Multi-layered model-based characterisation of the local-Universe galaxy data from the GAMA survey

Fan Dai,[1⋆] Ranjan Maitra,[2†] and Ivan K. Baldry[3‡]
[1] *Department of Statistics, North Dakota State University, 1340 Administration Ave, Fargo, ND 58105, USA*
[2] *Department of Statistics, Iowa State University, 2438, Osborn Drive, Ames, Iowa 50011-1090, USA*
[3] *Astrophysics Research Institute, Liverpool John Moores University, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool L3 5RF, UK*

Accepted 2026 May 21. Received 2026 May 20; in original form 2026 April 21

ABSTRACT

Understanding the formation and evolution of galaxy populations requires robust classification and characterisation techniques that jointly account for internal galaxy properties and environment. We analyse 5,306 galaxies from the Galaxy And Mass Assembly (GAMA) survey, described by stellar mass, specific star formation rate, $u-r$ colour, half-light radius, Sérsic index, and a combined environmental measure given by the optimal density. Unlike distance-based unsupervised clustering methods, our framework provides a probabilistic characterisation of galaxy populations, accommodates heavy-tailed feature distributions, and captures dependence among observables through latent factors. We model the sample using a t-mixture of factor analysers with group-specific latent structures (MtFAD), and then apply model-estimated overlap-based syncytial clustering (MOBSynC) to merge weakly separated groups and recover higher-level population structure. The first stage identifies eight simple clusters. The third and the fourth groups lie on the red, low-star-forming sequence and correspond to environmentally quenched and mass-quenched systems, respectively, while the sixth group traces the massive end of the star-forming sequence, and the seventh group appears to represent a more heterogeneous population that may include transition objects. The remaining groups populate the low- to intermediate-mass blue sequence, including both compact and more extended star-forming galaxies. The second MOBSynC stage merges the simple clusters into two compound groups: a red sequence formed by the third and the fourth groups, and the rest merging to form a broad blue sequence. Our results show that the familiar red-blue bimodality of local galaxies contains additional physically meaningful substructure linked to quenching pathway, morphology, and environment.

Key words: methods: statistical - methods: data analysis - surveys - galaxies: clusters: general - galaxies: fundamental parameters - galaxies: formation

1 INTRODUCTION

Identifying and distinguishing diverse galaxies in the local Universe has long been of major interest in astrophysics, providing key insight into the formation and evolution of galaxy populations under the influence of their inhabited environments (Postman & Geller 1984; Moore et al. 1995; Naab & Burkert 2001; Park et al. 2007; Kelvin et al. 2014; Turner et al. 2019). Traditional approaches for classifying the local-Universe galaxies rely on predefined morphological schemes, such as the Hubble sequence (Hubble 1926; Sandage 2005), that broadly separate galaxies into disc-dominated and spheroid-dominated systems, followed by further characterization using individual astrophysical properties including star formation rates (Smethurst et al. 2015), galaxy colours (Kelvin et al. 2014), stellar masses (Baldry et al. 2006; van der Wel et al. 2014), and Sérsic indices over time (Lange et al. 2014). While informative, these approaches often treat galaxy properties in isolation or impose rigid boundaries that may obscure more complex, multi-dimensional population structure; more importantly, with the rapidly increasing size and dimensionality of galaxy samples, such descriptive classification schemes become increasingly inadequate and impractical. In contrast, studies highlighting the joint influence of mass and environment—such as the separation of mass-driven and environment-driven

⋆ E-mail: fan.dai@ndsu.edu (FD)
† E-mail: maitra@iastate.edu (RM)
‡ E-mail: i.baldry@ljmu.ac.uk (IB)

quenching processes identified by Peng et al. (2010)—motivate the use of clustering-based methods that can simultaneously integrate multiple correlated features and reveal latent groupings of galaxies shaped by both intrinsic properties and environmental effects.

Advanced statistical techniques have therefore been adopted to improve the classification of large and complex galaxy samples, moving beyond traditional, descriptive schemes toward data-driven approaches capable of handling multidimensional feature spaces. Existing literature on galaxy classification has focused on supervised learning approaches built on visually labelled training samples. For example, Ball et al. (2004) utilizes the supervised artificial neural networks with Hubble-type labelled samples to classify 104619 galaxies from the Sloan Digital Sky Survey (SDSS), while Aguerri et al. (2010) conducts the morphological classification of around 70,000 galaxies from the SDSS DR7 spectroscopic sample using algorithms trained with visual classification results and Gravet et al. (2015) applies the convolutional neural networks to classify about 50,000 galaxies and again, based on training samples that are visually classified.

Supervised learning methods are capable of distinguishing massive samples; however, by definition, they require labelled observations to train the classification algorithms and consequently, are inapplicable to cases with no existing labels, as often arises in many scientific studies. Identifying groups of galaxies, for instance, in the context of the studies in this paper, is done by cluster analysis, an unsupervised learning tool that has many different approaches and algorithms. One common technique is hierarchical clustering that builds a hierarchy of clusters based on dissimilarity measures between sets of observations, applied in Ellis et al. (2005) that identifies two (early and late) types of galaxies from the Millennium Galaxy Catalogue. Another extremely common approach is k-means clustering, that iteratively assigns observations to k clusters based on the nearest centroid (the mean of all the data points within the cluster). The k-means algorithm was used by Sanchez Almeida et al. (2010) to identify major and minor classes of all the galaxy spectra in the seventh and final SDSS data release, and Turner et al. (2019) to cluster around 7,000 galaxies from the Galaxy And Mass Assembly (GAMA) survey. Other unsupervised approaches have also been used to organise or represent complex galaxy populations, including self-organizing maps (SOM), which display similarities among galaxies in a multidimensional feature space using a two-dimensional representation (Holwerda et al. 2022); manifold-based representations for morphological classification (Cooray et al. 2023); and the Fisher expectation-maximization algorithm, which has been used to distinguish galaxies based on magnitudes and spectroscopic redshifts (Siudek et al. 2018). These clustering methods, relying on measures of similarities or distance between sample points, while easy to implement, are unable to fully describe the underlying distributions of the grouped data.

In contrast, model-based clustering (MBC) (see, for instance Anderson 2003; McLachlan & Peel 2000; Mardia et al. 2006; Melnykov & Maitra 2010; Chattopadhyay & Maitra 2017, 2018) is an attractive approach to clustering because it provides a principled probabilistic-based characterisation of groups in a dataset. Typically, the probabilistic model is a mixture of component distributions with parameters to capture the central tendency and variation within the group, each of which characterises different desired aspects of each group. MBC approaches have been applied to cluster galaxies in Kelly & Mckay (2003) by means of a Gaussian mixture model (GMM) to separate around 3000 galaxies in SDSS data, or in Black & Evrard (2024) where the GMM is used to characterise the red and blue sequences of DES galaxies in the COSMOS field. Separately, Black & Evrard (2022) proposed an error-corrected GMM developed in the space of broad-band optical colours across redshift for galaxy population characterization while Zhang et al. (2023) modelled the conditional galaxy property distribution via the GMM. In each case, the GMM provides estimated ellipsoidal clusters of galaxies, which extends the isomorphic-cluster idea underlying k-means by providing a likelihood-based framework where each galaxy population is represented by a Gaussian distribution with its own mean and dispersion matrix. This yields ellipsoidal clusters, posterior membership probabilities, component-wise uncertainty estimates, and likelihood-based model comparison.

A GMM is however less effective in describing samples with longer tails, as exhibited by most of the features (after $\log_{10}$ transformation except the $u-r$ colour) of the 7,187 galaxies (see Fig 1 of our dataset that is described in greater detail in Section 3). Moreover, further investigation is required into the interdependencies among galaxy features and their dependencies on the local environment. For instance, local environmental density has been found to correlate with galaxy colour (Baldry et al. 2006; van der Burg et al. 2018; Reeves et al. 2021; Bhambhani et al. 2023) and star formation rate (Schaefer et al. 2018; Barsanti et al. 2018; Trussler et al. 2019; van de Sande et al. 2021; Sotillo-Ramos et al. 2021). Further, the existence of larger clusters that could arise from poorly separated groups remains unclear. Addressing these open questions demands probabilistic, flexible clustering frameworks that can capture intrinsic variability among galaxies while accounting for environmental effects and potential hierarchical organisation within galaxy populations.

In this paper, we cluster and characterise the galaxy data using a generalised t-mixture of factor analysers with variable numbers of factors (MtFAD) developed by Kareem & Dai (2025), plus a model-estimated overlap-based syncytial clustering (MOBSynC) adapted from Almodovar-Rivera & Maitra (2020); Chattopadhyay et al. (2022); Dai & Maitra (2024). The use of multivariate t-distributions offers a more robust modelling of the mixture components, and the factor analysers embedded in each t-distributed component can summarise all the data parameters using a few latent variables called factors, providing a better characterisation for the dispersion of the observations within the cluster. MOBSynC, on the other hand, uses measures of pairwise and generalised overlaps (Maitra & Melnykov 2010; Melnykov & Maitra 2011; Melnykov et al. 2012) to create compound or composite groups by merging the simple clusters obtained from MtFAD. We apply the methods to the galaxy samples described in Section 3, and identify eight simple clusters characterised by distinct sets of factors, and two major clusters of the red and blue sequences, where the merged blue sequence is further described via its underlying factors.

The remainder of this article is organised as follows.

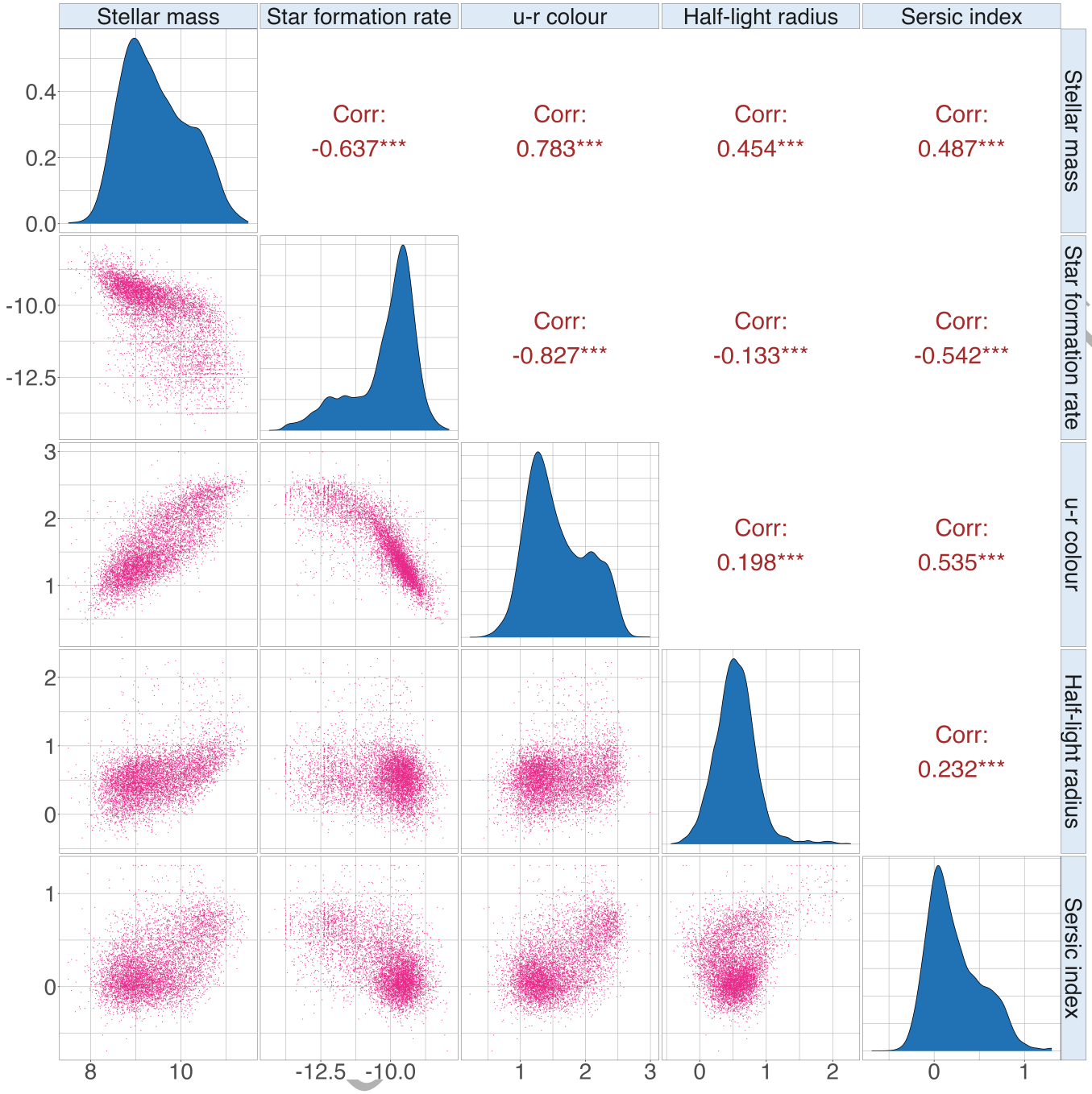


Figure 1. Densities and scatter plots of the five features (after $\log_{10}$ transformation except the $u-r$ colour) for the 7,187 local-Universe galaxies from the GAMA survey. Correlations between features are shown in the upper panel.

In Section 2, we introduce the M*t*FAD algorithm and the MOBSynC procedure. Section 3 describes the galaxy dataset that is analysed in Section 4. Finally, Section 5 summarises the paper and discusses possible avenues for further work.

2 A MULTI-LAYERED CHARACTERISATION FRAMEWORK

In this section, we present the statistical methodology underlying the clustering analysis used in this paper. Our approach begins with model-based clustering of the data using a *t*-mixture of group-specific factor analysers, together with efficient computational procedures for model parameter estimation. Building on the resulting initial partition, we then adapt the overlap-based merging framework of Chattopadhyay et al. (2022); Dai & Maitra (2024) to combine clusters according to pairwise and generalized overlap measures. The proposed methods are implemented in our `MtFAD` (*t*-Mixture of Factor Analysers in Data) and `MOBSynC` programs, both written in the open-source statistical software `R` (R Core Team 2024) and available at `https://github.com/fanstats/MBC-GAMA`.

2.1 Clustering with t-mixtures

A t-mixture model (tMM; Chattopadhyay & Maitra 2018) has the same general mixture-model structure as a GMM, but replaces the Gaussian components with p-variate t distributions, denoted by $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \nu)$, with mean vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, and degrees of freedom ν that allows each component to accommodate heavier tails and potential outliers more robustly than a Gaussian distribution. Specifically, let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ be p-dimensional observations arising from a tMM with K components, where $\boldsymbol{x}_i$ belongs to the kth component with probability η_k, for $k = 1, 2, \ldots, K$. Then, the observed data loglikelihood is

$$\ell(\boldsymbol{\Theta}; \boldsymbol{X}) = \sum_{i=1}^{n} \log \Big\{ \sum_{k=1}^{K} \eta_k f_t(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k) \Big\}, \tag{1}$$

where $\boldsymbol{X}=[\boldsymbol{x}_1 \vdots \boldsymbol{x}_2 \vdots \ldots \vdots \boldsymbol{x}_n]^\top$ is the data matrix, $\boldsymbol{\Theta}=\{(\eta_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k=1, 2, \ldots, K\}$ is the set of parameters characterizing the mixture model, and $f_t(\cdot)$ denotes the multivariate t probability density function (PDF) for the kth mixture component.

For model parameter estimation, direct maximisation of Eq. (1) is generally intractable, but can be carried out using the expectation-maximisation (EM) algorithm (Dempster et al. 1977; Rubin & Thayer 1982; McLachlan & Krishnan 2008). To do so, an unobserved component indicator z_i and a latent component-specific Gamma random variable u_i are introduced for each observed $\boldsymbol{x}_i$, with $\mathbb{P}\mathrm{r}(z_i = k) = \eta_k$ and the conditional distribution of u_i given $z_i = k$ specified to be the Gamma $\mathcal{G}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right)$ distribution. Then, conditional on u_i and that $z_i = k$, $\boldsymbol{x}_i$ is normally distributed as a p-variate $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k/u_i)$ random vector. Then, the complete (or augmented) data loglikelihood for tMM is

$$\ell(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{U}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}(z_i = k) \Big\{ \log \eta_k + \log \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \frac{\boldsymbol{\Sigma}_k}{u_i}) + \log f_G(u_i; \frac{\nu_k}{2}, \frac{\nu_k}{2}) \Big\}, \tag{2}$$

where $\boldsymbol{Z} = (z_1, z_2, \ldots, z_n)$, $\boldsymbol{U} = (u_1, u_2, \ldots, u_n)$, $\mathbb{1}(\cdot)$ denotes the indicator function, $\phi(\cdot)$ and $f_G(\cdot)$ represent the Gaussian and Gamma density functions, respectively.

Starting from an initial value of $\boldsymbol{\Theta}$, and given K, the EM algorithm alternates between the E (or expectation)-step and the M (or maximisation)-step until convergence to a locally maximum likelihood (ML) solution, as outlined in Algorithm 1.

From the ML estimates obtained via EM, each observation is allocated to the kth component for which γ_{ik} in Algorithm 1 is maximised. Having introduced the t-mixture model and its clustering framework, we now extend it to the t-mixture of factor analysers for richer characterisation of the component structure.

The above formulation has provided the most general setup for the tMM. In many cases, the variability in each mixture component can be specified by a few unobservable factors that also serves to simplify the model by reducing the number of parameters. We introduce such a model next.

Algorithm 1 EM for tMM

1: Propose a set of initial model parameter values.
2: Given current model parameters, compute the conditional expectations of z_i and u_i as follows.

$$\gamma_{ik} = \mathbb{E}[\mathbb{1}(z_i = k)|\mathbf{X}] = \frac{\eta_k f_t(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k)}{\sum_{k=1}^{K} \eta_k f_t(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \nu_k)};$$

$$\zeta_{ik} = \mathbb{E}[u_i|\mathbf{X}] = \frac{\nu_k + p}{\nu_k + (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)}.$$

3: Obtain model parameter estimates by maximising the expected complete data loglikelihood function, with updates computed as follows.

$$\eta_k^* = \frac{\sum_{i=1}^{n} \gamma_{ik}}{n},$$

$$\boldsymbol{\mu}_k^* = \frac{\sum_{i=1}^{n} \gamma_{ik} \zeta_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{n} \gamma_{ik} \zeta_{ik}},$$

$$\boldsymbol{\Sigma}_k^* = \frac{\sum_{i=1}^{n} \gamma_{ik} \zeta_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^*)(\boldsymbol{x}_i - \boldsymbol{\mu}_k^*)^\top}{\sum_{i=1}^{n} \gamma_{ik} \zeta_{ik}},$$

ν_k^*: obtained numerically by solving the equation

$$-\psi\left(\frac{\nu_k}{2}\right) + \log\left(\frac{\nu_k}{2}\right) + 1 + \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{ik} \left(\log \zeta_{ik} - \eta_{ik}\right) + \psi\left(\frac{\nu_k + p}{2}\right) - \log\left(\frac{\nu_k + p}{2}\right) = 0.$$

4: Iterate between Steps 2 and 3 until Eq. (1) converges.

2.1.1 A tMM with group-specific factor analysers

We further characterise the K components in the tMM by adopting a group-wise factor-analytic representation (Thurstone 1931, 1935; Anderson 2003), in which the p observed variables in each (kth) group are explained by q_k group-specific latent factors, with $q_k < \min(n, p)$ and $(p - q_k)^2 > p + q_k$ for identifiability. Specifically, for the kth component,

$$\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^\top + \boldsymbol{\Psi}_k, \tag{3}$$

where $\boldsymbol{\Lambda}_k$ is a $p \times q_k$ factor loading matrix whose (j, l) entry represents the strength and direction of the relationship between the jth variable and the lth latent factor, where $j = 1, 2, \ldots, p$ and $l = 1, 2, \ldots, q$. $\boldsymbol{\Psi}_k$ is a diagonal matrix of feature-specific variances. Then, conditional on u_i and $z_i = k$, $\boldsymbol{x}_i$ can be modelled through a linear equation

$$\boldsymbol{x}_i = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{F}_i + \boldsymbol{\epsilon}_i, \tag{4}$$

where given u_i and that $z_i = k$, $\boldsymbol{F}_i$ is $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{q_k}/u_i)$ distributed and represents the q_k latent factors, and conditionally independent of $\boldsymbol{\epsilon}_i$ that is $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_k/u_i)$ distributed.

Similar to the setup in Section 2.1, ML estimation of the model parameters may be carried out by combining the EM framework for tMM above with the classical EM approach for factor analysis (Dempster et al. 1977; Chen & Chen 2008). However, in practice, these iterative procedures may converge slowly, and are often sensitive to local maxima. In addition, standard formulations typically assume a common latent dimension (that is, q_k=q) across all K components. To address these issues, we adopt the M*t*FAD algorithm of Kareem & Dai (2025) that extends the t-mixture

of factor analysers to a more flexible version by allowing component-specific numbers of factors (q_k). The methodology incorporates (1) a profile likelihood strategy (Dai et al. 2020, 2021) for efficient joint updating of $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Psi}_k$ using matrix-free computations and (2) a stochastic initialization scheme (Maitra 2013; Goren & Maitra 2022) to reduce sensitivity to local maxima. The main steps of M*t*FAD are summarised in Algorithm 2.

Algorithm 2 M*t*FAD for *t*-mixture of group-specific factor analysers

1: Propose a set of initial model parameter values.
2: Given current model parameters, compute the conditional expectations of z_i and u_i as follows.

$$\mathbb{E}[\mathbb{1}(z_i = k)|\mathbf{X}] = \frac{\eta_k f_t(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^\top + \boldsymbol{\Psi}_k, \nu_k)}{\sum_{k=1}^K \eta_k f_t(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^\top + \boldsymbol{\Psi}_k, \nu_k)};$$

$$\mathbb{E}[u_i|\mathbf{X}] = \frac{\nu_k + p}{\nu_k + (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top(\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^\top + \boldsymbol{\Psi}_k)^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)},$$

where $(\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^\top + \boldsymbol{\Psi}_k)^{-1}$ is evaluated using the Woodbury matrix identity (Henderson & Searle 1981).
3: Update $\eta_k, \boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and ν_k as outlined in Algorithm 1.
4: Given $\boldsymbol{\Sigma}_k$ and q_k, update $\boldsymbol{\Lambda}_k, \boldsymbol{\Psi}_k$ as follows.

$$\boldsymbol{\Psi}_k^* = \arg\max_{\boldsymbol{\Psi}_k} \mathrm{Q}_p(\boldsymbol{\Psi}_k), \quad \boldsymbol{\Lambda}_k^* = h(\boldsymbol{\Psi}_k^*),$$

where

- $\mathrm{Q}_p(\boldsymbol{\Psi}_k)$ is the expected complete data loglikelihood obtained after profiling out $\boldsymbol{\Lambda}$ via $h(\boldsymbol{\Psi}_k)$.
- $h(\boldsymbol{\Psi}_k)$ is derived from the score equations and determined by the q largest eigenvalue–eigenvector pairs of $\boldsymbol{\Psi}_k^{-1/2}\boldsymbol{\Sigma}_k^*\boldsymbol{\Psi}_k^{-1/2}$ (Kareem & Dai 2025).

5: Repeat Steps 2, 3 and 4 until Eq. (1) converges.

2.1.2 Number of clusters and factors

Our framework so far has assumed that K and all the q_ks are known, a largely unrealistic scenario in most practical settings. We therefore choose K and q_k by using the Bayesian information criterion (BIC; Schwarz 1978) calculated by running Algorithm 2 for each of the candidate K and q_ks, and then determine the optimal values to be the set that yields the smallest BIC.

2.2 The MOBSynC algorithm for general-shaped groups

The *t*-mixture of factor analysers models data as a collection of ellipsoidally shaped groups characterised by latent factors. However, as pointed out, for example, by Almodovar-Rivera & Maitra (2020), by Chattopadhyay et al. (2022) or by Dai & Maitra (2024), some weakly separated groups may in fact represent subgroups within a larger compound cluster. Indeed, Chattopadhyay et al. (2022) provided a MOBSynC algorithm for *t*MM clusters, while Dai & Maitra (2024) developed a similar algorithm for groups obtained using a Gaussian mixture of factor analysers model. Here, we adapt the same overlap-based principle to examine the propensity for merging among the clusters identified by M*t*FAD by computing pairwise and generalized overlaps using the fitted multivariate *t* component densities. The procedure is described in Algorithm 3.

Algorithm 3 MOBSynC for clusters from M*t*FAD

1: Given the K estimated clusters from M*t*FAD, compute the $K(K-1)/2$ pairwise overlaps ω_{k_1,k_2} and the generalised overlap $\ddot{\omega}$ as follows.

$$\begin{aligned}\omega_{k_1,k_2} =& \mathbb{P}\mathrm{r}\left(\frac{\eta_{k_1} f_t(\boldsymbol{x}_i|z_i = k_2; \boldsymbol{\mu}_{k_1}, \boldsymbol{\Sigma}_{k_1}, \nu_{k_1})}{\eta_{k_2} f_t(\boldsymbol{x}_i|z_i = k_2; \boldsymbol{\mu}_{k_2}, \boldsymbol{\Sigma}_{k_2}, \nu_{k_2})} > 1\right) + \\ & \mathbb{P}\mathrm{r}\left(\frac{\eta_{k_2} f_t(\boldsymbol{x}_i|z_i = k_1; \boldsymbol{\mu}_{k_2}, \boldsymbol{\Sigma}_{k_2}, \nu_{k_2})}{\eta_{k_1} f_t(\boldsymbol{x}_i|z_i = k_1; \boldsymbol{\mu}_{k_1}, \boldsymbol{\Sigma}_{k_1}, \nu_{k_1})} > 1\right), \\ \ddot{\omega} =& (\lambda_{\boldsymbol{\Omega}}^* - 1)/(K-1),\end{aligned}$$

where $\lambda_{\boldsymbol{\Omega}}^*$ is the largest eigenvalue of the $K \times K$ matrix $\boldsymbol{\Omega}$ with pairwise overlaps ω_{k_1,k_2} as its entries.
2: Merge the k_1th and k_2th clusters if $\omega_{k_1,k_2} > \kappa\ddot{\omega}$, where κ is a positive integer indicating merging reluctance, with a larger value indicating that fewer pairs are merged in this step. Selection of the κ value was discussed in Dai & Maitra (2024).
3: Compute the probability that $\boldsymbol{x}_i$ from the m_1th compound cluster $\mathcal{C}_{m_1}$ is misclassified to the m_2th compound cluster $\mathcal{C}_{m_2}$ as follows.

$$\omega_{m_2|m_1} = \mathbb{P}\mathrm{r}\left(\frac{\sum_{j\in\mathcal{C}_{m_2}} \eta_j f_t(\boldsymbol{x}_i|z_i = m_1; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu_j)}{\sum_{l\in\mathcal{C}_{m_1}} \eta_l f_t(\boldsymbol{x}_i|z_i = m_1; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \nu_l)} > 1\right),$$

which can be approximated via the following Monte Carlo methods (Chattopadhyay et al. 2022):

(i) Generate random samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_M$ (M is set to be 10^6 in Section 4) from the mixture distribution of $\mathcal{C}_{m_1}$, which is defined as

$$\sum_{l\in\mathcal{C}_{m_1}} \eta_l^* f_t(\cdot; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \nu_l), \quad \eta_l^* = \eta_l / \sum_{h\in\mathcal{C}_{m_1}} \eta_h.$$

(ii) Estimate $\omega_{m_2|m_1}$ as

$$\widehat{\omega}_{m_2|m_1} = \frac{1}{M}\sum_{i=1}^M \mathbb{1}\left\{\frac{\sum_{j\in\mathcal{C}_{m_2}} \eta_j f_t(\boldsymbol{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu_j)}{\sum_{l\in\mathcal{C}_{m_1}} \eta_l f_t(\boldsymbol{x}_i; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \nu_l)} > 1\right\}.$$

4: Compute the pairwise overlap between $\mathcal{C}_{m_1}$ and $\mathcal{C}_{m_2}$ as $\omega_{m_1,m_2} = \widehat{\omega}_{m_1|m_2} + \widehat{\omega}_{m_2|m_1}$, and update the $\ddot{\omega}$ following the approach described in Step 1.
5: Repeat the merging steps 2, 3 and 4 until the current $\ddot{\omega}$, or its change compared to the previous stage, is below 10^{-3}.

The result of applying Algorithm 3 is that we can get a detailed multi-layered characterisation of the galaxies in the GAMA database.

3 GALAXY SAMPLES

We are interested in clustering and characterising the galaxy data using the methods outlined in this paper. For this, we aim to use a similar set of data as used by Turner

et al. (2019). Those data were taken from the GAMA survey (Baldry et al. 2018) by selecting a galaxy sample at $z < 0.06$. We use the same five features as Turner et al. (2019) but add one additional environmental feature. To do this, we use a slightly higher redshift range ($0.05 < z < 0.08$) that is better suited for environmental measurements. The dataset therefore consists of local-Universe galaxies ($0.05 < z < 0.08$) from GAMA DR4 (Driver et al. 2022; `https://gama-survey.org/dr4/data/cat`). This corresponds to 7,187 objects after removing observations with incomplete features and outliers. Each local object is described by five astrophysical features that capture the essential properties associated with the formation and evolution process of galaxies, including stellar mass (in $\log_{10}\mathrm{M}_\odot$), specific star formation rate (in $\log_{10}\mathrm{yr}^{-1}$), $u-r$ colour (in mags), half-light radius (in $\log_{10}$ kpc) and Sérsic index (in $\log_{10}$ n). All of which, except the $u-r$ colour (a logarithmic flux ratio), are analysed after $\log_{10}$ transformation because of the high skewness in their measurements.

The features were obtained from the tables called MagPhysv06 (Driver et al. 2018), StellarMassesPanChromv24 (Taylor et al. 2011) (for rest-frame $u-r$), and BDModelsv05 (Casura et al. 2022). Notably we have updated the structural fitting of galaxy profiles to those obtained by Casura et al. (2022) except we only use the single Sérsic profile fits.

As the formation and evolution of galaxies are strongly influenced by their surrounding environment, three key measurements of the local galaxy environment are considered in our cluster analysis. These data, obtained from the GAMA DR4 file server (EnvironmentMeasuresv06), include the surface density, cylindrical count, and adaptive Gaussian environment parameter. The descriptions of these environmental features are as follows:

- *Surface density* Σ - based on the distance to the 5th nearest neighbour among the density defining population in a velocity cylinder of $+/-1000$ km/s, i.e. $\frac{5}{\pi d_5^2}$ (Brough 2020; Bhambhani et al. 2023).
- *Cylindrical count* CC - measured as the number of (other) galaxies from the density defining population within a cylinder of co-moving radius 1 Mpc and a velocity range of $+/-1000$ km/s. The overdensity is given by $N_{\rm cyl}/(\bar{n}_{\rm ref}V_{\rm cyl})$, where $\bar{n}_{\rm ref} = 0.00734$ Mpc^{-3} is the average number density of the density defining population (Brough 2020; Bhambhani et al. 2023).
- *Adaptive Gaussian environment parameter* AGE - computed as

$$\mathrm{AGE} = \frac{1}{\sqrt{2\pi}\sigma}\sum_i \exp\left\{-\frac{1}{2}\left(\frac{r_{a,i}^2}{\sigma^2} + \frac{r_{z,i}^2}{(\mathrm{AGEScale}\cdot\sigma)^2}\right)\right\},$$

where r_a and r_z are the distances from the centre of the adaptive Gaussian ellipsoid in the plane of sky and along the line-of-sight in co-moving Mpc, respectively, $\sigma = 2$ Mpc, and AGEScale is the adaptive scaling factor used to scale the value of σ along the redshift axis by up to a factor of 3 for the highest density environments to compensate for the "finger-of-God" effect. This parameter is equivalent to a weighted local volume density of galaxies, where closer galaxies receive more weight than more distant ones (Brough 2020; Bhambhani et al. 2023).

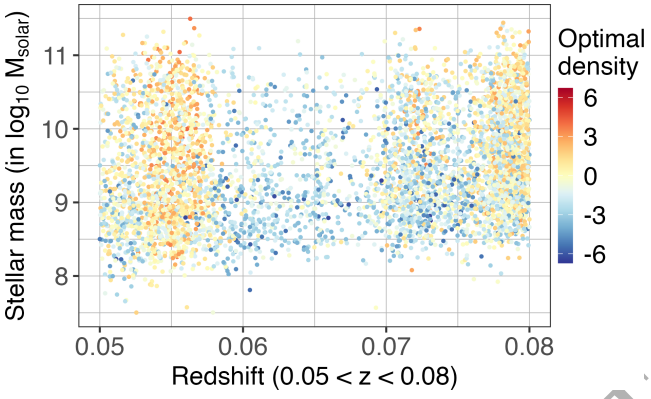


Figure 2. Redshift–stellar mass distribution of the 5,306 GAMA galaxies with complete environmental measurements. Points are colour-coded by optimal density, the combined environmental measure. The sample is restricted to $0.05 < z < 0.08$.

Within the GAMA sample, 5,306 galaxies possess complete measurements for the three environmental parameters. To summarise these effects, we adopt a combined environmental measure known as optimal density, proposed by Bhambhani et al. (2023) and detailed below, which effectively captures the variation in the red galaxy fraction and provides a more reliable measurement of local environmental influence. We therefore use

Optimal density Λ - computed as a linear combination of the surface density, cylindrical count, and adaptive Gaussian environment parameter:

$$\log\Lambda = \log\Sigma + \alpha\log\mathrm{CC} + \beta\log\mathrm{AGE}, \quad (5)$$

where $\log\mathrm{CC} = -1$ at $\mathrm{CC} = 0$. α and β are numerically determined to maximize the red fraction range of Λ. For our fully observed sample of 5,306 galaxies, the best results are $\alpha = 0.76$ and $\beta = 1.41$, with the local optimal density Λ achieving the highest red fraction range (0.469) compared to Σ (0.397), CC (0.425), and AGE (0.391).

In sum, our final dataset consists of 5,306 local-Universe galaxies with complete records on six parameters: Stellar mass, star formation rate, $u-r$ colour, half-light radius, Sérsic index, and optimal density. Fig 2 shows that these galaxies lie within the redshift range $0.05 < z < 0.08$ and span approximately 7.5–11.5 in log stellar masses. The colour gradient indicates that environmental density varies across the full sample, with higher optimal density values appearing more frequently among relatively massive galaxies.

4 RESULTS AND ANALYSIS

We applied the M*t*FAD algorithm to the 5,306 fully observed GAMA galaxies, each described by five intrinsic astrophysical properties and one combined environmental variable—the optimal density. The modelling considered up to fifteen mixture components and up to two latent factors (consistent with the maximum permissible number of factors for six observed features). The resulting clustering structure and latent characterisations are summarised below.

Table 1: Model-selection results for M*t*FAD over candidate numbers of groups $K = 1, 2, \ldots, 15$. For each K, the minimum BIC is taken over all possible group-specific q_k configurations $\mathbf{q}_K = (q_1, \ldots, q_K)$, with $q_k \in \{1, 2\}$ for $k = 1, 2, \ldots, K$. Here $\Delta\text{BIC} = \text{BIC}_{K-1} - \text{BIC}_K$ for $K \geq 2$. The selected model is highlighted in bold.

K	Selected $\mathbf{q}_K$	Min. BIC	ΔBIC
1	(2)	43757.92	–
2	(2, 2)	41452.73	2305.19
3	(2, 2, 2)	40188.36	1264.37
4	(2, 2, 2, 2)	39250.83	937.53
5	(2, 2, 2, 2, 1)	38500.60	750.23
6	(2, 2, 2, 2, 1, 2)	37960.88	539.72
7	(1, 2, 2, 2, 2, 2, 2)	37857.39	103.49
8	**(1, 2, 2, 2, 2, 2, 2, 1)**	**37489.09**	368.30
9	(2, 1, 2, 2, 2, 2, 2, 2, 1)	37747.81	-258.72
10	(2, 1, 2, 1, 1, 2, 2, 2, 1)	37847.07	-99.26
11	(2, 1, 1, 2, 2, 2, 2, 2, 1, 2, 1)	37921.59	-74.52
12	(2, 1, 2, 1, 1, 2, 2, 2, 1, 2, 2, 2)	37977.30	-55.71
13	(1, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 1)	38025.69	-48.39
14	(2, 1, 1, 2, 2, 2, 2, 2, 1, 2, 1, 1, 1, 1)	38006.58	19.11
15	(2, 1, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1)	37994.32	12.26

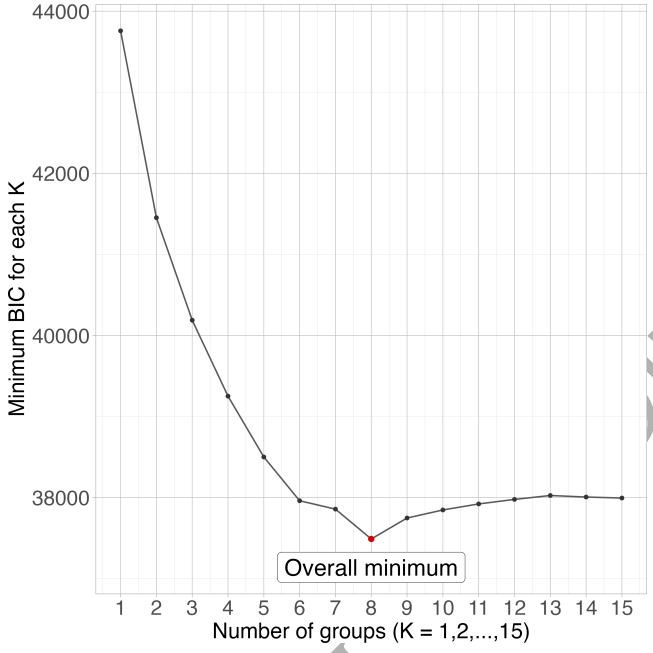


Figure 3. Minimum BIC values for the M*t*FAD model over candidate numbers of groups $K = 1, 2, \ldots, 15$. For each fixed K, the plotted value is the smallest BIC obtained over all group-specific q_k configurations, with $q_k \in \{1, 2\}$ for $k = 1, 2, .., K$. The overall minimum is obtained at $K = 8$.

4.1 M*t*FAD grouping

Our algorithm when applied with BIC selected $K = 8$ simple clusters (see Table 1 and Fig 3). The optimal number of factors for each group was $q_k = 1$ for Groups 1 and 8, and $q_k = 2$ for the remaining groups. We also see that Groups 2, 3, 5, 6 and 8 are essentially normally-distributed given that $\widehat{\nu_k} > 100$ for these groups. Table 2 lists the galaxy counts of the eight estimated groups, where we see that Group 7 contains the fewest number of galaxies which is around 12% of the largest Group 5 in terms of the sample size. Table 3 presents the mean and standard deviation for each of the six parameters within the identified groups, while Fig 4 provides a visual summary of their distributional patterns. Groups 3 and 4 are characterised by comparatively higher stellar masses, redder $u - r$ colours, and lower average star formation rates, consistent with more evolved galaxy populations. Among them, Group 4 is particularly distinguished by the reddest colours, the most suppressed star formation activity, and relatively large Sérsic indices with less variations, while Group 3 stands out as the cluster with the highest optimal environmental density. In contrast, Groups 1, 2, 5, and 8 are generally associated with lower stellar masses, bluer colours, smaller Sérsic indices, and higher average star formation rates. Among these, Groups 2 and 8 show the lowest optimal environmental densities. Groups 6 and 7 also exhibit relatively low optimal densities together with the largest half-light radii. In particular, Group 7 has the largest mean half-light radius and one of the highest Sérsic indices, with the greatest variation in both quantities.

The simple clusters are further illustrated in Fig. 5 using the 3D visualization framework of Zhu et al. (2021), where the cluster locations relative to the projected feature directions highlight their main distinguishing characteristics, as summarised in Table 2. The red-sequence galaxies are preferentially located toward higher stellar mass and larger Sérsic index, while suppressed star formation rates. Conversely, the blue-sequence systems occupy regions characterized by lower stellar mass and Sérsic index, and exhibit elevated star formation activity. Collectively, the estimated simple clusters demonstrate clear physical and environmental differentiation associated with galaxy formation and evolution, and indicate that massive and red galaxies preferentially inhabit denser environments, while lower-mass, star-forming galaxies are more common in lower-density regions.

4.1.1 Latent structure analysis

We further examined the latent structure within each simple cluster using the estimated factor analysers. Table 4 presents both the numerical values and visual representations of the group-wise factor loadings, corresponding to the columns of the loading matrix $\mathbf{\Lambda}_k$ described in Section 2.1.1. To enhance interpretability, an oblimin rotation (Costello & Osborne 2005) was applied to each loading matrix. Only loadings with magnitudes greater than 0.1 are displayed in the table. Each loading value reflects the contribution of an individual feature to a specific latent factor, with the sign (+ or −) indicating the direction of the relationship between the feature and the factor.

Group 1 is characterised by a single factor that is primarily driven by stellar mass, with additional smaller contributions from half-light radius and $u - r$ colour, which are opposed by smaller to minor components from star formation rate and optimal density.

In Group 2, the first factor is dominated by half-light radius, with a minor part of stellar mass and small opposing contribution from Sérsic index. While the second factor is primarily defined by a strong negative loading from $u - r$ colour, accompanied by substantial to minor components from stellar mass, Sérsic index and optimal density on one

Table 2: Data with Optimal Density: Galaxy counts and astrophysical interpretation of the simple clusters. The descriptions are based on the feature distributions in Fig 4 and the 3D representations relative to the six feature directions in Fig 5.

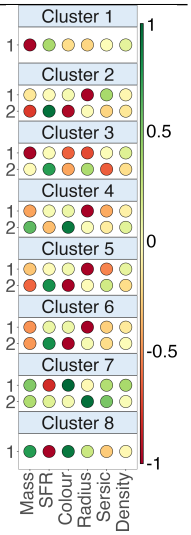
Group	Galaxy count	Approximate population	Main distinguishing characteristics
1	285	Low-mass blue/star-forming sequence	Low stellar mass, blue colour, relatively high SFR, low Sérsic index.
2	874	Compact low-mass blue/star-forming sequence	Low stellar mass, blue colour, high SFR, compact sizes, low environmental density.
3	604	Environmentally quenched red sequence	Red colour, low SFR, intermediate-to-high stellar mass, highest optimal density.
4	664	Mass-quenched red sequence	Highest stellar mass, reddest colour, lowest SFR, high Sérsic index.
5	1,662	Extended low- to intermediate-mass blue/star-forming sequence	Blue colour, high SFR, larger sizes than other blue groups, low Sérsic index.
6	496	High-mass end of a star-forming sequence	High stellar mass, relatively large size, moderate colour, SFR above the quenched groups.
7	209	Transition population and/or sources with large uncertainties	Broad feature distributions, largest sizes, high Sérsic index, intermediate colour and SFR.
8	512	Low- to intermediate-mass blue/star-forming systems	Blue colour, high SFR, low Sérsic index, lowest environmental density.

Table 3: Data with Optimal Density: Estimated feature means and standard deviations (in parenthesis) for simple clusters.

Feature / Cluster	Stellar mass	Star formation rate	$u-r$ colour	Half-light radius	Sérsic index	Optimal density
1	8.77(0.35)	−10.15(1.00)	1.39(0.31)	0.47(0.17)	0.03(0.13)	−0.72(1.90)
2	8.94(0.44)	−9.59(0.54)	1.32(0.31)	0.21(0.21)	0.26(0.27)	−1.27(1.99)
3	9.67(0.50)	−11.25(0.96)	2.01(0.24)	0.43(0.20)	0.23(0.19)	1.16(1.38)
4	10.45(0.42)	−12.17(0.72)	2.30(0.16)	0.61(0.29)	0.66(0.14)	0.01(1.92)
5	9.21(0.50)	−9.58(0.35)	1.27(0.20)	0.64(0.16)	0.05(0.15)	−0.97(1.81)
6	10.39(0.32)	−10.28(0.46)	1.86(0.24)	0.77(0.28)	0.36(0.23)	−0.94(1.83)
7	9.67(0.72)	−10.73(1.17)	1.74(0.43)	1.10(0.61)	0.80(0.26)	−0.39(1.60)
8	9.07(0.50)	−9.48(0.37)	1.35(0.33)	0.46(0.20)	0.00(0.12)	−1.80(1.92)

Table 4: Data with Optimal Density: Estimated factor loadings (in the correlation scale) for simple clusters, along with a heatmap for reference. For clarity of presentation, values in the interval (-0.1,0.1) are suppressed in the table, but displayed using light colours in the heatmap representation.

k	q	Stellar mass	Star formation rate	$u-r$ colour	Half-light radius	Sérsic index	Optimal density	Heatmap
1	1	−0.997	0.368	−0.213	−0.249		0.121	
2	1	−0.158			−0.995	0.373		
	2	−0.776	0.945	−0.974		−0.253	−0.159	
3	1	−0.993		−0.650	−0.724		0.149	
	2		0.749	−0.434	0.355	−0.658	−0.210	
4	1	−0.442			−0.999	−0.418	−0.102	
	2	0.603	−0.345	0.909		−0.128	0.134	
5	1	−0.315			−0.994	−0.537	0.112	
	2	−0.639	0.814	−0.995		0.217	−0.216	
6	1	−0.492			−0.996	−0.274	−0.147	
	2	−0.479	0.817	−0.995		−0.219		
7	1	0.496	−0.831	0.984		0.339	0.370	
	2	0.346	0.146		0.951	0.521		
8	1	0.758	−0.997	0.910	0.244	−0.296		



side, against a dominant contribution from star formation rate on the other.

For Group 3, the first factor is mainly explained by stellar mass, together with substantial to moderate contributions from half-light radius and $u-r$ colour, opposed by a minor component from optimal density. The second fac-

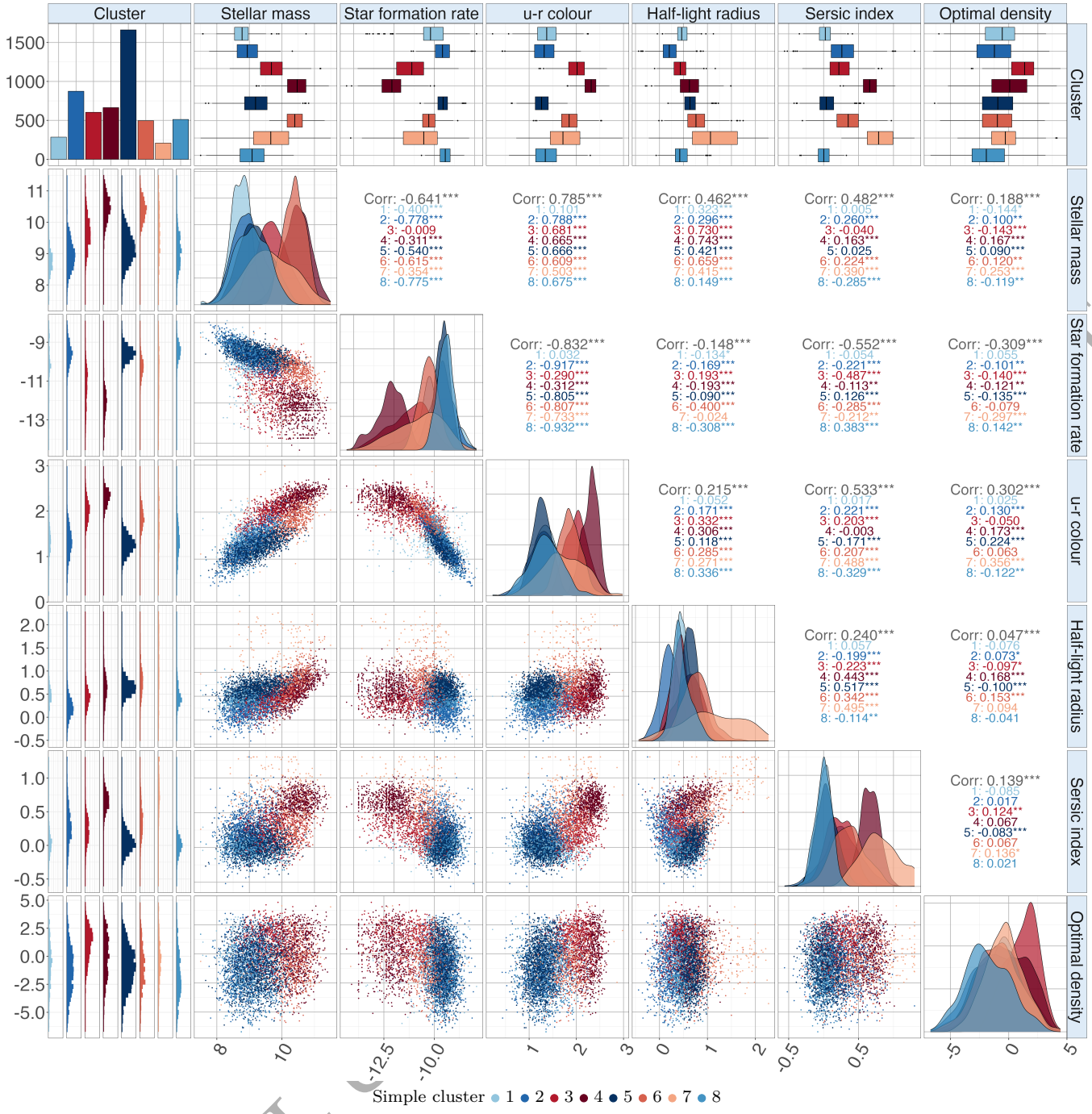


Figure 4. Data with Optimal Density: Densities and scatter plots of the six features: Stellar mass, star formation rate, $u-r$ colour, half-light radius, Sérsic index, and optimal density, for simple clusters (indicated by colours). Correlations between features are shown in the upper panel.

tor contrasts a strong contribution from star formation rate and a moderate part from half-light radius, against opposing loadings from Sérsic index, $u-r$ colour, and optimal density, whose magnitudes decrease from moderate to smaller to minor.

The first factor in Group 4 is primarily driven by half-light radius, with additional moderate and smaller contributions from stellar mass, Sérsic index, and optimal density on the same side. The second factor reflects a contrast between strong positive contributions from $u-r$ colour and stellar mass, together with a minor part from optimal density, and opposing smaller to minor components from star formation rate and Sérsic index.

Group 5 has the first factor dominated by half-light radius, along with moderate to smaller contributions from Sérsic index and stellar mass, and a minor opposing component from optimal density. The second factor contrasts substantial to small contributions from star formation rate and Sérsic index, against a dominant part of $u-r$ colour, which

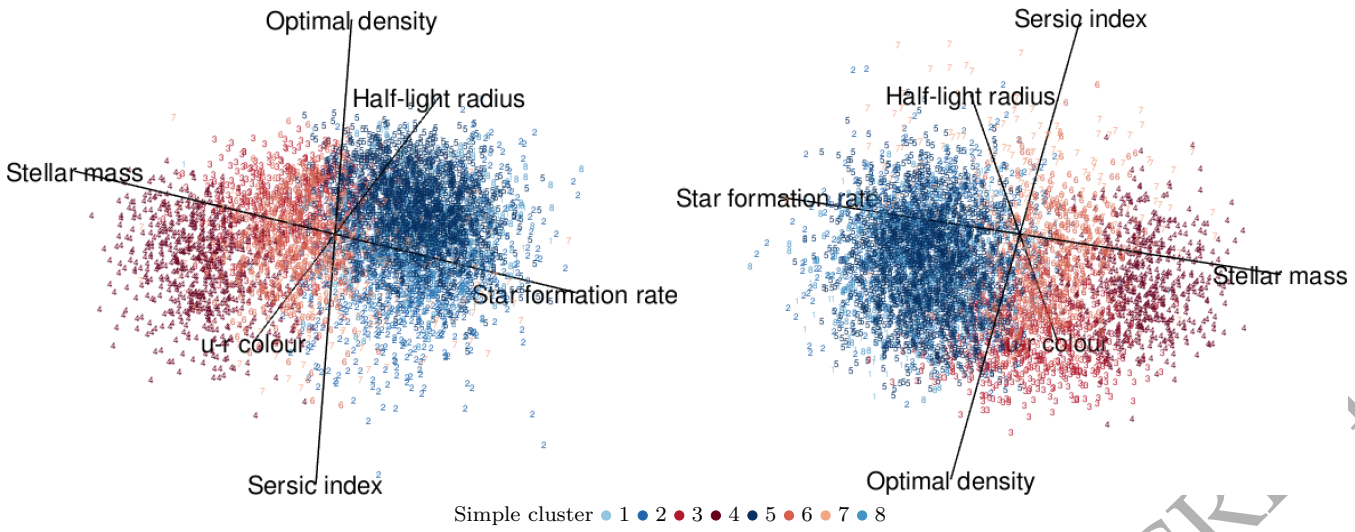


Figure 5. Data with Optimal Density: 3D star coordinates plots for simple clusters.

is accompanied by moderate to small components of stellar mass and optimal density.

In Group 6, the first factor is mainly explained by half-light radius, together with moderate to minor additional contributions from stellar mass, Sérsic index and optimal density on the same side. The second factor is dominated by a strong loading from $u-r$ colour, accompanied by moderate to small contributions from the stellar mass and Sérsic index, which are opposed by a major part of star formation rate.

For Group 7, the first factor contrasts star formation rate against $u-r$ colour, stellar mass, optimal density, and Sérsic index, whose contributions decrease from dominant to moderate to smaller. The second factor is primarily driven by half-light radius, with an additional moderate contribution from Sérsic index and smaller to minor components from stellar mass and star formation rate.

Finally, Group 8 has a single factor that is dominated by star formation rate and a smaller part of Sérsic index, contrasted with strong contributions from $u-r$ colour and stellar mass, together with a smaller component from half-light radius.

We also computed the unbiased estimates of the factor scores $\boldsymbol{F}_i$ specified in Eq. (4) using the Bartlett method (Bartlett 1937; Hershberger 2005; Distefano et al. 2009). The results are given in Table 5, where the mean score values represent the average importance of the latent factors as "rated" by its galaxies members within the group. For Groups 2, 3, 4 and 6, Factor 1 contributes more strongly than Factor 2, while Groups 5 and 7 show the reverse trend. Overall, the distinct factor-loading patterns across the eight simple clusters reveal diverse variability within the galaxy group and reinforce the colour–star-formation dichotomy as a key latent dimension among the clusters.

Table 5: Data with Optimal Density: Mean factor scores for simple clusters.

Cluster	1	2	3	4	5	6	7	8
Factor 1	0.07	0.22	0.03	0.01	−0.03	−0.03	0.01	0.01
Factor 2	—	−0.01	0.01	−0.02	0.05	−0.07	0.18	—

4.1.2 Physical picture of the eight simple clusters

The eight simple clusters are displayed in feature space in Fig. 4, and their approximate astrophysical interpretations are summarized in Table 2. Specifically, Groups 3 and 4 stand out as being both red in $u-r$ and with low specific star formation rate (SFR). These can be associated with quenched populations (Peng et al. 2010, 2012). Given the higher values for the environmental measure, Group 3 represents environmentally quenched galaxies while Group 4, with log stellar masses $\gtrsim$10, represents mass-quenched galaxies (Cochrane & Best 2018).

There are two intermediate clusters. Group 6 has high mass and while it is quite red, the SFRs are significantly higher than the quenched population. This cluster represents the high-mass end of a star-forming sequence (Brinchmann et al. 2004). Group 7 straddles a wide range of physical feature space values. This may represent a combination of effects, for example, a transition population and/or sources with large uncertainties in feature space. The high Sérsic index and large sizes for some of this space may indicate poor single Sérsic fits. Note this is the smallest cluster in number.

The four remaining clusters of blue galaxies (Groups 1, 2, 5 and 8) cover low to intermediate masses of the star-forming sequence. The two largest in number of these clusters separate into a compact galaxy sample (Group 2) and a more extended galaxy sample (Group 5). There is no obvious difference in environmental density between these two groups suggesting this size difference is related to secular evolution.

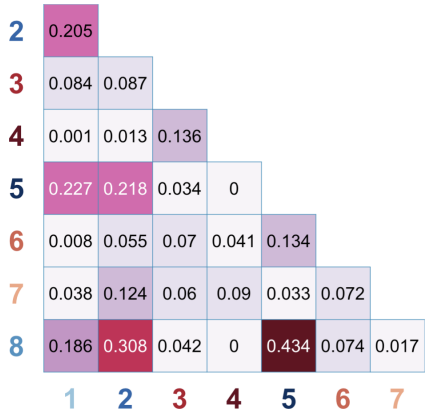


Figure 6. Data with Optimal Density: Pairwise overlap measures between any two of the simple clusters. The generalised overlap is $\ddot{\omega} = 0.123$.

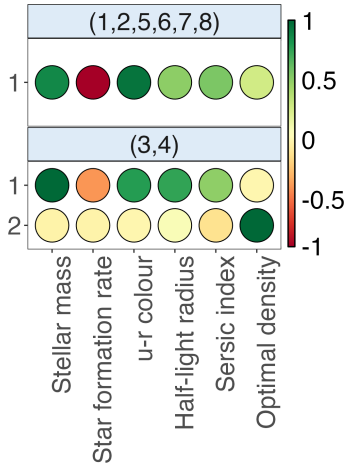


Figure 7. Data with Optimal Density: Heatmap of the estimated factor loadings for compound clusters.

The absence of a cluster corresponding to the classical "green-valley" population is also informative. Galaxies with intermediate colour or star-formation properties are not recovered as a distinct group, but are mainly distributed across Groups 6 and 7. This suggests that, in the present six-dimensional feature space, green-valley galaxies occupy a transitional region between the quenched and star-forming populations.

4.2 MOBSynC grouping of the M*t*FAD simple clusters

Using the simple clusters identified in Section 4.1, we further investigated the presence of compound clusters using MOBSynC described in Section 2.2. Fig 6 shows the pairwise overlaps among the eight clusters, with a generalised overlap of $\ddot{\omega} = 0.123$. Using the selected threshold $\kappa = 1$, Groups 3 and 4 merge to form one compound cluster, while the remaining groups combine into another cluster, resulting in two compound clusters at the final stage. The merging phases and outputs are visualized in Fig 8, where at each phase, the clusters are ordered vertically by average $u - r$ colour so that the reddest to bluest galaxy groups are shown from top to bottom.

Fig 9 shows the final feature distributions, demonstrating the two main galaxy populations. Specifically, the compound cluster formed by merging Groups 3 and 4 exhibits substantially higher stellar masses, redder $u-r$ colours, and larger Sérsic indices and optimal densities, together with a much lower specific star formation rate, compared to the other compound cluster (formed by Groups 1, 2, 5, 6, 7 and 8). Both compound clusters show the largest variation in optimal density, whereas the half-light radius appears to be the least distinctive galaxy property between them.

We further characterised the two compound clusters using a factor model. Because the groups formed by merging simple clusters is no longer normally distributed, we first applied a Gaussian distributional transform (GDT; Zhu et al. 2021; Dai & Maitra 2024) to normalize the results, and then performed factor analysis in the Gaussianised space. Fig 7 presents the resulting factor loadings. For the blue sequence (given by the compound cluster formed by merging Groups 1, 2, 5, 6, 7 and 8), the single factor reflects a contrast between star formation rate and strong to moderate to smaller opposing contributions from $u-r$ colour, stellar mass, Sérsic index, half-light radius, and optimal density in that order. For the red sequence (that is the compound cluster formed by merging Groups 3 and 4), the first factor contrasts star formation rate against stellar mass, $u-r$ colour, half-light radius, and Sérsic index, whose magnitudes decrease in that order. The second factor is largely dominated by optimal density, opposed by a minor part from Sérsic index. In sum, MOBSynC clearly distinguishes between the red and blue galaxy systems, consistent with the well-known colour bimodality.

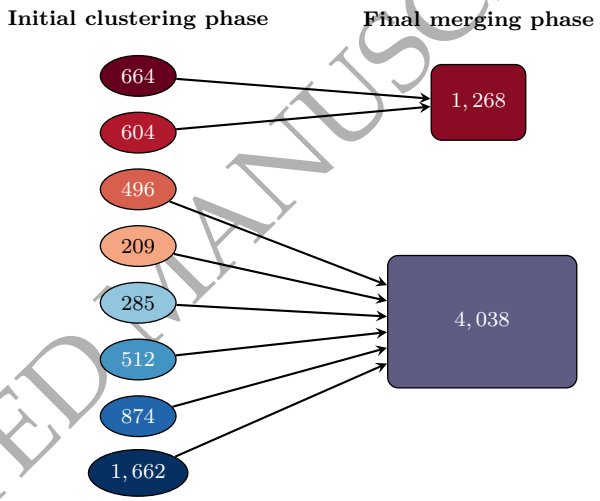


Figure 8. Data with Optimal Density: Flowchart illustrating the application of MOBSynC on simple clusters. Clusters are ordered vertically at each stage according to the average $u-r$ colour of their member galaxies.

5 CONCLUSIONS

In this paper, we applied a t-mixture of group-specific factor analysers to cluster local-Universe galaxies from the GAMA survey, characterised by five astrophysical features and one

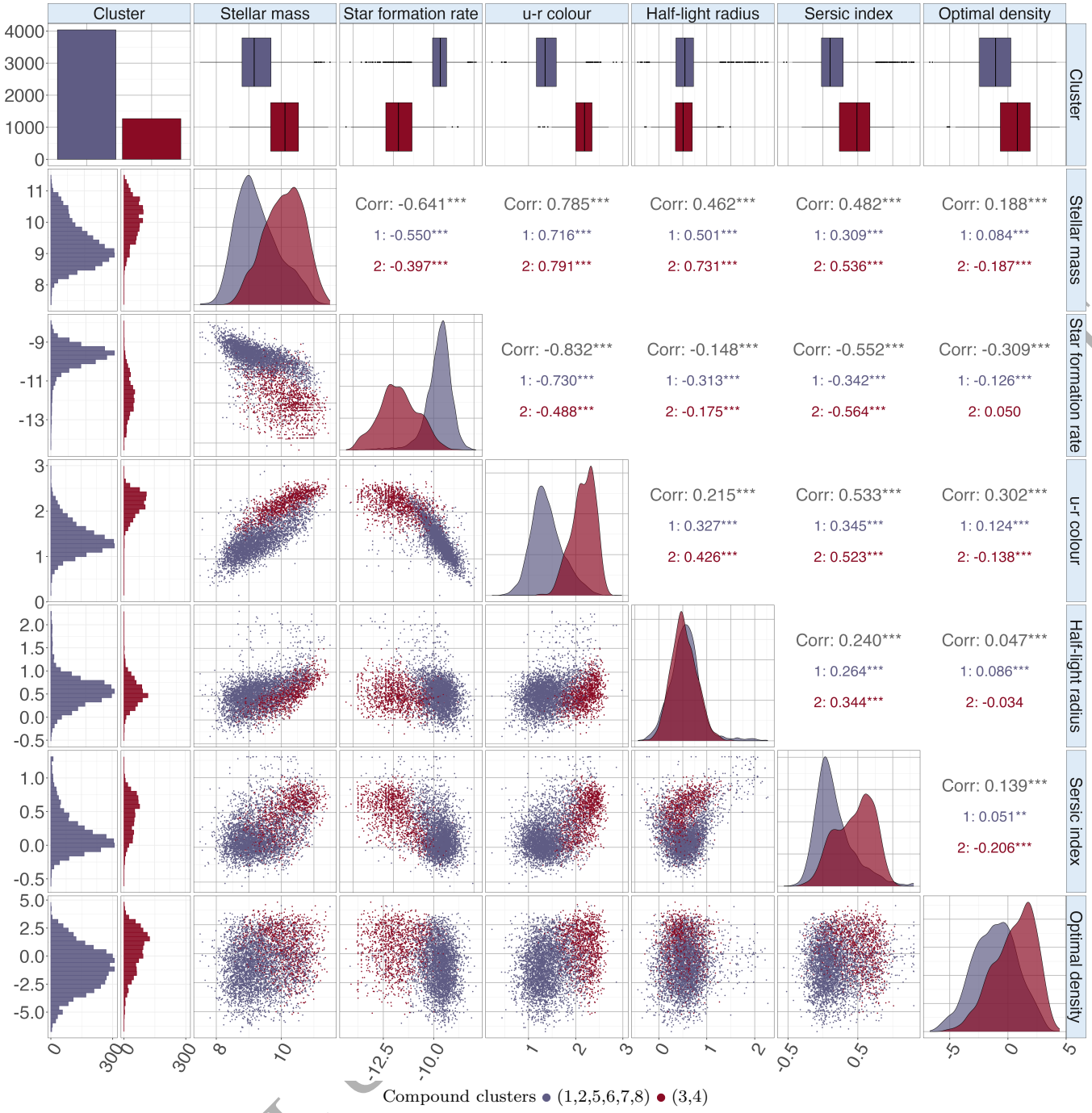


Figure 9. Data with Optimal Density: Densities and scatter plots of the six features: Stellar mass, star formation rate, $u-r$ colour, half-light radius, Sérsic index, and optimal density, for compound clusters (indicated by colours). Correlations between features are shown in the upper panel.

combined environmental parameter. We identified eight simple clusters that exhibit distinctivenesses in galaxy properties and in optimal density, revealing valuable insights into the diversity of galaxy populations. These simple clusters provide a finer-level description of the galaxy population, separating, for example, environmentally quenched and mass-quenched red-sequence systems, as well as several star-forming sequence groups that differ in stellar mass, size, morphology, and environment. We further employed MOB-SynC to identify larger composite structures by merging less well-separated groups, ultimately yielding two major galaxy classes corresponding to the red and blue sequences. Thus, the familiar red–blue bimodality is recovered as a higher-level structure, while the eight simple clusters reveal additional substructure within this broader bimodality. Each identified cluster was further characterised by latent factors, revealing additional differences in internal galaxy variability and the underlying physical processes.

The two-stage analysis therefore supports a hierarchical interpretation of the local-Universe GAMA galaxy pop-

ulation. At the first level, M*t*FAD identifies simple clusters that capture detailed variation in stellar mass, star formation activity, colour, size, morphology, and environment. At the second level, MOBSynC merges overlapping clusters and recovers the dominant red–blue division. This suggests that the global bimodality remains a major organising feature of local galaxies, but that each side of the bimodality contains physically meaningful substructure associated with quenching pathway, morphology, size, and environment.

From a machine learning perspective, the M*t*FAD clusters are interpreted as model-based probabilistic components in the 6D feature space. This differs from k-means clustering, which produces a distance-based partition, and from SOM, which provides a primarily visualization-focused low-dimensional organization of the data. In the present analysis, each M*t*FAD component represents a region of feature space with its own centre, covariance structure, tail behaviour, and latent factor representation. The subsequent MOBSynC analysis then assesses the overlap among these components and merges weakly separated groups to capture the higher-level structure of the feature space.

Although the final galaxy samples excluded incomplete features and outliers, these components can provide useful information on the intrinsic data structure and generative mechanisms. Incorporating measurement errors, missing values, and outliers into statistical modelling therefore remains important. Future work should focus on developing robust, model-based clustering approaches that naturally accommodate these complexities, offering a more complete and realistic characterisation of the data.

DATA AVAILABILITY STATEMENT

The galaxy data used in this article comprise local-Universe galaxies from GAMA DR4 (Driver et al. 2022); available at `https://gama-survey.org/dr4/data/cat`. The galaxy features were obtained from the following GAMA DR4 tables: `MagPhysv06` (Driver et al. 2018), `StellarMassesPanChromv24` (Taylor et al. 2011) (for rest-frame $u - r$ colour), and `BDModelsv05` (Casura et al. 2022). The environmental data used in this article consist of three local galaxy environment measures from the GAMA DR4 file server, `EnvironmentMeasuresv06`, available at `https://gama-survey.org/dr4/data/cat/EnvironmentMeasures/v06/`. The processed datasets, together with the `R` code used to analyse the sample and generate the tables and figures, are publicly available at `https://github.com/fanstats/MBC-GAMA`. Software implementing `MtFAD` will be made publicly available as an `R` (R Core Team 2024) package of the same name, while `MOBSynC` will be released as part of the publicly available `SynClustR` package in `R` (R Core Team 2024).

References

Aguerri J., Bernardi M., Mei S., Sanchez Almeida J., 2010, Astronomy and Astrophysics, 525

Almodovar-Rivera I. A., Maitra R., 2020, Journal of Machine Learning Research, 21, 1

Anderson T. W., 2003, An Introduction to multivariate statistical analysis. Wiley Series in Probability and Statistics, Wiley

Baldry I., Balogh M., Bower R., G. G., Nichol R., Bamford S., Budavari T., 2006, MNRAS, 373

Baldry I. K., et al., 2018, MNRAS, 474, 3875

Ball N., Loveday J., Fukugita M., Nakamura O., Okamura S., Brinkmann J., Brunner R., 2004, Monthly Notices of the Royal Astronomical Society, 348, 1038

Barsanti S., et al., 2018, The Astrophysical Journal, 857, 71

Bartlett M., 1937, British Journal of Psychology. General Section, 28, 97

Bhambhani P. C., Baldry I. K., Brough S., Hill A. D., Lara-Lopez M. A., Loveday J., Holwerda B. W., 2023, MNRAS, 522, 4116

Black W., Evrard A., 2022, Monthly Notices of the Royal Astronomical Society, 516

Black W., Evrard A., 2024, The Open Journal of Astrophysics, 7

Brinchmann J., Charlot S., White S., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151

Brough S., 2020, EnvironmentMeasures, `https://gama-survey.org/dr4/data/cat/EnvironmentMeasures/v06/EnvironmentMeasuresv06.notes`

Casura S., et al., 2022, Monthly Notices of the Royal Astronomical Society, 516, 942

Chattopadhyay S., Maitra R., 2017, Monthly Notices of the Royal Astronomical Society, 469, 3374

Chattopadhyay S., Maitra R., 2018, Monthly Notices of the Royal Astronomical Society, 481, 3196

Chattopadhyay S., Kawaler S. D., Maitra R., 2022, Publications of the Astronomical Society of Australia, 39, 1

Chen J., Chen Z., 2008, Biometrika, 95, 759

Cochrane R. K., Best P. N., 2018, MNRAS, 480, 864

Cooray S., Takeuchi T. T., Kashino D., Yoshida S. A., Ma H.-X., Kono K. T., 2023, Monthly Notices of the Royal Astronomical Society, 524, 4976

Costello A. B., Osborne J., 2005, Practical Assessment, Research & Evaluation, 10, 1

Dai F., Maitra R., 2024, Monthly Notices of the Royal Astronomical Society

Dai F., Dutta S., Maitra R., 2020, Journal of Computational and Graphical Statistics, 29, 675

Dai F., Dorman K. S., Dutta S., Maitra R., 2021, Exploratory Factor Analysis of Data on a Sphere, doi:10.48550/ARXIV.2111.04940, `https://arxiv.org/abs/2111.04940`

Dempster A. P., Laird N. M., Rubin D. B., 1977, Jounal of the Royal Statistical Society, Series B, 39, 1

Distefano C., Zhu M., Mindrila D., 2009, Practical Assessment, Research and Evaluation, 14, 20

Driver S. P., et al., 2018, MNRAS, 475, 2891

Driver S. P., et al., 2022, MNRAS, 513, 439

Ellis S., Driver S., Allen P., Liske J., Bland-Hawthorn J., Propris R., 2005, Monthly Notices of the Royal Astronomical Society, 363, 1257

Goren E. M., Maitra R., 2022, Stat, 11, 416

Gravet R., et al., 2015, The Astrophysical Journal Supplement Series, 221

Henderson H. V., Searle S. R., 1981, SIAM Review, 23, 53

Hershberger S. L., 2005, Encyclopedia of Statistics in Behavioral Science, pp 636–644

Holwerda B. W., et al., 2022, MNRAS, 513, 1972

Hubble E. P., 1926, ApJ, 64, 321

Kareem K., Dai F., 2025, A Hybrid Mixture of t-Factor Analyzers for Clustering High-dimensional Data (`arXiv:2504.21120`), `https://arxiv.org/abs/2504.21120`

Kelly B., Mckay T., 2003, The Astronomical Journal, 127

Kelvin S., et al., 2014, Monthly Notices of the Royal Astronomical Society, 439

Lange R., et al., 2014, Monthly Notices of the Royal Astronomical Society, 447

Maitra R., 2013, Sankhyā: The Indian Journal of Statistics, Series

B, 75, 293
Maitra R., Melnykov V., 2010, Journal of Computational and Graphical Statistics, 19, 354
Mardia K. V., Kent J. T., Bibby J. M., 2006, Multivariate analysis. Elsevier, Amsterdam
McLachlan G., Krishnan T., 2008, The EM Algorithm and Extensions, second edn. Wiley, New York, doi:10.2307/2534032
McLachlan G., Peel D., 2000, Finite Mixture Models. John Wiley and Sons, Inc., New York, doi:10.1002/0471721182
Melnykov V., Maitra R., 2010, Statist. Surv., 4, 80
Melnykov V., Maitra R., 2011, Journal of Machine Learning Research, 12, 69
Melnykov V., Chen W.-C., Maitra R., 2012, Journal of Statistical Software, 51, 1
Moore B., Katz N., Lake G., Dressler A., Oemler A., 1995, Nature, 379
Naab T., Burkert A., 2001, The Astrophysical Journal, 597
Park C., III J., Choi Y.-Y., 2007, The Astrophysical Journal, 674
Peng Y.-j., et al., 2010, ApJ, 721, 193
Peng Y.-j., Lilly S. J., Renzini A., Carollo M., 2012, ApJ, 757, 4
Postman M., Geller M. J., 1984, ApJ, 281, 95
R Core Team 2024, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/
Reeves A. M. M., et al., 2021, Monthly Notices of the Royal Astronomical Society, 506, 3364
Rubin D. B., Thayer D. T., 1982, Psychometrika, 47, 69
Sanchez Almeida J., Aguerri J., Munoz-Tunon C., De Vicente A., 2010, The Astrophysical Journal, 714, 487
Sandage A., 2005, Annual Review of Astronomy and Astrophysics, 43, 581
Schaefer A. L., et al., 2018, Monthly Notices of the Royal Astronomical Society, 483, 2851
Schwarz G. E., 1978, The Annals of Statistics, 6, 461
Siudek M., et al., 2018, A&A, 617, A70
Smethurst R., et al., 2015, Monthly Notices of the Royal Astronomical Society, 450
Sotillo-Ramos D., et al., 2021, Monthly Notices of the Royal Astronomical Society, 508, 1817
Taylor E. N., et al., 2011, MNRAS, 418, 1587
Thurstone L., 1931, Psychological Review, 38, 406
Thurstone L. L., 1935, The Vectors of Mind: Multiple-factor Analysis for the Isolation of Primary Traits.. University of Chicago Press
Trussler J., Maiolino R., Maraston C., Peng Y., Thomas D., Goddard D., Lian J., 2019, Monthly Notices of the Royal Astronomical Society, 491, 5406
Turner S., et al., 2019, Monthly Notices of the Royal Astronomical Society, 482, 126
Zhang Y., Pullen A., Somerville R., Breysse P., Forbes J., Yang S., Li Y., Maniyar A., 2023, The Astrophysical Journal, 950, 159
Zhu Y., Dai F., Maitra R., 2021, Visualization of Labeled Mixed-featured Datasets, doi:10.48550/ARXIV.1904.06366, https://arxiv.org/abs/1904.06366
van der Burg R. F. J., McGee S. L., Aussel H., Dahle H., Arnaud M. D., Pratt G. W., Muzzin A., 2018, Astronomy & Astrophysics
van der Wel A., et al., 2014, The Astrophysical Journal, 788
van de Sande J., et al., 2021, Monthly Notices of the Royal Astronomical Society, 508, 2307