IMF of a resolved population of individual stars. Formally, let $g(M, T, A_V | \theta)$ be the joint distribution of mass, age, and visual extinction for the underlying population, which depends on a vector of parameters θ .¹ For example, if we were to assume that the mass, age, and A_V distributions are separable power laws, then θ would contain the minimum, maximum, and slope of each power law. We wish to infer a posterior distribution for θ . Since the true size of the cluster population is not known a priori, and our

¹Metallicity is another potential physical parameter, but for simplicity, we will assume that the cluster-to-cluster variation in metallicity is small enough that its effects can be neglected. This assumption is particularly likely to be valid for optical data, since metallicity has relatively little effect on optical bands, and mostly affect near-infrared colours (Anders et al. 2004). The generalization to include metallicity as a parameter is straightforward.

observations are inevitably incomplete at the low-luminosity end of the distribution, we must also treat the number of clusters present N_c as a parameter of the model, although we will see below that it is more convenient to transform to a different variable.

The data from which we will make this inference consist of a set of N_{obs} observed star clusters, and for the i th star cluster we observe its absolute magnitude or luminosity $L_{F,i}$ in N_F different photometric filters F , measured with some photometric error $\sigma_{F,i}$, which we take to be known and Gaussian-distributed. For notational compactness, let \mathbf{L}_i and $\boldsymbol{\sigma}_i$ be the luminosities / absolute magnitudes and the corresponding errors for the i th cluster in all N_F filters, and $\{\mathbf{L}_i\}$ and $\{\boldsymbol{\sigma}_i\}$ be the set of all such observed luminosities and uncertainties for all clusters in every filter.

Finally, let us assume that each of our observations has a known completeness function described by $P_{\text{obs},i}(\mathbf{L}')$. This function is the probability that a cluster of intrinsic luminosity \mathbf{L}' , observed in the same manner as observed cluster i (i.e. with the same integration time and set of filters, in a field at the same distance) will be included in the sample.² Note that \mathbf{L}' is distinct from the quantity \mathbf{L} introduced in the previous paragraph: the former is the true luminosity of a cluster, while the latter is the measured luminosity, which is slightly different due to observational error. A simple magnitude limit corresponds to $P_{\text{obs},i}(\mathbf{L}')$ being a step function. In practice, this function must be determined by artificial cluster tests or the like. Note that we explicitly allow for the possibility that different sets of observations may have different completeness limits, for example if we are combining data from two different galaxies at different distances, or from two fields within the same galaxy that were observed to different depths.

2.2 Posterior probability for a cluster population

As usual in a Bayesian approach, we write the posterior probability distribution of the model parameters $(\boldsymbol{\theta}, N_c)$ given the data as the product of the prior probabilities with the likelihood function, i.e.³

$$p(\boldsymbol{\theta}, N_c | \{\mathbf{L}_i\}, \{\boldsymbol{\sigma}_i\}, N_{\text{obs}}) \propto p(\{\mathbf{L}_i\}, N_{\text{obs}} | \boldsymbol{\theta}, N_c, \{\boldsymbol{\sigma}_i\}) p_{\text{prior}}(\boldsymbol{\theta}, N_c). \quad (1)$$

We remind readers that, in this equation $\boldsymbol{\theta}$ is the vector of parameters describing the distribution of star cluster properties, N_c is the true number of clusters in the observed region, $\{\mathbf{L}_i\}$ and $\{\boldsymbol{\sigma}_i\}$ are the vector of observed cluster luminosities or magnitudes and their corresponding uncertainties, and N_{obs} is the number of observed clusters. The likelihood function $p(\{\mathbf{L}_i\}, N_{\text{obs}} | \boldsymbol{\theta}, N_c, \{\boldsymbol{\sigma}_i\})$ is simply the probability density of the data given the model and the observational errors, while $p_{\text{prior}}(\boldsymbol{\theta}, N_c)$ is the prior probability distribution for the parameters $\boldsymbol{\theta}$ and N_c , and $p(\boldsymbol{\theta}, N_c | \{\mathbf{L}_i\}, \{\boldsymbol{\sigma}_i\}, N_{\text{obs}})$ is the posterior PDF that we are attempting to compute. To evaluate it, we assume that the observed luminosity of each cluster represents an independent draw from an underlying distribution of star cluster luminosities, $p_L(\mathbf{L} | \boldsymbol{\theta}, \boldsymbol{\sigma})$; note that \mathbf{L} here is the observed luminosity, not the intrinsic one, and because the uncertainties $\boldsymbol{\sigma}$ are not

the same from one measurement to another, the luminosity distributions for each cluster are not identical. We defer a calculation of $p_L(\mathbf{L} | \boldsymbol{\theta}, \boldsymbol{\sigma})$ to the next section.

We assume that both the intrinsic luminosities of clusters and the observational errors on them are uncorrelated.⁴ Under this assumption, and since the number of observed clusters N_{obs} is also an independent variable, we can write the likelihood function for the cluster population as a product of the probability distributions for individual clusters and for N_{obs} ,

$$p(\{\mathbf{L}_i\}, N_{\text{obs}} | \boldsymbol{\theta}, N_c, \{\boldsymbol{\sigma}_i\}) \propto P_N(N_{\text{obs}} | \boldsymbol{\theta}, N_c) \prod_{i=1}^{N_{\text{obs}}} p_L(\mathbf{L}_i | \boldsymbol{\theta}, \boldsymbol{\sigma}_i). \quad (2)$$

Here $P_N(N_{\text{obs}} | \boldsymbol{\theta}, N_c)$ is the probability that we will observe N_{obs} clusters from a population of N_c whose intrinsic luminosity distribution is parametrized by $\boldsymbol{\theta}$.

To determine $P_N(N_{\text{obs}} | \boldsymbol{\theta}, N_c)$, first consider the simplest case where the observed clusters all come from a single field imaged with a single set of filters and uniform sensitivity across it. In this case, there is a single completeness function $P_{\text{obs}}(\mathbf{L}')$, and for a cluster population with a distribution of intrinsic luminosities $p_{L'}(\mathbf{L}' | \boldsymbol{\theta})$, for any set of population parameters $\boldsymbol{\theta}$ there is a single probability

$$P_{\text{obs}}(\boldsymbol{\theta}) = \int P_{\text{obs}}(\mathbf{L}') p_{L'}(\mathbf{L}' | \boldsymbol{\theta}) d\mathbf{L}' \quad (3)$$

that a randomly selected cluster will be observed. In this case, the number of clusters we expect to observe is $N_{\text{ex}} = P_{\text{obs}}(\boldsymbol{\theta}) N_c$, and since each observation of one of the N_c clusters present is an independent experiment, the actual number observed must be Poisson-distributed:

$$P_N(N_{\text{obs}} | N_{\text{ex}}) = \frac{N_{\text{ex}}^{N_{\text{obs}}} e^{-N_{\text{ex}}}}{N_{\text{obs}}!}. \quad (4)$$

In this expression, we have suppressed the dependence of N_{ex} on $\boldsymbol{\theta}$ for the sake of compactness.

Now consider the more general case where we have multiple fields with different sensitivities and filter sets, and thus different completeness functions. For each such observation j , there will be some number of clusters $N_{\text{ex},j}$ that one would expect to detect, which is a function of both the true number of clusters in the observed field and the observational completeness function for it. The number of clusters $N_{\text{obs},j}$ that is actually observed in each field will then be Poisson-distributed per equation (4), and the total number of clusters expected in the full catalogue of all fields is just $N_{\text{ex}} = \sum_j N_{\text{ex},j}$. However, the sum of a number of random variables that are each drawn from a Poisson distribution is itself Poisson-distributed. Thus, $P_N(N_{\text{obs}} | \boldsymbol{\theta}, N_c)$ must be distributed following equation (4) even for heterogeneous observations.

Because $P_N(N_{\text{obs}} | \boldsymbol{\theta}, N_c)$ depends only on N_{ex} , it is convenient to eliminate N_c in favour of N_{ex} as the variable for which we will seek a posterior PDF. That is, rather than trying to compute $p(\boldsymbol{\theta}, N_c |$

²Here and throughout, we use lowercase p to denote PDFs, and uppercase P to denote simple, dimensionless probabilities.

³A note on notation, since all probability distributions can be properly normalized by requiring that their integrals be unity, in what follows we usually omit normalization constants and write out all dependencies as proportionalities. The only exceptions are cases where we retain the normalization constant for clarity.

⁴The assumption of uncorrelated noise may not be strictly true in a real observation, since the dominant uncertainty in real observations is usually the aperture correction. This may lead to errors that depend on the level of crowding or background, and thus are correlated with respect to the locations of clusters within the target galaxy. However, this would represent a correlation of error with cluster position. As long as there is no correlation of the error with cluster physical properties, this does not matter for our purposes.

$\{\mathbf{L}_i\}$, $\{\boldsymbol{\sigma}_i\}$, N_{obs}), we will instead compute

$$p(\boldsymbol{\theta}, N_{\text{ex}} | \{\mathbf{L}_i\}, \{\boldsymbol{\sigma}_i\}, N_{\text{obs}}) \propto P_N(N_{\text{obs}} | N_{\text{ex}}) \prod_{i=1}^{N_{\text{obs}}} p_L(\mathbf{L}_i | \boldsymbol{\theta}, \boldsymbol{\sigma}_i). \quad (5)$$

We have therefore written the likelihood function for our observed cluster population in terms of the luminosity distribution for a single cluster and the expected number of observed clusters, with the dependence on N_{ex} separable from that on $\boldsymbol{\theta}$. Of course, once one has determined the posterior distributions of $\boldsymbol{\theta}$ and N_{ex} , one could use these to obtain the posterior distribution of N_c . In practice, however, this is unlikely to be interesting, for the following reason: star cluster mass functions are invariably observed to be steep, such that by number most clusters have low masses. Thus, the value of N_c will depend strongly on the shape of the mass function at low masses. Since real extragalactic observations invariably become incomplete at masses significantly larger than the smallest possible star cluster mass, any parameters we introduced to describe the shape of the lower part of the mass function (e.g. a lower mass cut-off), will not be constrained by the observations, and since N_c depends critically upon them, it will be unconstrained by the observations as well.

2.3 The distribution of observed luminosities for individual clusters

The final step in our derivation is to compute the distribution of observed luminosities for an individual cluster, $p_L(\mathbf{L} | \boldsymbol{\theta}, \boldsymbol{\sigma})$, where we remind readers that \mathbf{L} is the vector of *observed* luminosities, which are the result of taking the true luminosity \mathbf{L}' and measuring it with some finite error $\boldsymbol{\sigma}$. To do so, we assume that there is a distribution of *intrinsic* cluster luminosities \mathbf{L}' that is identical for every cluster, and that depends only the model parameters: $p_{L'}(\mathbf{L}' | \boldsymbol{\theta})$. We can then obtain the observed luminosity distribution by marginalizing over the intrinsic luminosity of each cluster:

$$p_L(\mathbf{L} | \boldsymbol{\theta}, \boldsymbol{\sigma}) \propto \int p_{L'}(\mathbf{L}' | \boldsymbol{\theta}) p(\mathbf{L} | \mathbf{L}', \boldsymbol{\sigma}) P_{\text{obs}}(\mathbf{L}') d\mathbf{L}', \quad (6)$$

where $p(\mathbf{L} | \mathbf{L}', \boldsymbol{\sigma})$ is the probability that a cluster of intrinsic luminosity \mathbf{L}' will yield an observed luminosity \mathbf{L} when measured with uncertainty $\boldsymbol{\sigma}$. Under our assumption that the observational uncertainties are Gaussian, this is

$$p(\mathbf{L} | \mathbf{L}', \boldsymbol{\sigma}) = \frac{(2\pi)^{-N_F/2}}{\prod_{n=1}^{N_F} \sigma_n} \exp \left[-\sum_{n=1}^{N_F} \frac{(L_n - L'_n)^2}{2\sigma_n^2} \right] \equiv \mathcal{N}(\mathbf{L} | \mathbf{L}', \boldsymbol{\sigma}), \quad (7)$$

where the sum runs over all N_F filters, and we have introduced the notation $\mathcal{N}(\mathbf{x} | \mathbf{x}_0, \boldsymbol{\sigma})$ to represent the usual multidimensional normal distribution centred on \mathbf{x}_0 with standard deviation $\boldsymbol{\sigma}$ and no covariance, evaluated at position \mathbf{x} .

We estimate the intrinsic luminosity distribution convolved with the probability of being observed using the method described by Krumholz et al. (2015b), and implemented in the `cluster.slug` module in the SLUG software package. Specifically, given a library of simulated clusters, where cluster j has a mass M_j , age T_j , extinction $A_{V,j}$, and a vector of luminosities \mathbf{L}'_j , we write the intrinsic luminosity distribution using a kernel density

estimation model,

$$p_{L'}(\mathbf{L}' | \boldsymbol{\theta}) P_{\text{obs}}(\mathbf{L}') \propto \sum_{j=1}^{N_{\text{lib}}} w_j(\boldsymbol{\theta}) \mathcal{N}(\mathbf{L}' | \mathbf{L}'_j, \mathbf{h}). \quad (8)$$

Here, the sum runs over all N_{lib} clusters in the simulation library, \mathbf{h} is the bandwidth of the kernel density estimation, and the weights w_j are given by

$$w_j(\boldsymbol{\theta}) = P_{\text{obs}}(\mathbf{L}'_j) \frac{g(M_j, T_j, A_{V,j} | \boldsymbol{\theta})}{p_{\text{lib}}(M_j, T_j, A_{V,j})}, \quad (9)$$

where $g(M, T, A_V | \boldsymbol{\theta})$ is the proposed distribution of mass, age, and extinction, and $p_{\text{lib}}(M, T, A_V)$ is the distribution from which the library was sampled.

For the purposes of developing intuition, it is helpful to examine the weight factors $w_j(\boldsymbol{\theta})$ factors in a bit more detail. The meaning of the first term, $P_{\text{obs}}(\mathbf{L}'_j)$, is simple: it simply downweights the contribution of each library cluster to the observed luminosity distribution by the probability that will actually be observed. The factor $p_{\text{lib}}(M, T, A_V)$ in the denominator of equation (9) simply represents the frequency with which we drew a particular combination of (M, T, A_V) in the process of constructing the library; that is, the number of sample library clusters that fall into a particular infinitesimal range in mass, age, and extinction is just proportional to $p_{\text{lib}}(M, T, A_V)$. By contrast, $g(M, T, A_V | \boldsymbol{\theta})$ is the number of sampled points that we *would* have had in that bin if our library had been drawn from the distribution described by the parameters $\boldsymbol{\theta}$. Thus, the ratio of these two terms, to which $w_j(\boldsymbol{\theta})$ is proportional, simply represents the ratio of the number of clusters we *should* have for a particular set of parameters $\boldsymbol{\theta}$ to the number we *actually* used when we constructed our library. For example, if $g(M, T, A_V | \boldsymbol{\theta}) = (1/2)p_{\text{lib}}(M, T, A_V)$ at some particular point (M, T, A_V) , this means that our library has twice as many clusters in that neighbourhood as it should give the value of $\boldsymbol{\theta}$, and thus when attempting to compute the luminosity distribution $p_L(\mathbf{L}' | \boldsymbol{\theta})$, we should only count our library samples as half a cluster each, $w_j(\boldsymbol{\theta}) = 1/2$. Note that our procedure imposes a restriction on $p_{\text{lib}}(M, T, A_V)$: it must be non-zero at any point in (M, T, A_V) -space where $g(M, T, A_V | \boldsymbol{\theta})$ is non-zero for any set of parameters $\boldsymbol{\theta}$, i.e. the support of the library must encompass the support of all candidate distributions describing the population. If this condition is not satisfied, then $w_j(\boldsymbol{\theta})$ diverges.

We now evaluate equation (6) using equation (7) for $p(\mathbf{L} | \mathbf{L}', \boldsymbol{\sigma})$ and equation (8) for $p_{L'}(\mathbf{L}' | \boldsymbol{\theta}) P_{\text{obs}}(\mathbf{L}')$. This gives

$$\begin{aligned} p_L(\mathbf{L} | \boldsymbol{\theta}, \boldsymbol{\sigma}) &\propto \int \sum_{j=1}^{N_{\text{lib}}} w_j(\boldsymbol{\theta}) \mathcal{N}(\mathbf{L}' | \mathbf{L}'_j, \mathbf{h}) \mathcal{N}(\mathbf{L} | \mathbf{L}', \boldsymbol{\sigma}) d\mathbf{L}' \\ &\propto \sum_{j=1}^{N_{\text{lib}}} w_j(\boldsymbol{\theta}) \int \mathcal{N}(\mathbf{L}' | \mathbf{L}'_j, \mathbf{h}) \mathcal{N}(\mathbf{L} | \mathbf{L}', \boldsymbol{\sigma}) d\mathbf{L}' \\ &= \mathcal{A}(\boldsymbol{\theta}) \sum_{j=1}^{N_{\text{lib}}} w_j(\boldsymbol{\theta}) \mathcal{N}(\mathbf{L} | \mathbf{L}'_j, \mathbf{h}'), \end{aligned} \quad (10)$$

where $\mathbf{h}' = \sqrt{\mathbf{h}^2 + \boldsymbol{\sigma}^2}$, with the sum is computed elementwise. In the second step, we use the linearity of the integration operator to exchange the sum and the integral and take the weights $w_j(\boldsymbol{\theta})$ out of the integral because they do not depend on \mathbf{L}' ; in the final step, we make use of the standard result for the integral of the product of normal distributions. The quantity $\mathcal{A}(\boldsymbol{\theta})$ that we have added to the final line is a normalization constant chosen to ensure

that $\int p_{\mathbf{L}}(\mathbf{L} | \boldsymbol{\theta}, \boldsymbol{\sigma}) d\mathbf{L} = 1$, and is given by

$$\mathcal{A}(\boldsymbol{\theta}) = \left[\sum_{j=1}^{N_{\text{lib}}} w_j(\boldsymbol{\theta}) \right]^{-1}. \quad (11)$$

Inserting this into equation (2) gives the complete specification of the likelihood function,

$$p(\{\mathbf{L}_i\}, N_{\text{obs}} | \boldsymbol{\theta}, N_{\text{ex}}, \{\boldsymbol{\sigma}_i\}) \propto P_{\text{N}}(N_{\text{obs}} | N_{\text{ex}}) \mathcal{A}(\boldsymbol{\theta})^{N_{\text{obs}}} \prod_{i=1}^{N_{\text{obs}}} \left[\sum_{j=1}^{N_{\text{lib}}} w_j(\boldsymbol{\theta}) \mathcal{N}(\mathbf{L}_i | \mathbf{L}_j, \mathbf{h}') \right], \quad (12)$$

where $P_{\text{N}}(N_{\text{obs}} | N_{\text{ex}})$ is a Poisson distribution with expectation value N_{ex} . As noted above, since in practice we cannot constrain N_{ex} from observations, we can regard it as a nuisance parameter to be marginalized over. The remaining problem of determining the best-fitting parameters $\boldsymbol{\theta}$, and exploring the shape of the posterior probability distribution in the vicinity of this maximum in order to determine uncertainties, can then be solved using any number of methods. Our implementation uses the EMCEE package (Foreman-Mackey et al. 2013), a Markov Chain Monte Carlo (MCMC) algorithm.

Numerical evaluation of equation (12) requires some care, because the right-hand side involves a very large number of terms. A typical catalogue might contain several thousand observed clusters, and the SLUG libraries we use contain 10^7 sample clusters; thus equation (12) involves 10^{10} – 10^{11} terms. Since any method of finding the maximum likelihood invariably involves evaluating the likelihood function hundreds of thousands of times, brute force evaluation of equation (12) is impractically slow. We can avoid this problem by noting that the normal distribution $\mathcal{N}(\mathbf{L}_i - \mathbf{L}_j | \mathbf{h}')$ is negligibly small for most combinations of \mathbf{L}_i and \mathbf{L}_j , because for a the great majority of clusters in the library $|\mathbf{L}_i - \mathbf{L}_j|^2 / 2\mathbf{h}'^2 \gg 1$. That is, only a tiny fraction of \mathbf{L}_j values are near any given \mathbf{L}_i , and these nearby clusters completely dominate the inner sum in equation (12). In Appendix A, we describe an algorithm that exploits this fact to evaluate the sum in order $\ln N_{\text{lib}}$ rather than order N_{lib} time. Combined with openMP parallelization over the outer product, this algorithm enables us to evaluate equation (12) for each value of $\boldsymbol{\theta}$ and five-filter photometry in a roughly 1 s on a workstation, making MCMC optimization of the fit parameters practical.

3 MOCK CATALOGUE TESTS

3.1 Generation of mock catalogues

To demonstrate the capabilities of our new method, we carry out a series of tests on mock data. We generate mock star cluster data sets by running SLUG to draw a certain number of clusters from specified mass, age, and extinction distributions, and for each cluster to calculate the photometric magnitude in the *Hubble Space Telescope* (HST) WFC3 filters F275W, F336W, F438W, F555W, and F814W; for shorthand below, we refer to these filters as *UV*, *U*, *B*, *V*, and *I*. Although our method can handle heterogenous data without difficulty, for simplicity in this demonstration of it, we assume that all fields are images in these same five filters. For all the tests presented in this section, unless otherwise noted, we adopt a Chabrier (2005) IMF for the stars, we compute stellar evolution using the MESA Isochrones and Stellar Tracks (MIST) version 1.0 tracks for stars born rotating at 40 per cent of breakup (Dotter 2016; Choi

et al. 2016),⁵ and using SLUG’s default option (‘sb99’) for stellar atmospheres (Leitherer et al. 1999; Vázquez & Leitherer 2005). We include extinction with an extinction law given by SLUG’s Milky Way extinction curve, and nebular emission using SLUG’s default treatment, with a ratio of nebular to stellar extinction drawn from a Gaussian distribution with a mean of 2.1 and a dispersion of 0.5, based on the empirically determined distribution found by Kreckel et al. (2013). All tests use Solar metallicity, and all assume a constant star formation rate of $\dot{M}_* = 1 M_{\odot} \text{ yr}^{-1}$, with star formation at this rate having begun a time T_{sf} in the past, and continuing to the present. For simplicity, and because extinction is a nuisance parameter, we assume that all catalogues have an extinction distribution of the form

$$p(A_{\text{V}}) \propto \exp\left(-\frac{A_{\text{V}}^2}{2\sigma_{A_{\text{V}}}^2}\right) \quad (13)$$

with $\sigma_{A_{\text{V}}} = 0.5$ mag, and A_{V} restricted to be >0 .

3.1.1 Mass and age distributions

The different mock data sets differ only in their assumed distributions of mass and age. We consider a population of clusters born with a mass distribution

$$p(M_i) \propto M_i^{\alpha_{\text{M}}} \exp\left(-\frac{M_i}{M_{\text{break}}}\right) \quad (14)$$

above some minimum mass M_{min} . The corresponding expectation value for the cluster mass is

$$\langle M_i \rangle = M_{\text{break}} \frac{\Gamma(2 + \alpha_{\text{M}}, M_{\text{min}}/M_{\text{break}})}{\Gamma(1 + \alpha_{\text{M}}, M_{\text{min}}/M_{\text{break}})}, \quad (15)$$

where $\Gamma(a, z)$ is the incomplete Γ function. Assuming that a fraction f_{c} of stars are born in star clusters, the total number of clusters formed is

$$N_{\text{form}} = f_{\text{c}} T_{\text{sf}} \frac{\dot{M}_*}{\langle M_i \rangle}. \quad (16)$$

For the age distributions, we consider two possibilities that have been advocated in the literature. Some authors (e.g. Fall & Chandar 2012; Chandar, Fall & Whitmore 2015; Chandar et al. 2017) argue for mass-independent (mid) cluster disruption, whereby the probability that a given cluster survives to a particular time is independent of its mass, at least for ages below a few Gyr. In this formulation, the probability that a cluster survives to age T is described by a power law,

$$p_{\text{s,mid}} = \begin{cases} 1, & T < T_{\text{mid}} \\ (T/T_{\text{mid}})^{\alpha_{\text{T}}}, & T > T_{\text{mid}} \end{cases} \quad (17)$$

for $\alpha_{\text{T}} \leq 0$. In this case the corresponding joint distribution of cluster mass and age is

$$\frac{d^2 N}{dM dT} \propto M^{\alpha_{\text{M}}} \exp\left(-\frac{M}{M_{\text{break}}}\right) \max(T, T_{\text{mid}})^{\alpha_{\text{T}}}, \quad (18)$$

so that at ages $T > T_{\text{mid}}$, we have the usual power-law age distribution $dN/dT \sim T^{\alpha_{\text{T}}}$ usually adopted in mid-models. We require that this distribution only apply at ages $T > T_{\text{mid}}$, and be flat at younger ages,

⁵The MIST models make use of the MESA stellar evolution code (Paxton et al. 2011, 2013, 2015).

because otherwise the distribution would diverge as $T \rightarrow 0$.⁶ For this distribution, the fraction of clusters formed that have survived to the present day, assuming $T_{\text{sf}} \geq T_{\text{mid}}$, is

$$f_{s,\text{mid}} = \begin{cases} (1 + \ln \chi) / \chi, & \alpha_T = -1 \\ (1/\chi^{\alpha_T} - \alpha_T/\chi)/(1 - \alpha_T), & \alpha_T \neq -1 \end{cases} \quad (19)$$

where $\chi = T_{\text{sf}}/T_{\text{mid}}$ is the number of disruption times for which star formation has been ongoing.

The other possibility is that star clusters undergo mass-dependent disruption (mdd), as proposed for example by Lamers et al. (2005) and Gieles (2009). In this model, clusters lose mass at a rate that varies as a power law with their current mass, $dM/dT \propto M^{\gamma_{\text{mdd}}}$ with $0 \leq \gamma_{\text{mdd}} \leq 1$, so that at age T a cluster born with mass M_i will have a mass

$$M = M_i \left[1 - \gamma_{\text{mdd}} \left(\frac{M_{\text{min}}}{M_i} \right)^{\gamma_{\text{mdd}}} \frac{T}{T_{\text{mdd},\text{min}}} \right]^{1/\gamma_{\text{mdd}}} \quad (20)$$

If the second term in square brackets is >1 , then the cluster is considered to have disrupted completely. Here $T_{\text{mdd},\text{min}}$ is the time-scale over which a cluster of initial mass M_{min} loses all its mass and disappears.⁷ In this case, the distribution of present-day masses and ages is

$$\frac{d^2 N}{dM dT} \propto \frac{d^2 N}{dM_i dT} \frac{dM}{dM_i} \quad (21)$$

$$\propto M^{\alpha_M} \eta^{\alpha_M + 1 - \gamma_{\text{mdd}}} \exp\left(-\eta \frac{M}{M_{\text{break}}}\right) \quad (22)$$

where

$$\eta(M, T) \equiv \left[1 + \gamma_{\text{mdd}} \left(\frac{M_{\text{min}}}{M} \right)^{\gamma_{\text{mdd}}} \frac{T}{T_{\text{mdd},\text{min}}} \right]^{1/\gamma_{\text{mdd}}} \quad (23)$$

is the ratio of the initial mass to the present mass for a cluster of present-day mass M and age T . Since a cluster born of age T must have been formed with a mass larger than $M_{s,\text{min}} = M_{\text{min}}(\gamma_{\text{mdd}} T / T_{\text{mdd},\text{min}})^{1/\gamma_{\text{mdd}}}$ to have survived, the fraction of clusters of age T that have survived mdd is

$$f_{s,\text{mdd}}(T) = \frac{\int_{\max(M_{s,\text{min}}(T), M_{\text{min}})}^{\infty} M^{\alpha_M} e^{-\frac{M}{M_{\text{break}}}} dM}{\int_{M_{\text{min}}}^{\infty} M^{\alpha_M} e^{-\frac{M}{M_{\text{break}}}} dM} \quad (24)$$

$$= \frac{\Gamma\left(1 + \alpha_M, \frac{\max(M_{s,\text{min}}(T), M_{\text{min}})}{M_{\text{break}}}\right)}{\Gamma\left(1 + \alpha_M, \frac{M_{\text{min}}}{M_{\text{break}}}\right)} \quad (25)$$

⁶Physically, the assumption that the age distribution is flat below some minimum age T_{mid} is expected on dynamical grounds. Even if there is a disruption mechanism that unbinds clusters on time-scales below T_{mid} , there is no way to determine from photometry that stars have become unbound until they begin to disperse, and they cannot disperse on time-scales less than a cluster crossing time. Thus, regardless of the nature of any physical disruption mechanism, the observed cluster age distribution must match the star formation rate distribution (i.e. must be independent of age) at times less than the typical cluster crossing time.

⁷Note that the conventional choice for mdd models is to normalize to the disruption time for a cluster of initial mass $10^4 M_{\odot}$, denoted t_4 , or an initial mass of $1 M_{\odot}$, denoted t_0 . We have instead chosen to normalize at M_{min} instead, simply to avoid introducing an extra parameter. Since the disruption time is simply a power law in the mass, our $T_{\text{mdd},\text{min}}$ parameter is related to the more usual t_4 or t_0 trivially: $T_{\text{mdd},\text{min}} = t_4(M_{\text{min}}/10^4 M_{\odot})^{\gamma_{\text{mdd}}}$, and similarly for t_0 .

Averaged over all ages, the total fraction of surviving clusters is

$$f_{s,\text{mdd}} = \frac{1}{T_{\text{sf}}} \int_0^{T_{\text{sf}}} f_{s,\text{mdd}}(T) dT \quad (26)$$

The integral cannot be evaluated in closed form, but is trivial to evaluate numerically for any specified set of parameters.

3.1.2 Mock catalogues

Our joint mass–age distribution is fully characterized by a choice to use mdd or mid, and by six parameters: M_{min} , T_{sf} , α_M , M_{break} , and either α_T and T_{mid} (for mid-models) or γ_{mdd} and $T_{\text{mdd},\text{min}}$ (for mdd models). The parameters M_{min} and T_{sf} cannot generally be determined from observations of the type we are considering because clusters near the minimum mass or maximum age are invariably too dim to observe; they enter the problem only by changing the total number of clusters in the catalogue. Since this effect is degenerate with changes to the value of f_c (the fraction of stars formed in clusters) or \dot{M}_* (the total star formation rate), we simply set $M_{\text{min}} = 100 M_{\odot}$ and $T_{\text{sf}} = 10$ Gyr for all catalogues, and do not explore variations in these parameters further. For the remaining parameters, we consider three mock catalogues, whose parameters are summarized in Table 1 that illustrate different possible combinations of them. We tune our parameters so that each catalogue produces a comparably sized sample of observable clusters, with the size chosen to be about the size of the catalogue for NGC 628 presented by Adamo et al. (2017), which contains approximately 3700 clusters (though not all are visually confirmed). The cases are:

(i) **Powerlaw**: this case uses a mock catalogue with clusters drawn from a distribution similar to that proposed by, e.g. Fall & Chandar (2012), and Chandar et al. (2015, 2017), whereby the mass function is a pure powerlaw whose upper limit set only by size of sample effects, and disruption is mass independent. Specifically, we adopt for this catalogue $\alpha_M = -2$, $\alpha_T = -1$, $M_{\text{break}} = 10^{6.5} M_{\odot}$, and $T_{\text{mid}} = 10^{6.5}$ yr. Note that for this choice of M_{break} , the expected number of clusters that form with $M > M_{\text{break}}$ is $\ll 1$, so the mass distribution is effectively a pure power law truncated only by finite sample size. Also adopting $f_c = 1$ for this case, the expected number of clusters surviving to the present day is

$$N = N_{\text{form}} f_{s,\text{mid}} = 2.93 \times 10^4 \quad (27)$$

We therefore draw this number clusters for the mock catalogue.

(ii) **Truncated**: this case is similar to the results obtained by Fouesneau et al. (2014) and Johnson et al. (2017) for Andromeda. Disruption is mass-independent, but no disruption occurs until ages above 100 Myr, i.e. $T_{\text{mid}} = 10^8$ yr. The mass function is truncated at a lower mass, $M_{\text{break}} = 10^5 M_{\odot}$. Finally, only 10 per cent of stars form in clusters, so $f_c = 0.1$. All other parameters are the same as for Powerlaw. For this distribution, the mean cluster mass is $\langle M \rangle = 638 M_{\odot}$ and the number of surviving clusters is $N = 8.78 \times 10^4$.

(iii) **Mass-dependent disruption (MDD)**: this catalogue uses parameters chosen to be similar to those obtained by Adamo et al. (2017). For this case we use mdd, with $\gamma_{\text{mdd}} = 0.65$ and $T_{\text{mdd},\text{min}} = 9.5$ Myr.⁸ The mass distribution at birth is the same as for Truncated, i.e. $\alpha_M = -2$ and $M_{\text{break}} = 10^5 M_{\odot}$. We take $f_c = 0.3$ in this case, so the total number of clustered formed is $N_{\text{form}} = 4.71 \times 10^6$. Using equation (26), the fraction of clusters

⁸Note that $T_{\text{mdd},\text{min}} = 9.5$ Myr is equivalent to Adamo et al.'s parameter $t_4 = 190$ Myr.

Table 1. Parameters of mock catalogues, and results of fits to these parameters using SLUG. The first row lists the number of observed clusters in each mock catalogue. The second row specifies whether the catalogue was generated using mid or mdd, and the values $w(\text{mid})$ and $w(\text{mdd})$ that we report are the Akaike weights of the mid and mdd models as determined from our MCMC fits; see the main text for details. For all other parameters, we give the true value used in generating the catalogue, and we list fit values in the form $(q_{50})^{+(q_{84}-q_{50})}_{-(q_{50}-q_{16})}$ where q_N is the N th percentile estimate for q . Thus, the value reported is the 50th percentile, and the + and – error range indicates the range from the 16th to the 84th percentile. While we calculate fit parameters for both mid and mdd models for each mock catalogue, in the table above we report the fits only for whichever of the two models has the higher Akaike weight.

Parameter		Powerlaw	Truncated	MDD	DoubleErr	LibMismatch	CompMismatch
N_{obs}		5629	5167	5479	5255	5088	5093
mid / mdd	True	mid	mid	mdd	mid	mid	mid
	$w(\text{mid})$	1.0	1.0	$<10^{-10}$	1.0	1.0	1.0
	$w(\text{mdd})$	$<10^{-10}$	$<10^{-10}$	1.0	$<10^{-10}$	$<10^{-10}$	$<10^{-10}$
α_M	True	–2	–2	–2	–2	–2	–2
	Fit	$-2.00^{+0.018}_{-0.021}$	$-2.01^{+0.028}_{-0.033}$	$-2.05^{+0.019}_{-0.039}$	$-2.00^{+0.019}_{-0.021}$	$-1.91^{+0.030}_{-0.033}$	$-1.97^{+0.020}_{-0.023}$
$\log(M_{\text{break}}/M_{\odot})$	True	6.5	5.0	5.0	5.0	5.0	5.0
	Fit	$6.36^{+0.51}_{-0.39}$	$4.90^{+0.17}_{-0.12}$	$5.20^{+0.16}_{-0.09}$	$4.90^{+0.10}_{-0.08}$	$4.76^{+0.10}_{-0.08}$	$4.96^{+0.11}_{-0.11}$
α_T	True	–1	–1	–	–1	–1	–1
	Fit	$-1.02^{+0.021}_{-0.022}$	$-1.01^{+0.071}_{-0.056}$	–	$-0.97^{+0.040}_{-0.038}$	$-0.95^{+0.040}_{-0.035}$	$-1.04^{+0.042}_{-0.040}$
$\log(T_{\text{mid}}/\text{yr})$	True	6.5	8.0	–	8.0	8.0	8.0
	Fit	$6.49^{+0.029}_{-0.031}$	$8.00^{+0.047}_{-0.050}$	–	$7.96^{+0.031}_{-0.030}$	$7.90^{+0.042}_{-0.041}$	$7.97^{+0.035}_{-0.032}$
γ_{mdd}	True	–	–	0.65	–	–	–
	Fit	–	–	$0.61^{+0.030}_{-0.036}$	–	–	–
$\log(T_{\text{mdd, min}}/\text{yr})$	True	–	–	6.98	–	–	–
	Fit	–	–	$7.05^{+0.134}_{-0.049}$	–	–	–

that survive mdd is $f_{s, \text{mdd}} = 0.00370$, so the expected number of clusters at the present day is $N = 1.74 \times 10^4$.

(iv) `DoubleErr`: this catalogue is identical to `Truncated`, except that the assumed photometric errors added to the true cluster luminosities are twice as large – see below. Its purpose is to test how our results depend on the size of the photometric error.

(v) `LibMismatch`: this catalogue is identical to `Truncated` in its parameters, but instead of generating the mock catalogue using MIST models for stellar evolution and a Milky Way extinction curve, we generate them using SLUG’s Padova tracks (Girardi et al. 2000) and starburst attenuation curve. The goal of this catalogue is to test the robustness of our method in a case where the models for stellar evolution and dust are not a perfect match to the underlying data.

(vi) `CompMismatch`: this catalogue is identical to `Truncated` in its parameters, but uses a different completeness function (see below). The goal of this catalogue is to test how our method behaves when our estimated completeness function is not exactly correct.

3.1.3 Completeness and photometric error

To test the effects of observational completeness and photometric error, and show how our method copes with them, we next add noise to our mock catalogues, and apply completeness cuts to them. We set the photometric noise level for all catalogues except `DoubleErr` to 0.1 mag in all bands,⁹ based on typical levels of photometric accuracy in recent large surveys such as Legacy Extragalactic UV Survey (LEGUS) (Adamo et al. 2017). To test the sensitivity of our results to the noise level, for `DoubleErr` we set the noise level to 0.2 mag instead, comparable to the poorest levels of accuracy in LEGUS. For either noise level, we generate the observed magnitudes of all clusters by taking the true magnitudes in each band

⁹Here and throughout we use Vega magnitudes.

calculated by SLUG and adding a random offset drawn from a Gaussian distribution with a dispersion 0.1 or 0.2 mag, as appropriate for that catalogue.

We next apply a completeness cut, using a completeness function comparable to that obtained by Adamo et al. (2017) for the galaxy NGC 628 based on their mock cluster tests. Specifically, for all the catalogues except `CompMismatch`, we take the probability that a given cluster makes it into the catalogue to be 100 per cent for $V \leq -5$ mag, 0 for $V \geq -4$ mag, and linearly varying between these two limits for $-5 \text{ mag} < V < -4 \text{ mag}$, i.e. $P_{\text{obs}} = -V - 4$ mag. For `CompMismatch`, we instead use a completeness function that is 100 per cent for $V \leq -4.75$ mag, 0 for $V \geq -4$ mag, and varies in between these two limits as $P_{\text{obs}} = [(-V - 4)/0.75]^2$. For each cluster, we assign a flag of ‘observed’ or ‘not observed’ based on the V magnitude; for clusters in the partially complete range, we randomly assign them one flag or the other with probability $P_{\text{obs}}(V)$. This process yields a list of ≈ 5000 observed clusters for of our mock catalogues; the exact number in each case is given in Table 1.

As an example of the effects of noise and the completeness cut, Fig. 1 shows the true distribution of cluster physical properties in the `Powerlaw` catalogue, and the corresponding distribution for those clusters flagged as observed. Figs 2 and 3 show the corresponding observed CMD and colour–colour diagram. As the plots shows, observational completeness truncates both the mass and age distributions, and does so in a way that is correlated – clusters are more likely to remain in the catalogue if they are either young or massive, and are mostly removed if they are old and low mass. However, the cut-off imposed by observational limits is not sharp in either mass or age due to the effects of stochastic sampling, varying extinction, and partial completeness. For example, for a fully sampled (i.e. non-stochastic), unextincted stellar population with a mass of $300 M_{\odot}$, the ages corresponding to 100 per cent, 50 per cent, and 0 per cent completeness ($V = -5, -4.5$, and -4 mag, respectively) are 12.0, 19.1, and 53.7 Myr, respectively. However, in our `Powerlaw` mock catalogue, we find that there are 13 observed clusters with mass $< 300 M_{\odot}$ at ages above 53.7 Myr, and 388 non-observed clusters larger than $300 M_{\odot}$ with ages below 12.0 Myr.

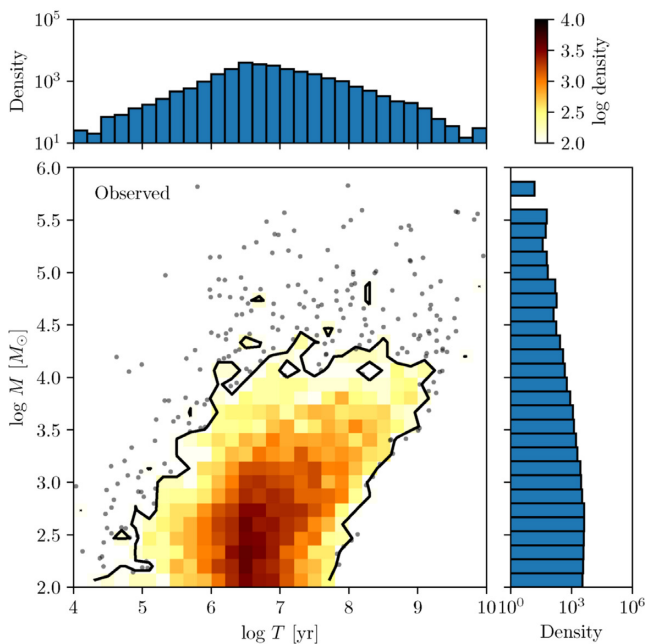
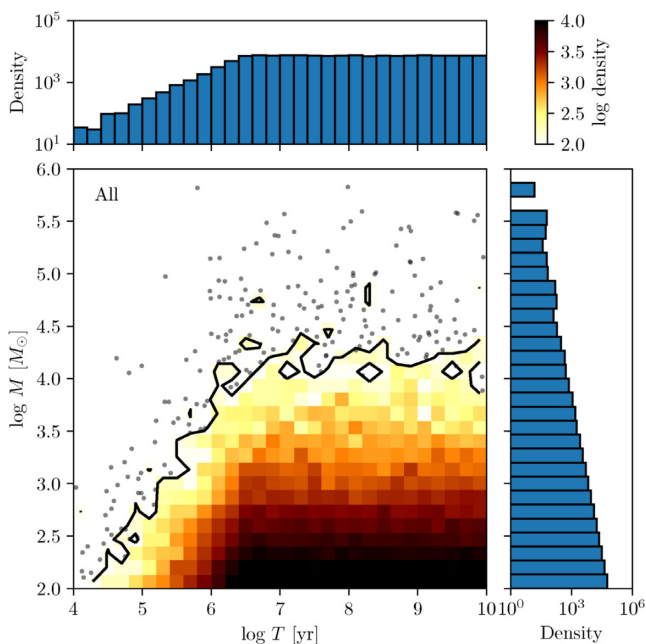


Figure 1. Distribution of clusters in the Powerlaw mock catalogue. The top set of panels shows the distribution of all clusters, while the bottom set shows the distribution of those clusters that are observed. In each set of panels, the central one shows the density of clusters (in clusters per dex²), as indicated by the colour bar; points mark individual clusters in sparsely populated regions. The enclosing contour corresponds to a number of clusters per dex² equal to the lowest value in the colour bar. Above and to the right of the central panel, we show 1D histograms of the mass and age distributions, in units of clusters per dex (i.e. the quantities plotted are $dN/d\log M$ and $dN/d\log T$).

The conventional means of avoiding this complication in analysing cluster populations is to impose fairly severe cuts on the data so as to ensure that the sample that is retained is well within the zone of completeness, and massive enough to be relatively unaffected by stochasticity. However, this approach is undesirable

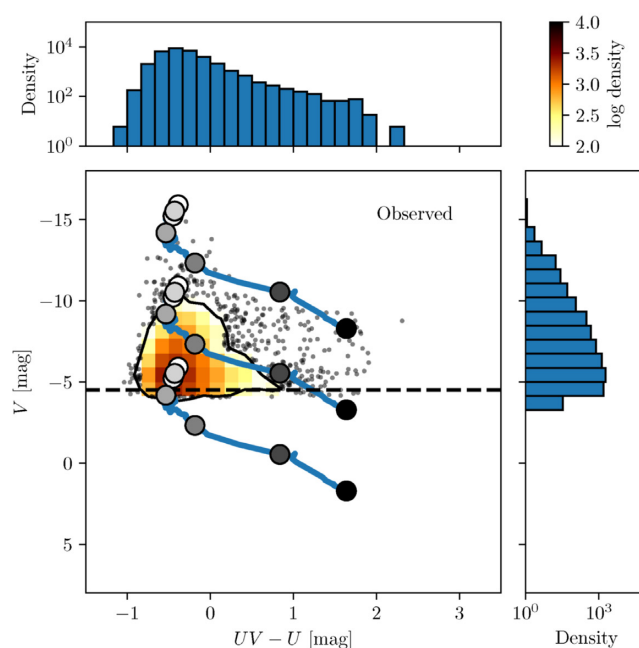
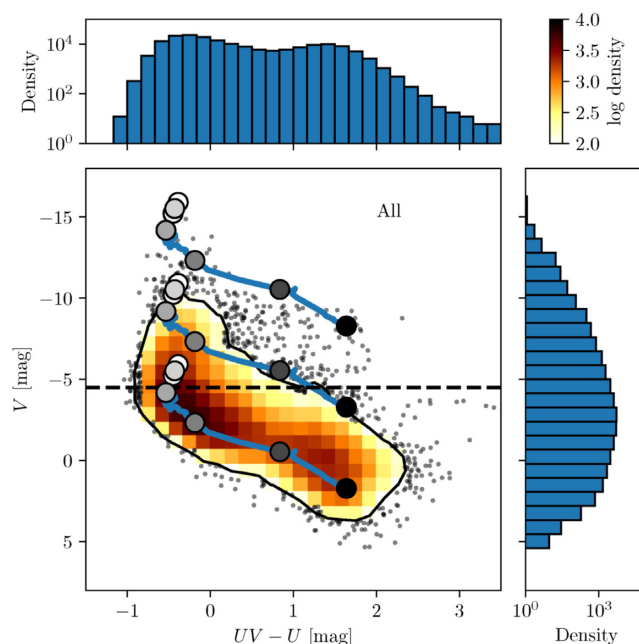


Figure 2. Same as Fig. 1, except that now we show the distribution of clusters in colour and magnitude rather than in their physical properties. The blue lines with large grey points show evolutionary tracks for unextincted clusters with fully sampled (i.e. non-stochastic) stellar populations over the age range from 10^5 to 10^{10} yr. From top to bottom, the lines correspond to cluster masses of 10^6 , 10^4 , and $10^2 M_{\odot}$. The points are logarithmically spaced in age from 10^5 yr (lightest) to 10^{10} yr (darkest) at intervals of 1 dex. The dashed black line indicates the 50 per cent completeness limit.

because it both discards much of the available information and restricts the range of applicability of the resulting fits. For example, our Powerlaw mock catalogue contains 5629 observed clusters. Discarding all clusters with $V > -6$ mag, as is done for example in Adamo et al. (2017), leaves only 2445, and thus amounts to throwing out more than half the data. Any fits to the remaining data could be used to constrain the distributions of mass and age only

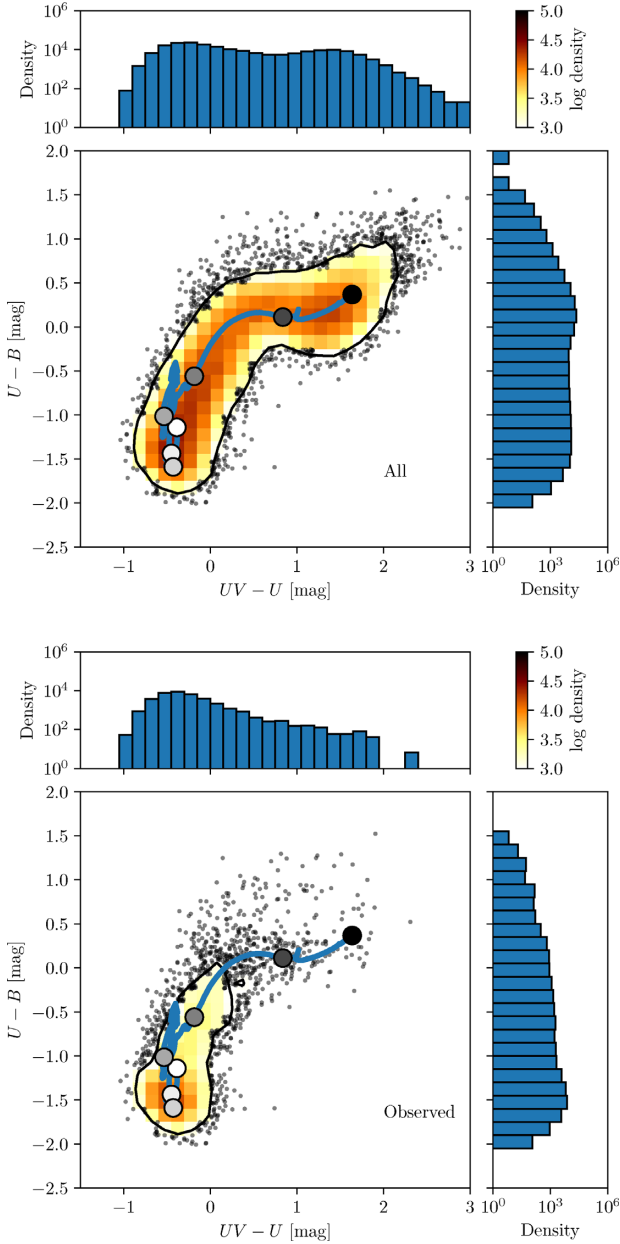


Figure 3. Same as Fig. 2, except that now we show the distribution in colour–colour space rather than colour–magnitude space.

for masses above $10^4 M_{\odot}$ and ages below 1 Gyr (or masses above $10^{3.7} M_{\odot}$ and ages below 200 Myr, the cuts used in Adamo et al. 2017), since only in this mass–age range are the data reasonably complete.

3.2 Libraries

To analyse the mock catalogues, we require libraries of model clusters. To produce these, we use SLUG to simulate 10^7 clusters using the same combination of tracks, atmospheres, and metallicity as for all the mock catalogues except `LibMismatch`. Cluster masses, ages, and extinctions in the library are drawn randomly from the distribution

$$p_{\text{lib}}(M, T, A_V) \propto p_M(M) p_T(T) p_{A_V}(A_V) \quad (28)$$

with

$$p_M(M) \propto \begin{cases} \left(\frac{M}{10^5 M_{\odot}}\right)^{-1}, & 10^2 < \frac{M}{M_{\odot}} \leq 10^5 \\ \left(\frac{M}{10^5 M_{\odot}}\right)^{-2}, & 10^5 < \frac{M}{M_{\odot}} \leq 10^7 \end{cases} \quad (29)$$

$$p_T(T) \propto \frac{1}{T}, \quad 10^5 \text{ yr} < T < 1.5 \times 10^{10} \text{ yr} \quad (30)$$

$$p_{A_V}(A_V) \propto \text{const}, \quad 0 < A_V < 3 \text{ mag}. \quad (31)$$

This sampling is chosen to maximize the density of samples at younger masses and ages where there is a larger amount of stochastic variation in cluster colour and luminosity; we remind the reader that the sampling density is explicitly accounted for in equation (9), and thus this choice affects the result only in so far as it provides a better or worse sampling of the underlying distribution. All other details of the sampling procedure are identical to that used in Krumholz et al. (2015a), and we refer readers to that paper for full details.

We adopt a bandwidth of $h = 0.05$ dex in the physical dimensions and $h = 0.05$ mag in the photometric directions. Note that the value of h enters the calculation only through evaluation of the likelihood function, equation (12). This means that the value of h in the physical dimensions has no effect on the results, because the likelihood function only makes use of the photometric dimensions. In the photometric dimensions, h enters only in quadrature sum with the uncertainties σ , and thus the value of h does not affect the results as long as h is significantly smaller than σ in all dimensions. Our choices satisfy this condition, and in general, the condition can always be satisfied as long as the library is large enough to allow a choice of h satisfying this condition. For a more detailed exploration of values of h and the density of sampling points in the library, we refer readers to Krumholz et al. (2015a).

3.3 Analysis of the mock catalogues

We first focus on the Powerlaw, Truncated, and Mdd catalogues, which have uniform tracks, completeness, and errors, and where the only differences are between the parameters describing the cluster distribution; we discuss the remaining cases in the next section. For each mock catalogue, we consider proposed distributions of cluster mass and age following the functional forms given by equations (18) and (22); the free parameters are α_M , M_{break} , and either α_T and T_{mid} (for equation 18) or γ_{mdd} , and $T_{\text{mdd, min}}$ (for equation 22). We leave M_{min} and T_{sf} fixed as above. Our priors on α_M , α_T , and γ_{mdd} are flat. For $\log M_{\text{break}}$, our priors are restricted by the range of masses sampled in our library, and thus we adopt a prior that is flat in $\log(M_{\text{break}}/M_{\odot})$ from 2 to 7. For the same reason we adopt priors on $\log(T_{\text{mid}}/\text{yr})$, and $\log(T_{\text{mdd, min}}/\text{yr})$ that are flat from 5 to 10.17.

Since we cannot assume that we know the functional form of the dust extinction, we choose to parametrize it with a simple piecewise-linear form over the range 0–3 mag in our library. Specifically, we define $A_{V,i} = i\Delta A_V$ mag for $i = 0, \dots, N$, where $\Delta A_V = 3/N$ mag, and our linear fit breaks the range from 0 to 3 mag into N intervals. For an extinction A_V in the range $[A_{V,i}, A_{V,i+1})$, we set $p(A_V) = p_{A_{V,i}} + (p_{A_{V,i+1}} - p_{A_{V,i}})(A_V - A_{V,i})/\Delta A_V$. The values of $p_{A_{V,i}}$ for $i = 0, \dots, N - 1$, representing the values of the extinction PDF at points $A_{V,i}$, are free parameters to be fit, while the value of $p_{A_{V,N}}$ is fixed by the requirement that $\int p(A_V) dA_V = 1$. In the experiments, we present here we adopt $N = 6$, corresponding to breaking the extinction PDF into bins 0.5 mag wide, but our code leaves this

as a free parameter to be set at run time. We adopt priors that are flat in $\log p_{A_{V,i}}$ for $i = 0, \dots, N - 1$, subject to the constraint that $p_{A_{V,N}} > 0$.

We carry out the optimization of the parameters for both the mid and mdd cases using 100 walkers and 500 iterations; visual inspection shows that the distribution of walkers stabilizes after ~ 150 iterations, so we discard the first 200 iterations and derive the posterior PDFs from the remainder. For our sample catalogues, each MCMC calculation requires ≈ 12 h on a workstation. To decide whether the mid or mdd model provides a better fit to each data set, we compute the Akaike information criterion (AIC) for each model (e.g. Sharma 2017). Specifically, for the mid and mdd cases, we find the largest value of the likelihood function $p(\{L_i\}, N_{\text{obs}} | \theta, N_{\text{ex}}, \{\sigma_i\})$ returned by any of the MCMC sample points, which we denote $\hat{\mathcal{L}}$, and compute

$$\text{AIC}_{(\text{mid}, \text{mdd})} = 2k - 2 \ln \hat{\mathcal{L}}_{(\text{mid}, \text{mdd})}, \quad (32)$$

where $k = 11$ is the number of parameters for both the mid and mdd models: six parameters to describe the dust extinction distribution, four to describe the joint mass–time distribution, and 1 to describe the number of clusters N_{ex} . The corresponding Akaike weight for the mid-model,

$$w(\text{mid}) = \frac{e^{-\Delta_{\text{mid}}/2}}{e^{-\Delta_{\text{mid}}/2} + e^{-\Delta_{\text{mdd}}/2}} \quad (33)$$

$$\Delta_{(\text{mid}, \text{mdd})} = \text{AIC}_{(\text{mid}, \text{mdd})} - \min(\text{AIC}_{\text{mid}}, \text{AIC}_{\text{mdd}}), \quad (34)$$

gives the probability that the mid-model is the better fit to the data. Note that this method automatically marginalizes over the unknown dust distribution.

We report values of $w(\text{mid})$ (and the analogously determined $w(\text{mdd}) = 1 - w(\text{mid})$), and posteriors on all parameters, in Table 1. We also show corner plots for the highest weight models for in Figs 4, 5, and 6, for the Powerlaw, Truncated, and MDD catalogues, respectively. Examining the plots and the table, we can draw a number of conclusions. First, the method does an extremely good job at distinguishing whether mid or mdd is a better fit to the data. The Akaike weights are unambiguous. Second, the method recovers the input parameters extremely accurately. The recovered mass function slopes α_M are accurate to better than 0.1 in all cases, and the power-law indices describing the age distribution (α_T for the mass-independent cases, γ_{mdd} for the mass-dependent ones) are recovered with similar accuracy. We also recover the locations of breaks in the mass or age distributions with accuracies of 0.1–0.3 dex, with the sole exception of the Powerlaw case, where by construction our sample should not be able to constrain M_{break} . Indeed, in this case we find a very broad posterior PDF that rules out a break mass below $\sim 10^{5.5} M_{\odot}$, but otherwise leaves the value unconstrained. In summary, we find that our new method offers excellent performance for data sets of a sample size and at an error level typical of modern observations.

3.4 Sensitivity of the method to errors

We next investigate how robust our method is against various types of error, using the mock catalogues `DoubleErr`, `LibMismatch`, and `CompMismatch`. To remind the reader, each of these mock catalogues has the same physical parameters as `Truncated`, but differs in the errors in some way. `DoubleErr` has photometric errors of 0.2 mag instead of 0.1 mag, `LibMismatch` uses Padova rather than MIST tracks, as well as a starburst attenuation curve

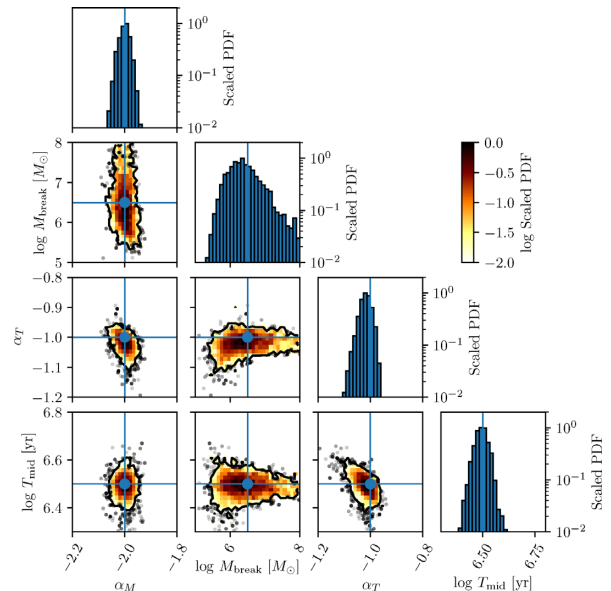


Figure 4. Corner plot showing the 1D and 2D histograms of the posterior PDFs of the parameters α_M , $\log M_{\text{break}}$, α_T , and $\log T_{\text{mid}}$ for the Powerlaw mock catalogue, as determined by MCMC optimization. Nuisance parameters describing the dust extinction distribution have been omitted. Blue histograms show 1D marginal PDFs for each parameter; red-coloured heat maps show 2D probability densities on a logarithmic scale, with all panels normalized to have a maximum of unity. The contour corresponds to a probability density of 10^{-2} on this scale, and scattered points show individual MCMC samples outside this contour. Blue lines and points indicate the true values for the input Powerlaw catalogue.

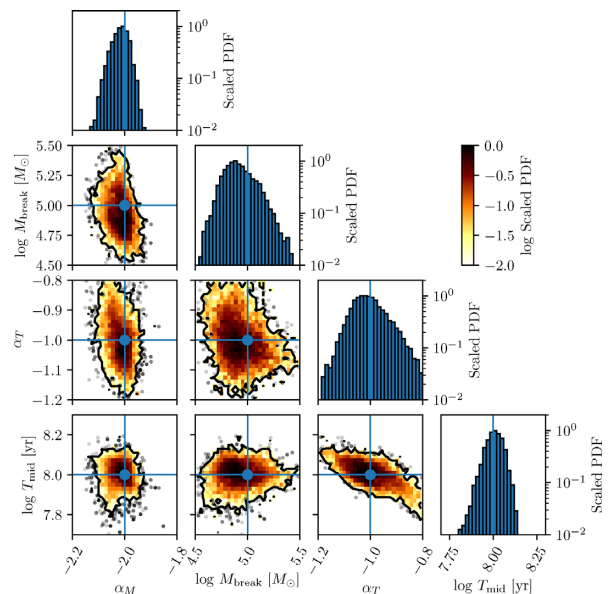


Figure 5. Same as Fig. 4 for the Truncated catalogue.

instead of a Milky Way extinction curve to model the effects of dust. Finally, `CompMismatch` uses a different completeness function than the one assumed in the analysis. We analyse each of these cases using the same procedure as described in Section 3.3; we run the MCMC in for 800 iterations instead of 500, and derive the posterior PDF from the final 300, since we find that it takes slightly

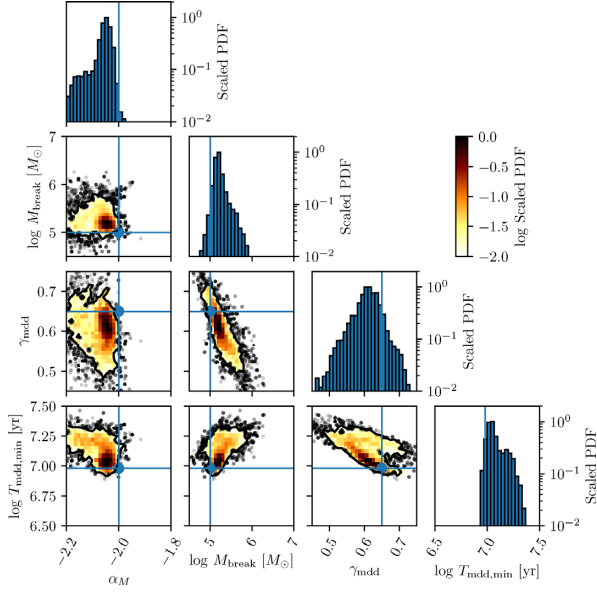


Figure 6. Same as Fig. 4 for the MDD catalogue. However, note that the last two columns plotted (for the quantities γ_{mdd} and $T_{\text{mdd, min}}$) are different than in Fig. 4, since the AIC indicates that the mdd model is a better match to this catalogue than the mid-model.

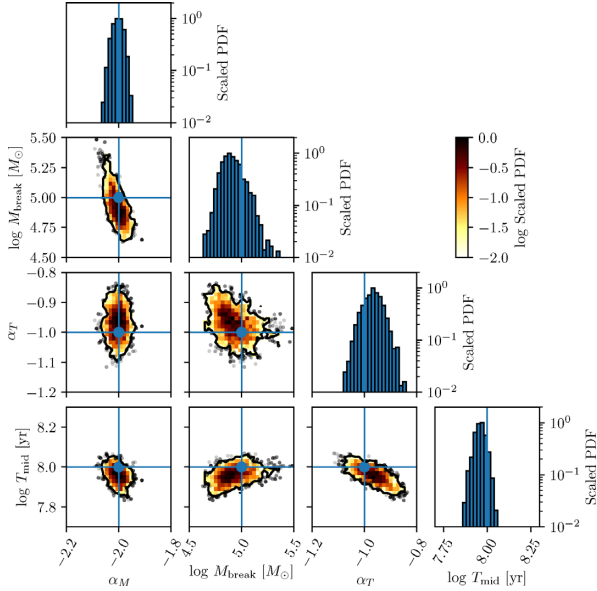


Figure 7. Same as Fig. 5 for the DoubleErr catalogue. Note that the ranges on the axes for this figure are identical to those used in Fig. 5, so the two may be compared directly.

longer for the posterior distributions to stabilize in at least some of these cases.

We report the marginalized posteriors in Table 1, and show corner plots for the posteriors in Figs 7, 8, and 9 for the DoubleErr, LibMismatch, and CompMismatch cases, respectively. As is clear from the table and from comparing these three figures to Fig. 5, in all three cases with increased errors we can still clearly distinguish between mid and mdd, and still clearly identify the truncation in the mass function and the break in the age distribution. Doubling the photometric errors has remarkably little effect on the accuracy of the resulting fits, likely because, given the large number

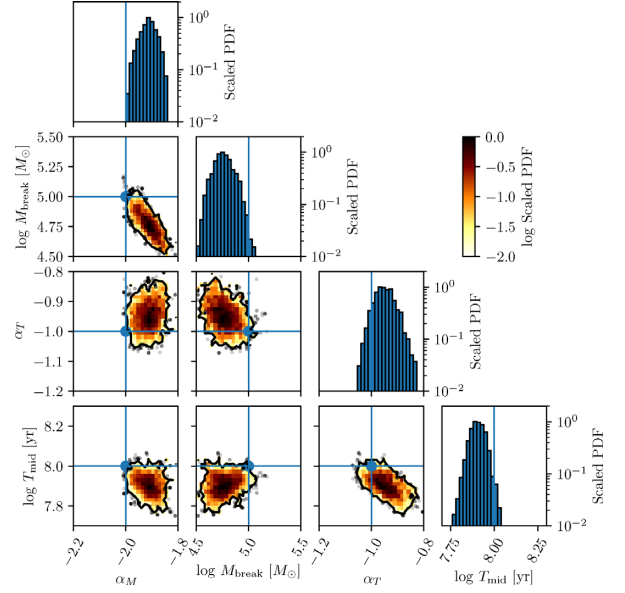


Figure 8. Same as Fig. 5 for the LibMismatch catalogue. Note that the ranges on the axes for this figure are identical to those used in Fig. 5, so the two may be compared directly.

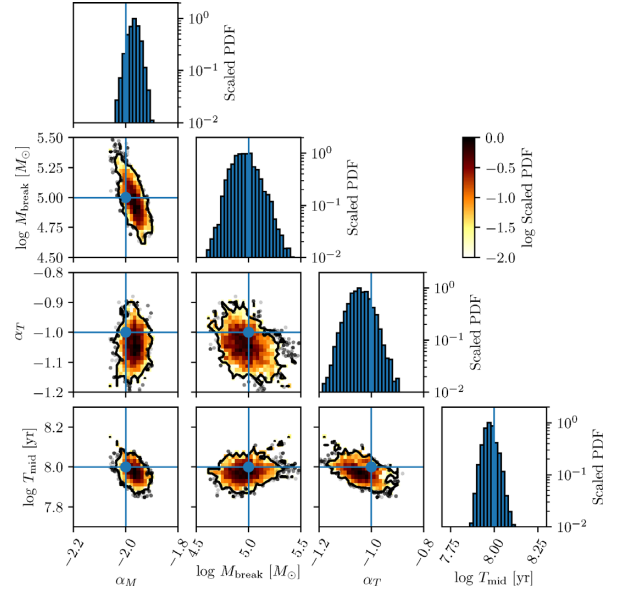


Figure 9. Same as Fig. 5 for the CompMismatch catalogue. Note that the ranges on the axes for this figure are identical to those used in Fig. 5, so the two may be compared directly.

of clusters in the catalogue, the limiting factor in the accuracy of the fits is stochastic sampling and the degeneracies it induces in the tracks, not the accuracy with which individual clusters' photometry can be measured. The main effect of using a library that does not precisely match the data is to induce a systematic shift in the posteriors, while leaving the shape and width of the posterior distribution largely unchanged. The best-fitting mass function and age distribution slopes for LibMismatch are displaced from their true values ($\alpha_M = -2$ and $\alpha_T = -1$) to $\alpha_M = -1.9$ and $\alpha_T = -0.95$, while the truncation mass and break in the age distribution are shifted to $\log(M_{\text{break}}/M_\odot) = 4.76$ (true value 5.0) and $\log(T_{\text{mid}}/\text{yr}) = 7.9$ (true value 8.0). Using an incorrect completeness function (as in

CompMismatch) has a similar effect, but smaller in magnitude. This may well be a function of our parametrization: the effect of using an incorrect estimate of our completeness is to induce artificial wiggles in the completeness-corrected luminosity function at low luminosity. However, since we are forcing the functional forms of our physical distributions to be power laws without any such wiggles, the wiggles do not greatly alter the best-fitting power-law slopes. Overall our results suggest that fits to cluster demographics likely to be limited to an accuracy of $\approx \pm 0.1$ in the age and mass function slopes, and a few tenths of a dex in age or mass truncations.

It is important to note that the results of our fit to the LibMismatch catalogue differ from both the true values and the estimates we find for the Truncated case by amounts that range from one to a few standard deviations. This implies that our method is sufficiently sensitive that, for our sample size of a few thousand clusters, the accuracy of the results is ultimately limited by the quality of the underlying physical models, rather than by the data quality or statistics. We emphasize that this is a highly non-trivial statement. Even for samples of this size, analysis using conventional methods generally yields results where the choice of tracks or libraries does not change the best-fitting values by more than the statistical error bars (e.g. Adamo et al. 2017), and even using a method that fully accounts for stochastic sampling and returns the full posterior distribution, estimates of for the mass and age of individual clusters are relatively insensitive to the choice of underlying stellar model (Krumholz et al. 2015a). It is only because our method is capable of exploiting all the information present in a realistically sized sample of clusters that we have reached a point where we are limited by systematic rather than statistical uncertainties.

4 COMPARISON TO CONVENTIONAL METHODS

We have now demonstrated that our new method is capable of recovering the parameters describing a cluster population with high accuracy, and that it is robust against plausible systematic and random errors. We now turn to the question of how well our method performs compared to more conventional approaches that do not involve full forward modelling. We therefore re-analyse the Powerlaw and Truncated catalogues using a conventional method. There are a wide range of such methods, which differ from one another significantly in their details. We have chosen a methodology based on those used in a number of recent publications, and that should be broadly representative of the strengths and weaknesses of current techniques. As we will see below, the conventional method performs far worse when operating on the same data.

4.1 Fitting individual clusters

In any conventional method to derive cluster population properties, the first step is to derive the properties of the individual star clusters from their photometry assuming that the IMF is fully sampled, so that the relationship between photometry and physical properties is deterministic and we can assign a single best-fitting mass and age. To facilitate this we generate a grid of SLUG simulations of star clusters at a range of ages and extinctions with a fully sampled IMF, i.e. with no stochasticity. We then find the best-fitting age and extinction for each catalogue cluster by calculating the minimum χ^2 between the colours of the model grid and the colours in the catalogue, and find the best-fitting age by scaling the absolute magnitudes of the cluster to those of the model grid. We give a full explanation of the procedure in Appendix B.

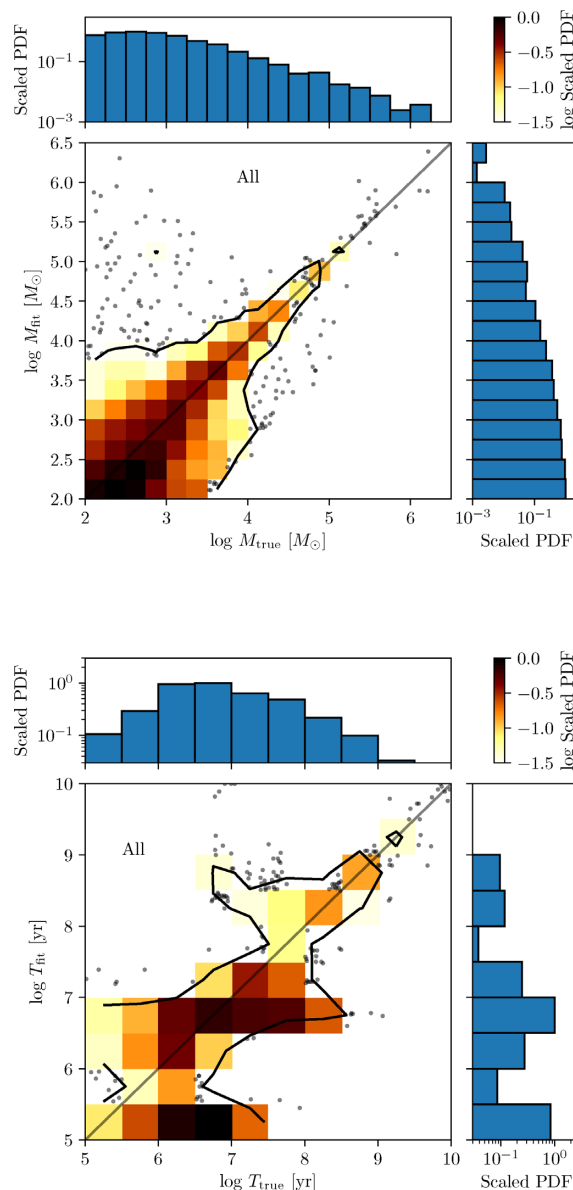


Figure 10. Comparison between true and best-fitting cluster masses and ages, for all clusters in the Powerlaw mock catalogue for which the reduced χ^2 value of the fit is < 5 . In the top panel, the heat map shows the density of clusters, measured in bins 0.25 dex wide, in the plane of $\log M_{\text{true}}$ versus $\log M_{\text{fit}}$, where the former is the true cluster mass and the latter is the best fit determined from the procedure outlined in the main text. The black contour marks a density of $10^{-1.5}$ relative to the maximum, points show individual clusters in low-density regions, and the grey line shows the one-to-one line where a perfect fit would lie. The flanking histograms show the distributions of true and best-fitting mass. The bottom panel shows the analogous comparison between actual and best-fitting clusters ages, with bins 0.5 dex wide for the histograms.

To test how well this method performs, in Fig. 10 we compare the best-fitting and true masses and ages of the clusters in the Powerlaw mock catalogue; results for other catalogues are qualitatively similar, and so for now we simply focus on the Powerlaw case. When making this figure, we only include clusters for which the reduced χ^2 value ($= \chi^2/2$, since we have five photometric bands and three model parameters) of the fit is < 5 on the grounds that larger values indicate a poor fit. This cut removes just under 20 per cent

of the clusters in the mock catalogue, and the results are not very sensitive to the exact threshold used to remove bad fits. We see that, in agreement with our expectations, the masses and ages recovered by the fitting procedure are for the most part accurate to within a factor of a few for clusters whose true mass is above $\sim 10^4 M_\odot$, or whose true age is above 10^8 yr. In this mass and age range (which are largely overlapping in the observed catalogue, since only a massive cluster will be bright enough to be detected at ages above $\sim 10^8$ yr), stochastic sampling of the IMF has minor effects, so full sampling is a reasonable approach. At lower masses and ages, the scatter is substantially larger, but the majority of the data still cluster around the one-to-one line.

There are also a few extreme outliers very far from the one-to-one line. These typically have a true mass is below $\sim 1000 M_\odot$ and age $\sim 10^7$ yr, but the best-fitting mass is $> 10^6 M_\odot$, and the best-fitting age above $\sim 10^9$ yr. This phenomenon occurs when stochastic sampling of the IMF in a low-mass, middle-aged cluster whose light is dominated by a small number of evolved stars happens to produce colours quite similar those of a much more massive, older stellar population whose light is dominated by a large number of low-mass stars. Because the conventional method does not include stochastic sampling, no models in the right age range match the colours as well as the older model, and thus the age and mass are catastrophically misestimated.

4.2 Mass and age cuts

To handle the problem of incompleteness, the next step in a conventional analysis is to cut the sample based on mass and age, restricting to ranges that are thought to be reasonably complete. We therefore next remove from the mock catalogue any clusters whose best-fitting age is $> 10^{8.5}$ yr or whose best-fitting mass is $< 10^{3.75} M_\odot$; these cuts are quite comparable to those normally used for star cluster analysis (e.g. Adamo et al. 2017; Chandar et al. 2017; Johnson et al. 2017), and roughly corresponds to the region in age–mass space where almost all clusters will lie at $V \leq -5$ mag, and thus in the range where our catalogue is complete. This combined with the cut on quality of fit reduces the `PowerLaw` catalogue to 366 clusters (from the original 5629), i.e. only 6.5 per cent of the original sample passes all the cuts. The reason that the mass–age cut so severely reduces the size of the available data is that, although any cluster with mass above $10^{3.75} M_\odot$ and age below $10^{8.5}$ yr is very likely to have $V < -5$ mag and thus be bright enough to make it into the observed catalogue, the converse is not true, i.e. having $V < -5$ mag in no way guarantees that the mass is above $10^{3.75} M_\odot$ and age below $10^{8.5}$ yr. Indeed, because low-mass clusters are intrinsically much more common than more massive ones, the majority of the clusters bright enough to be observed have best-fitting masses that place them below our mass limit. They are bright enough to be observed because they are much younger than $10^{8.5}$ yr. We could retain more of these clusters by using a lower mass cut, but only at the price of having an even more severe age cut in order to ensure completeness. The trade-off between age and mass cuts that we have made is comparable to the ones used by previous authors.

We plot the distribution of mass and age for the remaining clusters for the `PowerLaw` catalogue in Fig. 11. One particularly noticeable effect in this plot is that the lowest bins of fitted mass contain a very significant number of interlopers whose true mass is smaller, visible as the large tail of clusters extending into the grey region in the upper panel of Fig. 11. This is not a small effect: the true number of clusters in the mass range $\log M/M_\odot = 3.75\text{--}4.0$, only considering clusters that make it into the fitted sample (i.e. excluding those that have been dropped because their best-fitting mass is too small, or because

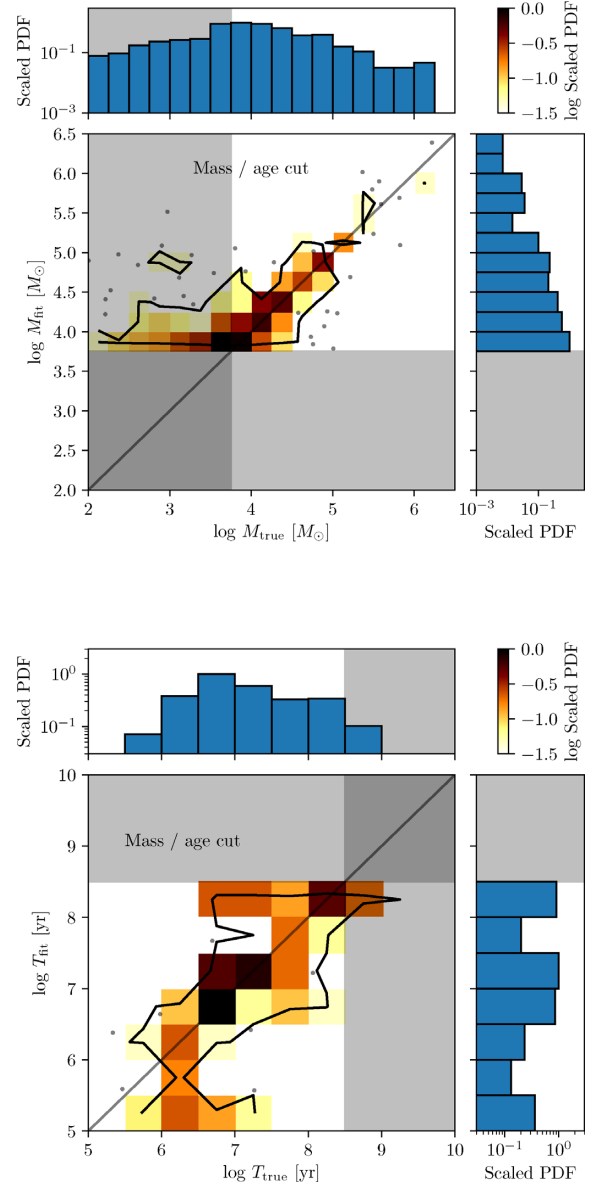


Figure 11. Same as Fig. 10, but now including only those clusters with a best-fitting mass above $\log (M/M_\odot) = 10^{3.75}$ and best-fitting age below $\log (T/\text{yr}) = 8.5$. The grey regions show the ranges of mass and age removed by the cut.

they do not have good χ^2 values) is 64, while the fitting procedure produces 141, i.e. more than double the correct count. This bias has the effect of slightly flattening the best-fitting mass distribution. Its origin lies in the fact that low-mass clusters are intrinsically more common than high-mass ones, and thus there are many more low-mass clusters scattering to high-mass fits than high-mass clusters scattering to lower masses. The problem is worsened by the use of a mass cut, which creates an asymmetry: clusters whose fitted masses scatter below their true ones are preferentially removed from the sample, while those whose fitted masses scatter above their true ones are preferentially included.

4.3 Fitting the distributions

The final step in a conventional analysis is to fit the remaining data to constrain the properties of the population. Numerous fitting

techniques exist in the literature, and it is not our goal here to perform a comprehensive comparison among them, so we select one representative example: we bin the data uniformly by mass and by age, using the binning shown in Fig. 11, then perform χ^2 fits to the resulting binned distributions using the bin centres as the independent variable and bin number counts with Poisson errors as the dependent variable. A number of recent publications have used this method (e.g. Silva-Villa et al. 2014; Chandar et al. 2015, 2017). For simplicity, we only consider mid for this exercise, since for mdd we would need to fit the joint mass–age distribution, or fit multiple age distributions at constant mass or mass distributions at constant age.

For the mass distribution, we fit to both exponentially truncated and pure power laws (i.e. to functional forms $dN/dM \propto M^{\alpha_M}$ and to $dN/dM \propto M^{\alpha_M} e^{-M/M_{\text{break}}}$). We plot the resulting best fits against the binned data and the true input distributions in Fig. 12. We see that binning in mass does allow one to qualitatively recover roughly the correct mass distributions, but with significant defects compared to the forward-modelling method. First, the truncated-power-law fits return $\alpha_M = -1.82 \pm 0.11$, $\log M_{\text{break}} = 6.47 \pm 0.93$ (1σ error bars) and $\alpha_M = -2.07 \pm 0.10$, $\log M_{\text{break}} = 5.22 \pm 0.2$ for the Powerlaw and Truncated cases, respectively; compared to the results given in Table 1 the central values for α_M are significantly further from the true value ($\alpha_M = -2$), and the error bars are larger by a factor of 3–5; one should also recall that even these error bars are underestimates, since they include only shot noise in the bin counts and not errors in the assigned masses.

Moreover, unlike the forward-modelling method, the binned method does not always successfully distinguish between pure- and truncated-power-law fits. For the Powerlaw catalogue, the reduced χ^2 values for the truncated- and pure-power-law fits are 2.6 and 2.1, respectively, so the pure-power law is a marginally better fit, as it should be since the data really do lack the statistical power to distinguish the two cases. However, for the Truncated catalogue, the truncated- and pure-power-law fits have reduced χ^2 values of 0.32 and 0.95 respectively, so the pure-power-law fit is actually preferred on the grounds that the truncated-power law is overfitting the data. Indeed, one could have guessed this simply from Fig. 12, since the pure-power-law fit falls within the Poisson error bars for every bin. For this catalogue, our new method successfully recovers an important result that the conventional one misses. The reasons for the new method’s superior performance are obvious: the conventional χ^2 method introduces a significant source of error by assigning each cluster a single best-fitting mass, discards more than 90 per cent of the sample due to the mass and age cuts it requires, then and discards even more information by binning the data that remain.

We show the analogous result for the binned age distributions in Fig. 13; the fits shown are to a distribution $dN/dT \propto T$ for $T < T_{\text{mid}}$ and $dN/dT \propto T^{\alpha_T}$ for $T > T_{\text{mid}}$, with T_{mid} and α_T as our fit variables. Again, we see that the results of χ^2 fitting have a rough qualitative resemblance to the input distributions, but that the properties recovered are considerably less accurate. For the Powerlaw catalogue, the best-fitting values are $\alpha_T = -1.34 \pm 0.39$ and $\log T_{\text{mid}} = 6.44 \pm 0.37$, while the true inputs are $\alpha_T = -1$, $\log T_{\text{mid}} = 6.5$; the slope is poorly determined in part due to the undercount in the $10^{7.5}–10^8$ yr bin (also visible in Fig. 11), which drags the best-fitting slope downward and results in a poor fit quality overall (reduced χ^2 of 40). This is an artefact of trajectory of a fully sampled cluster through colour space in this age range, which makes it easy to mistake an older cluster for a younger one; such artefacts are known peril in χ^2 fitting methods, as illustrated for example by

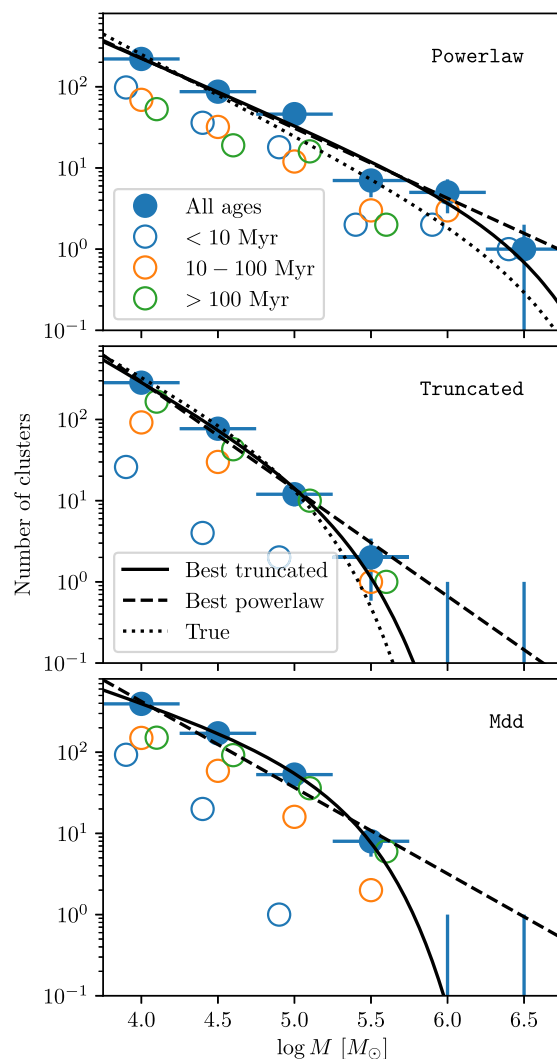


Figure 12. Binned mass distributions for the Powerlaw, Truncated, and Mdd catalogues (top to bottom, as indicated). Solid blue points indicate the number counts for clusters of all ages below $10^{8.5}$ yr; horizontal error bars show the width of each bin, and vertical error bars show Poisson errors on the number count. Open circles show the number counts for the same bins for clusters ages <10 , $10–100$, and >100 Myr, as indicated; the horizontal positions of these points have been perturbed slightly, and error bars on them suppressed, to reduce confusion. The black solid and dashed lines show the best truncated- and pure-power-law fits to the solid points (see the main text), while the black dotted line shows the true mass distribution, normalized to match the data in the lowest mass bin; we omit the true mass distribution for the Mdd case because it is not time-independent.

the age-stripping visible in fig. 10 of Chandar et al. (2010) or fig. 14 of Adamo et al. (2017). Our new method avoids this problem by retaining the full posterior PDF rather than simply using the best fit, and Table 1 shows that our method recovers α_T with uncertainties an order of magnitude smaller: ≈ 0.02 rather than ≈ 0.2 .

For the Truncated catalogue, the conventional χ^2 method returns a best-fitting slope of $\alpha_T = -0.31 \pm 0.29$ and is unable to determine a meaningful value of $\log T_{\text{mid}}$ (formally the uncertainty in this parameter diverges), whereas the true inputs are $\alpha_T = -1.0$ and $\log T_{\text{mid}} = 8$. It is not surprising that the method fails to recover $\log T_{\text{mid}}$, since the break in the age distribution in the input data lies very close to the age of $10^{8.5}$ yr above which we had to discard data due to incompleteness in the conventional method. The recovered

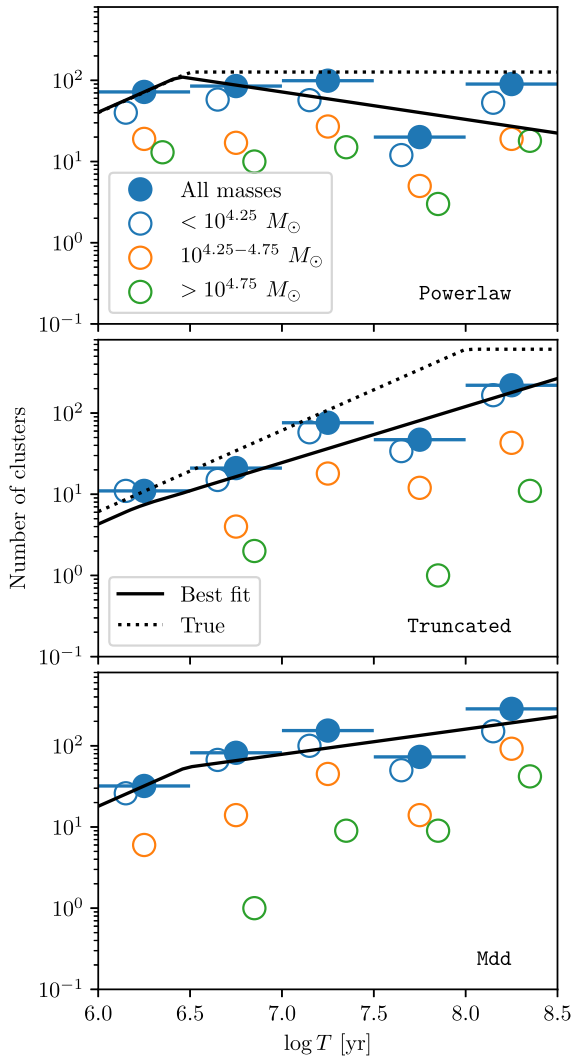


Figure 13. Same as Fig. 12, but showing the binned age distributions instead of mass distributions.

slope $\alpha_T = -0.31 \pm 0.29$ is essentially within 1σ of the correct slope (which is 0) for the age range below 10^8 yr covered by the remaining data. However, this only serves to illustrate a further advantage of our new method that it does not require such severe cuts on the data to cope with incompleteness.

In summary, we find that the new method we have introduced outperforms the traditional one across all dimensions. It returns more accurate parameter estimates with much smaller uncertainties, and across a wider domain in cluster mass and age, and it successfully distinguishes between truncated and non-truncated mass distributions in cases where conventional methods cannot.

5 SUMMARY AND CONCLUSION

We introduce a new forward-modelling method to determine the demographics of a population of star clusters from unresolved photometry. The basic idea of the method is to consider a proposed distribution of cluster masses and ages, apply kernel density estimation to a pre-computed library of models weighted by the observational completeness to predict the observed luminosity distribution, and adjust the proposed mass and age distributions to optimize agreement between the observed and predicted luminosity distributions.

Our method does not require that the data be binned, allows analysis of heterogeneous data sets where not all regions have been observed to the same depth or with the same filters, does not require that data be limited to a particular range of mass or age, and naturally accounts for the effects of stochastic sampling of the stellar IMF. The method is computationally efficient enough that a catalogue of star clusters comparable to those obtained via recent *HST* campaigns can be analysed in approximately half a day of computing time on a workstation.

We test our method on synthetic data sets and show that we are able to recover correct demographics for the underlying populations with very high accuracy; typical statistical uncertainties on slopes of the mass and age distributions are only ~ 0.01 dex. Our method distinguishes between alternative models for star cluster age distributions, including truncated versus power-law mass functions, and mid versus mdd, with very high confidence. The performance of our method compares very favourably with that of traditional χ^2 fitting methods, whereby one obtains a best-fitting mass and age for each individual star cluster, and then fits the demographics of the population. We show that this method produces uncertainties on the slopes of mass and age distributions that are as much as an order of magnitude larger than our forward-modelling technique, and often lacks the statistical power to distinguish between alternative physical scenarios for star cluster demographics. By repeating our analysis with different set of stellar tracks and dust distributions, we find that systematic errors in the slopes of mass and age distributions are ≈ 0.1 dex, significantly larger than the statistical uncertainties our method produces, and thus our method is as accurate as possible given our current knowledge of stellar evolution and interstellar dust physics. Indeed, it is not beyond the realm of possibility that application of this method with different libraries of models could be used to diagnose which stellar evolution models are best able to match reality.

We implement our method as part of the SLUG suite of stochastic stellar population and statistics tools. The software for the method and for all the tests presented in this paper, along with the mock catalogues on which we performed our tests and the full outputs of our MCMC analysis, are available from the SLUG website at <http://www.slugsp.com/cluster-population-pipeline>.

In future work, we will apply the method presented here to the large sample of star clusters produced by the LEGUS survey (Calzetti et al. 2015; Adamo et al. 2017). In addition to providing an analysis of star cluster demographics with considerably greater accuracy than any previous method, this will enable us to obtain much better estimates for the properties of individual clusters. Krumholz et al. (2015a) showed that the largest uncertainty in the posterior probability distributions for the masses and ages for individual star clusters is the prior distribution, i.e. one's starting estimate of the frequency with which particular masses and ages arise in the population. By combining the methods outlined in Krumholz et al. (2015a) with the population demographics, we determine from the method implemented here, we will be able to mitigate this uncertainty considerably, thereby improving our estimates cluster by cluster as well for the population as a whole.

ACKNOWLEDGEMENTS

MRK acknowledges support from the Australian Research Council's *Discovery Projects* funding scheme (project DP160100695) and from NASA through grants from the Space Telescope Science Institute research (program numbers 13256 and 13364). This project was undertaken with the assistance of resources and services from

the National Computational Infrastructure (NCI), which is supported by the Australian Government. MF acknowledges support by the Science and Technology Facilities Council, grant number ST/P000541/1. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 757535).

REFERENCES

- Adamo A. et al., 2017, *ApJ*, 841, 131
- Adamo A., Östlin G., Zackrisson E., Hayes M., Cumming R. J., Micheva G., 2010, *MNRAS*, 407, 870
- Adamo A., Kruijssen J. M. D., Bastian N., Silva-Villa E., Ryon J., 2015, *MNRAS*, 452, 246
- Anders P., Bissantz N., Fritze-v. Alvensleben U., de Grijs R., 2004, *MNRAS*, 347, 196
- Avni Y., 1976, *ApJ*, 210, 642
- Bastian N. et al., 2012, *MNRAS*, 419, 2606
- Baumgardt H., Kroupa P., 2007, *MNRAS*, 380, 1589
- Borisova J. et al., 2011, *A&A*, 532, A131
- Calzetti D. et al., 2015, *AJ*, 149, 51
- Cerviño M., 2013, *New Astron. Rev.*, 57, 123
- Cerviño M., Luridiana V., 2004, *A&A*, 413, 145
- Cerviño M., Luridiana V., 2006, *A&A*, 451, 475
- Chabrier G., 2005, in Corbelli E., Palla F., Zinnecker H., eds, *Astrophysics and Space Science Library*, Vol. 327, *The Initial Mass Function 50 Years Later*. Springer, Dordrecht. 41– +
- Chandar R. et al., 2010, *ApJ*, 719, 966
- Chandar R., Fall S. M., Whitmore B. C., 2010, *ApJ*, 711, 1263
- Chandar R., Whitmore B. C., Calzetti D., Di Nino D., Kennicutt R. C., Regan M., Schinnerer E., 2011, *ApJ*, 727, 88
- Chandar R., Fall S. M., Whitmore B. C., 2015, *ApJ*, 810, 1
- Chandar R., Fall S. M., Whitmore B. C., Mulia A. J., 2017, *ApJ*, 849, 128
- Choi J., Dotter A., Conroy C., Cantiello M., Paxton B., Johnson B. D., 2016, *ApJ*, 823, 102
- Conroy C., van Dokkum P. G., 2016, *ApJ*, 827, 9
- de Meulenaer P., Narbutis D., Mineikis T., Vasevicius V., 2015, *A&A*, 574, A66
- da Silva R. L., Fumagalli M., Krumholz M., 2012, *ApJ*, 745, 145
- da Silva R. L., Fumagalli M., Krumholz M. R., 2014, *MNRAS*, 444, 3275
- Dolphin A. E., 2002, *MNRAS*, 332, 91
- Dotter A., 2016, *ApJS*, 222, 8
- Elmegreen B. G., Hunter D. A., 2010, *ApJ*, 712, 604
- Fall S. M., Chandar R., 2012, *ApJ*, 752, 96
- Fall S. M., Zhang Q., 2001, *ApJ*, 561, 751
- Fall S. M., Krumholz M. R., Matzner C. D., 2010, *ApJ*, 710, L142
- Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, *PASP*, 125, 306
- Fouesneau M. et al., 2014, *ApJ*, 786, 117
- Fouesneau M., Lançon A., Chandar R., Whitmore B. C., 2012, *ApJ*, 750, 60
- Gieles M., 2009, *MNRAS*, 394, 2113
- Gieles M., Lamers H. J. G. L. M., Portegies Zwart S. F., 2007, *ApJ*, 668, 268
- Girardi L., Bressan A., Bertelli G., Chiosi C., 2000, *A&AS*, 141, 371
- Goddard Q. E., Bastian N., Kennicutt R. C., 2010, *MNRAS*, 405, 857
- Harris J., Zaritsky D., 2009, *AJ*, 138, 1243
- Hunter D. A., Elmegreen B. G., Dupuy T. J., Mortonson M., 2003, *AJ*, 126, 1836
- Johnson L. C. et al., 2016, *ApJ*, 827, 33
- Johnson L. C. et al., 2017, *ApJ*, 839, 78
- Kreckel K. et al., 2013, *ApJ*, 771, 62
- Kruijssen J. M. D., 2009, *A&A*, 507, 1409
- Kruijssen J. M. D., 2012, *MNRAS*, 426, 3008
- Kruijssen J. M. D., Pelupessy F. I., Lamers H. J. G. L. M., Portegies Zwart S. F., Bastian N., Icke V., 2012, *MNRAS*, 421, 1927
- Krumholz M. R. et al., 2014, in Beuther H., Klessen R. S., Dullemond C. P., Henning Th., eds, *Protostars and Planets VI*. Univ. Arizona Press, Tucson, AZ, p. 243
- Krumholz M. R. et al., 2015a, *ApJ*, 812, 147
- Krumholz M. R., 2014, *Phys. Rep.*, 539, 49
- Krumholz M. R., Matzner C. D., 2009, *ApJ*, 703, 1352
- Krumholz M. R., Fumagalli M., da Silva R. L., Rendahl T., Parra J., 2015b, *MNRAS*, 452, 1447
- Lada C. J., Lada E. A., 2003, *ARA&A*, 41, 57
- Lamers H. J. G. L. M., 2009, *Ap&SS*, 324, 183
- Lamers H. J. G. L. M., Gieles M., Bastian N., Baumgardt H., Kharchenko N. V., Portegies Zwart S., 2005, *A&A*, 441, 117
- Larsen S. S., 2002, *AJ*, 124, 1393
- Leitherer C. et al., 1999, *ApJS*, 123, 3
- Longmore S. N. et al., 2014, in Beuther H., Klessen R. S., Dullemond C. P., Henning Th., eds, *Protostars and Planets VI*. Univ. Arizona Press, Tucson, AZ, p. 291
- Maschberger T., Kroupa P., 2009, *MNRAS*, 395, 931
- Messa M. et al., 2018, *MNRAS*, 473, 996
- Murray N., Quataert E., Thompson T. A., 2010, *ApJ*, 709, 191
- Parmentier G., Goodwin S. P., Kroupa P., Baumgardt H., 2008, *ApJ*, 678, 347
- Paxton B. et al., 2013, *ApJS*, 208, 4
- Paxton B. et al., 2015, *ApJS*, 220, 15
- Paxton B., Bildsten L., Dotter A., Herwig F., Lesaffre P., Timmes F., 2011, *ApJS*, 192, 3
- Popescu B., Hanson M. M., 2009, *AJ*, 138, 1724
- Popescu B., Hanson M. M., 2010a, *ApJ*, 724, 296
- Popescu B., Hanson M. M., 2010b, *ApJ*, 713, L21
- Popescu B., Hanson M. M., Elmegreen B. G., 2012, *ApJ*, 751, 122
- Rafelski M., Zaritsky D., 2005, *AJ*, 129, 2701
- Sharma S., 2017, *ARA&A*, 55, 213
- Silva-Villa E., Adamo A., Bastian N., Fouesneau M., Zackrisson E., 2014, *MNRAS*, 440, L116
- Vázquez G. A., Leitherer C., 2005, *ApJ*, 621, 695
- Weisz D. R. et al., 2013, *ApJ*, 762, 123
- Williams J. P., McKee C. F., 1997, *ApJ*, 476, 166
- Zackrisson E., Rydberg C.-E., Schaerer D., Östlin G., Tuli M., 2011, *ApJ*, 740, 13
- Zhang Q., Hunter T. R., Sridharan T. K., Cesaroni R., 1999, *ApJ*, 527, L117

APPENDIX A: ALGORITHM FOR FAST EVALUATION OF THE LIKELIHOOD FUNCTION

Here, we describe the algorithm we use for fast evaluation of the sum

$$s = \sum_{j=1}^{N_{\text{lib}}} w_j(\theta) \mathcal{N}(\mathbf{L}_i - \mathbf{L}_j | \mathbf{h}') \quad (\text{A1})$$

in the likelihood function (equation 12). The algorithm proceeds in three steps. First, we arrange the N_{lib} clusters in the library in a KD-tree of N_F dimensions based on their luminosities \mathbf{L}_j . (Recall that \mathbf{L}_j has N_F dimensions.) We maintain separate KD-trees for each set of filters that are present in the observed sample. For each node k in the KD-tree, we compute the bounding box, i.e. the smallest N_F -dimensional rectangular prism aligned with the cardinal axes that contains \mathbf{L}_j for each cluster luminosity \mathbf{L}_j in the node. We also record the weight $w_j(\theta)$ of each cluster, and the summed weight w_k of all the clusters contained in each node. This is an order $N_{\text{lib}} \ln N_{\text{lib}}$ operation, but need only be done once at the start of the calculation.

The second step in the algorithm is that, when we wish to change model parameters θ , we recalculate the weights $w_j(\theta)$ of each individual cluster, and the summed weights w_k of all nodes in the tree. This is an order N_{lib} operation, but need only be done once

for each set of trial parameters θ , not once per cluster, and can be parallelized trivially. On a workstation-level machine, using library of 10^7 clusters, this step requires a few tenths of a second.

The third step is evaluation of the sum given by equation (A1) for each observed cluster luminosity L_i . We carry out this step via a divide and conquer algorithm of order $\ln N_{\text{lib}}$:

(i) Let `nodes` be a list of nodes in the tree; for each node k in the list, we record an estimate s_k of its contribution to the sum and an upper bound Δs_k on the error in this estimate, computed whenever a node is added to the list in step (iii). That is, for any node k , the contribution of the clusters within that node to s is strictly bounded between $s_k - \Delta s_k$ and $s_k + \Delta s_k$.

(ii) Evaluate the current estimate of the sum $s = \sum_k s_k$ and the upper bound on the error $\Delta s = \sum_k \Delta s_k$, where the sums run over all nodes in the list `nodes`. If $\Delta s/s$ is smaller than some specified tolerance, stop iterating and return s . If not identify the node k with the largest value of Δs_k .

(iii) Remove the node with the largest Δs_k from `nodes`, and add its left and right children, which we denote ℓ and r , to `nodes`. Compute s_ℓ and Δs_ℓ as follows (and similarly for s_r and Δs_r):

(a) If node ℓ is a leaf (i.e. it has no children), set $s_\ell = \sum w(\theta) \mathcal{N}(L_i - L_j | h')$, where the sum runs over all clusters in the leaf. Set $\Delta s_\ell = 0$. That is, if the node is a leaf, directly evaluate the contribution to the sum of all clusters in that leaf, and set the maximum possible error to zero.

(b) If node ℓ is not a leaf, find the vectors ΔL_{near} and ΔL_{far} between L_i and the nearest and farthest points in the bounding box of node ℓ , where distances are measured in units of h' . That is, $\Delta L_{\text{near}} = L_i - L_b$ for the point L_b within the bounding box of node ℓ that minimizes $d = \sum_{n=1}^{N_F} (L_{i,n} - L_{b,n})^2 / h_n'^2$; similarly, ΔL_{far} is computed for the point L_b that maximizes d . Note that if L_i is inside the bounding box, then $\Delta L_{\text{near}} = \mathbf{0}$. Set

$$s_\ell = w_\ell \frac{\mathcal{N}(\Delta L_{\text{near}} | h') + \mathcal{N}(\Delta L_{\text{far}} | h')}{2} \quad (\text{A2})$$

$$\Delta s_\ell = w_\ell \frac{\mathcal{N}(\Delta L_{\text{near}} | h') - \mathcal{N}(\Delta L_{\text{far}} | h')}{2}. \quad (\text{A3})$$

Intuitively, this amounts to finding the maximum (minimum) possible contribution of the clusters in node ℓ to the sum, which would occur if all the clusters in that node were at the nearest (farthest) point within the bounding box. We then set the central estimate of the contribution of that node to the sum to the average of the minimum and maximum possible contributions, and the error to half the distance between them.

(iv) Go back to step (ii).

The algorithm begins by adding the root node of the KD-tree to the list `nodes`, with s_k and Δs_k for it evaluated as in step (iii). We then iterate until the convergence condition in step (ii) is satisfied. The algorithm is efficient because (1) it quickly eliminates parts of the KD-tree that are far from L_i , and thus make small contributions to s and Δs , and (2) it removes the need to examine individual clusters whose separation in luminosity space is $\ll h'$, since these will be grouped into the same node, and nodes whose bounding boxes are $\ll h'$ in size will have $\Delta s_k \ll s_k$. In practice, we find that evaluation of s to 1 per cent accuracy requires examining hundreds to thousands of nodes, depending on the number of filters N_F , the

bandwidth h' , the number of points N_{lib} in the library, and the density of library points in the vicinity of L_i . In our tests with five filters, a bandwidth of 0.05–0.1 mag, and 10^7 library points, typical evaluation times were $\sim 100 \mu\text{s}$ per cluster on a workstation.

APPENDIX B: METHOD FOR χ^2 FITTING

Here, we explain in detail the method we use to fit the mass, age, and extinction for each catalogue cluster for our conventional analysis. The first step is to generate a grid of SLUG models with a fully sampled IMF. We use a fixed cluster mass of $10^6 M_\odot$, since in the non-stochastic case the photometry can be rescaled trivially to any chosen mass; specifically, if our SLUG run with a mass of $10^6 M_\odot$ produces a magnitude $m_{i,\text{model}}$ in filter i , then the corresponding prediction for a cluster of mass M is simply $m_i = m_{i,\text{model}} - 2.5(\log M - 6)$. We output model predictions $m_{i,\text{model}}$ at a set of times from 10^5 to 10^{10} yr, with outputs spaced at 0.01 dex. We repeat this calculation at dust extinctions from $A_V = 0$ to 3 mag in steps of 0.05 mag. The result is a grid of 30 561 models, each with predicted photometry in the same five bands as our mock catalogues as a function of cluster age and extinction.

To fit obtain the best-fitting mass, age, and extinction, we proceed as follows. First, for each catalogue cluster, we find a best-fitting mass at each point in our age-extinction grid by finding the value M_{fit} that minimizes

$$\chi^2 = \sum_i \frac{[m_{i,\text{cat}} - m_{i,\text{model}} + 2.5(\log M_{\text{fit}} - 6)]^2}{\sigma_i^2}, \quad (\text{B1})$$

where $m_{i,\text{cat}}$ is the magnitude of the catalogue star cluster in band i , σ_i is the error on this value (0.1 mag for all our mock catalogues), and $m_{i,\text{model}}$ is the magnitude for the model grid point. Note that the required value of M_{fit} can be obtained analytically simply by solving the equation $d\chi^2/d\log M_{\text{fit}} = 0$. We record the corresponding minimum value of χ^2 at each grid point for each cluster. Second, for each cluster in the mock catalogue, we then find the grid point that produces the smallest χ^2 value, which we denote χ_{min}^2 . We assign the age and extinction of that point as the best-fitting values for that cluster, and the corresponding mass recorded for that grid point as its best-fitting mass. Third, to derive the 68 per cent confidence interval, we find all the grid points for which $\chi^2 < \chi_{\text{min}}^2 + 2.3$, where the factor 2.3 comes from the numerical experiments of Avni (1976), and find the minimum and maximum mass, age, and extinction among them.

Note that our method for deriving cluster mass and age identical to that used in Adamo et al. (2010), which has been used for star cluster analysis by a number of authors (e.g. Silva-Villa et al. 2014; Adamo et al. 2015, 2017). We have developed our own code rather than using the Adamo et al. (2010) code because that code uses simple stellar populations computed with YGGDRASIL (Zackrisson et al. 2011), whereas we wish to use photometry generated by SLUG so that our treatment of stellar evolution, stellar atmospheres, nebular emission, and dust extinction is identical to that used to generate the mock data and the libraries used for our new method. However, we have verified that the differences in the cluster masses and ages derived by our code versus the Adamo et al. (2010) code are for the most part within the error bars.

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.