

Bayesian Level Set Clustering

David Buch*

DAVIDBUCH42@GMAIL.COM

Two Sigma

New York, New York 10013, USA

Miheer Dewaskar*

MDEWASKAR@UNM.EDU

Department of Mathematics and Statistics

University of New Mexico

Albuquerque, New Mexico 87106, USA

David B. Dunson

DUNSON@DUKE.EDU

Department of Statistical Science

Duke University

Durham, North Carolina 27708, USA

Editor: Debdeep Pati

Abstract

Classically, Bayesian clustering interprets each component of a mixture model as a cluster. The inferred clustering posterior is highly sensitive to any inaccuracies in the kernel within each component. As this kernel is made more flexible, problems arise in identifying the underlying clusters in the data. To address this pitfall, this article proposes a fundamentally different approach to Bayesian clustering that decouples the problems of clustering and flexible modeling of the data density f . Starting with an arbitrary Bayesian model for f and a loss function for defining clusters based on f , we develop a Bayesian decision-theoretic framework for density-based clustering. Within this framework, we develop a Bayesian level set clustering method to cluster data into connected components of a level set of f . We provide theoretical support, including clustering consistency, and highlight performance in a variety of simulated examples. An application to astronomical data illustrates improvements over the popular DBSCAN algorithm in terms of accuracy, insensitivity to tuning parameters, and providing uncertainty quantification.

Keywords: Bayesian nonparametrics, DBSCAN, decision theory, density-based clustering, loss function, nonparametric density estimation

1. Introduction

In the Bayesian literature, when clustering is the goal, it is standard practice to model the data as arising from a mixture of unimodal probability distributions (Lau and Green, 2007; Wade and Ghahramani, 2018). The observations are then grouped according to their association with a mixture component. Bayesian clustering has potential advantages over algorithmic and frequentist approaches, providing natural hierarchical modeling, uncertainty quantification, and the ability to incorporate prior information (Wade, 2023). However, limitations appear in trying to apply the mixture model framework when clusters cannot

*. Joint first authors; MD is the corresponding author.

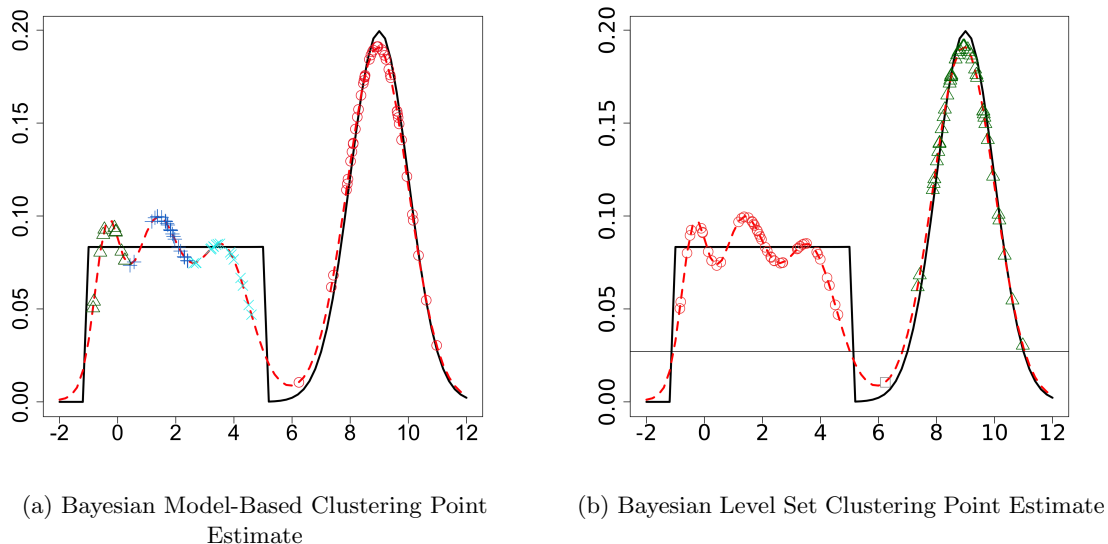


Figure 1: We see the cluster splitting phenomenon among the clusters obtained (left) by fitting a Dirichlet process mixture of Gaussian prior and finding the partition that minimizes expected VI loss under the posterior. Our Bayesian level set clustering (BALLET) point estimate based on the same prior (right) does not suffer from this phenomenon, despite the obvious bias in the posterior expectation of the density caused by the poor choice of prior distribution. We display a random subsample of the data in both plots, with their y-coordinates set to the expectation of the density at their locations, and with cluster assignments reflected by the color and shape of the points. The dashed red line is the expected density under the posterior. The solid line shown in black is the true data-generating density. The density level $\lambda = 0.028$ denoted by the horizontal line (right) was selected using an elbow heuristic in Section J (see Figure 28).

be well represented by simple parametric kernels. Even when clusters are *nearly* examples of simple parametric components, mixture model-based clustering can be brittle and result in *cluster splitting* (Miller and Dunson, 2019; Cai et al., 2021; Chaumeny et al., 2022). A potential solution is to use more flexible kernels (Malsiner-Walli et al., 2017). However, as the components are made more flexible, mixture models become difficult to fit and identify, since the multitude of reasonable models for a dataset tends to explode as the flexibility of the pieces increases (Ho and Nguyen, 2016, 2019).

Rather than avoid Bayesian clustering when the mixture approach fails, we propose decoupling the problems of modeling the data density and inferring clusters. Suppose that the data are drawn from the sample space $\mathcal{X}$, and denote by $\mathcal{D}(\mathcal{X})$ the density space on $\mathcal{X}$. Then, letting $\mathcal{P}(\mathcal{X})$ refer to the space of all possible partitions of $\mathcal{X}$, we can define functions $\Psi : \mathcal{D}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{X})$ that map from densities on $\mathcal{X}$ to partitions of $\mathcal{X}$. In the example in Figure 1 (b), $\Psi(f)$ was chosen as the partition of $\mathcal{X}$ induced by the connected components of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ at the $\lambda = 0.028$ level. Partitions of the sample space determine well-defined clusterings since, for any sample $\mathcal{X}_n = \{x_1, \dots, x_n\} \subseteq \mathcal{X}$, a partition of $\mathcal{X}$ induces

a partition on $\mathcal{X}_n$. For a particular Ψ and data set $\mathcal{X}_n$, we denote maps from the densities on $\mathcal{X}$ to the partition on $\mathcal{X}_n$ induced by Ψ with the lower case $\psi : \mathcal{D}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{X}_n)$. Here we have suppressed the dependence on the sample $\mathcal{X}_n$ to simplify the notation.

Let $D\{\psi(f), \mathcal{C}\}$ denote the loss for clustering $\mathcal{C} \in \mathcal{P}(\mathcal{X}_n)$ relative to clustering $\psi(f) \in \mathcal{P}(\mathcal{X}_n)$. If f_0 is the true data-generating density, then the target clustering is $\mathcal{C}_0 = \psi(f_0)$. In practice f_0 is unknown, so we represent uncertainty in the unknown density using a Bayesian posterior $f \sim P_M(\cdot | \mathcal{X}_n)$ based on the model M . This allows us to define a Bayesian decision-theoretic estimator $\hat{\psi}_M(\mathcal{X}_n)$, obtained by minimizing the expected posterior loss: $\hat{\psi}_M(\mathcal{X}_n) = \arg \min_{\mathcal{C} \in \mathcal{P}(\mathcal{X}_n)} E_{f \sim P_M(\cdot | \mathcal{X}_n)}[D\{\psi(f), \mathcal{C}\}]$, and to quantify the uncertainty in the clustering.

There is a substantial non-Bayesian literature on clustering based on the data-generating density f (Menardi, 2015; Campello et al., 2019; Bhattacharjee and Mitra, 2020). In this article, applying our decision-theoretic Bayesian paradigm for density-based clustering, we propose a new framework for Bayesian level set clustering. Level set clustering (Rinaldo and Wasserman, 2010; Sriperumbudur and Steinwart, 2012; Jiang, 2017; Jang and Jiang, 2019) is a popular approach that groups data points that fall into the same high-density region, while allowing these regions to have complex shapes. Our Bayesian approach has substantial advantages over current algorithmic approaches, such as DBSCAN (Ester et al., 1996; Schubert et al., 2017), which we will illustrate in various examples. Advantages include accuracy, less sensitivity to tuning parameters, and uncertainty quantification in clustering.

Our approach starts with the posterior under any nonparametric Bayesian model for f as the input, defines a loss function appropriate for level set clustering, and develops efficient algorithms for producing Bayes clustering estimates, while also providing a characterization of uncertainty in clustering. We develop supporting theory and demonstrate advantages over model-based and algorithmic level set clustering in various applications. The code for implementing our methodology is available at https://github.com/davidbuch/ballet_article and can be applied to data $\mathcal{X}_n$ and samples $f^{(1)}, \dots, f^{(s)}$ from the posterior distribution of f under any Bayesian model.

As a teaser motivating Bayesian level set clustering over a mixture-based approach, Figure 1 shows clusters produced by (a) a traditional Bayesian clustering approach and (b) our proposed approach. Here, the black line is the true density f_0 and both methods rely on fitting the same Dirichlet process mixture of Gaussians to the data to obtain a posterior for f . Although the use of Gaussian kernels leads to a noticeable bias in density estimation in the left mode of f_0 , our inferred level set clusters, which depend on the posterior distribution of the level set $\{x : f(x) \geq \lambda\}$ for our chosen level λ , are not affected by this. In contrast, an approach that equates clusters to mixture components sub-divides the uniform component into several subclusters. An interesting aspect of level set clustering is no attempt is made to cluster data points falling in low density regions; see Figure 4 for an example motivated by cosmology.

1.1 Contributions

The closest literature relevant to our work is that of Bayesian estimation of level sets of densities studied by Gayraud and Rousseau (2005, 2007) and the results in Li and Ghosal

(2021) on posterior contraction and credible regions for level curves. The frequentist estimation of level curves using bootstrap to characterize uncertainty is studied in Chen et al. (2017). Compared to the previous work on level set estimation, here we develop a practical method to compute a consistent Bayesian estimator of the induced clustering of the data and describe the associated uncertainty. Obtaining Bayesian clustering approaches that have appealing frequentist asymptotic properties is challenging under the predominant mixture model approach, particularly without making unrealistic assumptions such as correct kernel specification. Consequently, new Bayesian clustering methodologies based on the merging of components of an overfitted mixture of Gaussians (Dombowsky and Dunson, 2025; Aragam et al., 2020), the use of repulsive priors in the cluster means (Petrulia et al., 2012; Xie and Xu, 2020; Beraha et al., 2022) and the addition of entropic regularization (Franzolini and Rebaudo, 2024), have been proposed to improve the reliability of Bayesian clustering. With similar motivation, here we propose a Bayesian framework for *density-based clustering* (Menardi, 2015; Campello et al., 2019; Bhattacharjee and Mitra, 2020) that is consistent under suitable assumptions (Theorem 1). We show how the standard Bayesian decision-theoretic clustering machinery can be adapted to handle density-based clustering by modifying the loss function (8). Focusing on level set clustering, we leverage the current algorithmic and theoretical understanding (Schubert et al., 2017; Sriperumbudur and Steinwart, 2012) to implement our Bayesian level set clustering methodology **BALLET** and establish its consistency (Corollary 6). Finally, in illustrating the application of **BALLET** to various datasets, we discuss practical strategies to choose the level λ (Section J) and highlight the advantages offered by describing the clustering uncertainty associated with **BALLET** in a comprehensive analysis of astronomical sky survey data (Section 6).

2. Bayesian Level Set Clustering Methodology

2.1 Level Set Clusters and Sub-partitions

We start by expanding on the notational conventions of Section 1. Suppose that our data $\mathcal{X}_n = \{x_1, \dots, x_n\}$ are drawn independently from an unknown density $f_0 \in \mathcal{D}(\mathcal{X})$ on the sample space $\mathcal{X}$ taken to be $\mathbb{R}^d$ in much of this article, where $\mathcal{D}(\mathcal{X})$ denotes the space of densities on $\mathcal{X}$ with respect to the Lebesgue measure. Let $S_{\lambda, f_0} \doteq \{x \in \mathcal{X} : f_0(x) \geq \lambda\}$ denote the λ level set of f_0 , and temporarily let $W_1^{f_0}, \dots, W_{k^*}^{f_0}$ denote the *topologically connected components* of S_{λ, f_0} . In Figure 2, S_{λ, f_0} is the colored region on the x -axis, with colors corresponding to the different choices of λ indicated by the dashed lines. When $d = 1$, this region will either be a single interval $S_{\lambda, f_0} = W_1^{f_0}$ with $k^* = 1$, or more generally, be a union $S_{\lambda, f_0} = W_1^{f_0} \cup \dots \cup W_{k^*}^{f_0}$ of $k^* \in \{0, 1, 2, \dots\}$ disjoint intervals. The *level set clustering* $\mathcal{C}_0 = \psi_\lambda(f_0)$ of the data points $\mathcal{X}_n$ associated with f_0 is the collection $\mathcal{C}_0 = \{C_1^{f_0}, \dots, C_k^{f_0}\}$ of $k \leq k^*$ non-empty sets in $\{W_1^{f_0} \cap \mathcal{X}_n, \dots, W_{k^*}^{f_0} \cap \mathcal{X}_n\}$. For instance, the level set clustering corresponding to $\lambda = 0.1$ in Figure 2 is a grouping of data points $\mathcal{X}_n$ (not shown) based on whether they fall in a common blue interval or not. Data points that fall outside all of the blue intervals will be called *noise points*.

A level set clustering $\mathcal{C} = \{C_1, \dots, C_k\}$ of $\mathcal{X}_n$ is a *sub-partition*, since $C_i \cap C_j = \emptyset$ for all $i \neq j$ and $\cup_{i=1}^k C_i \subseteq \mathcal{X}_n$ but, unlike regular partitions, the presence of noise points not assigned to any cluster can lead to $\cup_{i=1}^k C_i \neq \mathcal{X}_n$. We call the observations in $A = \cup_{i=1}^k C_i$

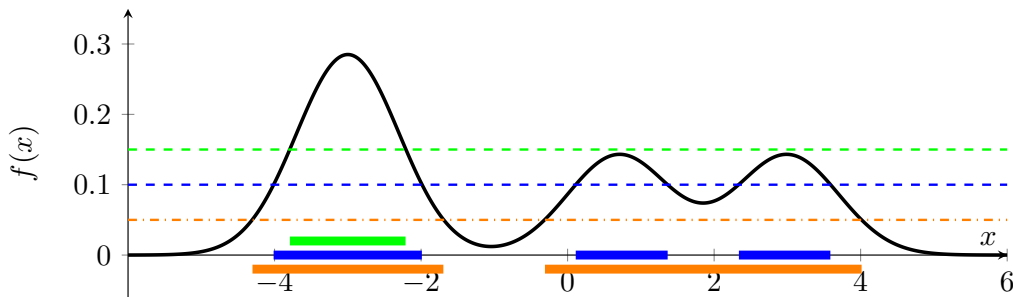


Figure 2: Topological connected components of the level set $\{x : f_0(x) \geq \lambda\}$ for a mixture f_0 (black curve) of Gaussians based on three colored choices for the level λ . Changing λ can result in discovery of anywhere from zero to three components (clusters).

as *active* or *core points*, while the remaining observations $I = \mathcal{X}_n \setminus A$ are *inactive* or *noise points*. In Figure 1(b) a noise point is shown in gray. Every sub-partition of size k with some noise points can be mapped to a unique partition of size $k + 1$, where the extra set in the partition consists of the noise points. However, this mapping is not one-to-one because the information on the identity of the noise cluster is lost (see example at the beginning of Section 2.5). Instead, to preserve information about noise points, we explicitly work with the non-standard setup of regarding a clustering as a sub-partition rather than a partition. To this end, we repurpose the notation $\mathcal{P}(\mathcal{X}_n)$ to denote the space of all *sub-partitions* of $\mathcal{X}_n$. Note that this is a strict expansion since $\mathcal{P}(\mathcal{X}_n)$ also contains all partitions of $\mathcal{X}_n$.

2.2 On the interpretation of level set clusters and the choice of level λ

Level set clustering is primarily meant to discover connected regions of high (population) density separated by regions of low density, and the parameter λ determines what ‘high’ means here. While a reasonable choice of λ may be apparent in certain applications (see Section 6), we now discuss strategies from the literature when this is not the case.

When the clusters are expected to be well-separated from each other (e.g. Figures 1 and 3), simple strategies to tune λ based on elbow plots (Ester et al., 1996) and deciding on a small fraction ($\nu = \int f(x)1_{\{f(x) < \lambda\}} dx$) of noise points in advance (Cuevas et al., 2001) are useful and robust. See Section J for our implementation.

In general, however, as seen in Figure 2, care is needed to select the level λ and in some cases a single appropriate λ does not exist (see Figure 31 and Menardi (2015); Campello et al. (2019)). In such scenarios, one should study the *cluster-tree* (Campello et al., 2015; Wang et al., 2019; Steinwart et al., 2023) obtained by running level set clustering across a range of values of $\lambda > 0$. It is common to visualize (Zappia and Oshlack, 2018) and process (Campello et al., 2015; Scrucca, 2016) this tree to extract clusters that remain stable across a range of values of λ . This motivates our persistent clustering implementation in Section K.

2.3 Decision-Theoretic Framework

We focus on finding the sub-partition of data $\mathcal{X}_n$ associated with the connected components of S_{λ, f_0} . We let $\psi_\lambda : \mathcal{D}(\mathcal{X}) \mapsto \mathcal{P}(\mathcal{X}_n)$ be the *level- λ clustering function*, by which we

mean that $\psi_\lambda(f)$ returns the sub-partition $\mathcal{C}$ of $\mathcal{X}_n$ associated with the level- λ connected components of f .

We start by choosing a Bayesian model M for the unknown density f . Examples of M include not only kernel mixture models but also Bayesian nonparametric approaches that do not involve a latent clustering structure, such as Polya trees (Lavine, 1992; Ma, 2017) and logistic Gaussian processes (Lenk, 1991; Tokdar, 2007). Under M , we obtain a posterior distribution $P_M(f|\mathcal{X}_n)$ for the unknown density of the data. This also induces a posterior on the λ level set of f . Based on this posterior, we define $\hat{\psi}_{\lambda,M}$ as an estimator of $\psi_\lambda(f_0)$.

Let $D\{\psi_\lambda(f), \mathcal{C}\}$ denote a loss function measuring the quality of sub-partition $\mathcal{C}$ relative to the ground truth $\psi_\lambda(f)$. The Bayes estimator (e.g. Berger, 2013, Section 4.4.1) of the sub-partition then corresponds to the value that minimizes the expectation of the loss under the posterior of f :

$$\hat{\psi}_{\lambda,M}(\mathcal{X}_n) \doteq \arg \min_{\mathcal{C} \in \mathcal{P}(\mathcal{X}_n)} E_{f \sim P_M(\cdot|\mathcal{X}_n)}[D\{\psi_\lambda(f), \mathcal{C}\}]. \quad (1)$$

In practice, we use a Monte Carlo approximation based on samples $f^{(1)}, \dots, f^{(S)}$ from $P_M(f|\mathcal{X}_n)$: $\hat{\psi}_{\lambda,M}(\mathcal{X}_n) \approx \arg \min_{\mathcal{C} \in \mathcal{P}(\mathcal{X}_n)} \sum_{s=1}^S D\{\psi_\lambda(f^{(s)}), \mathcal{C}\}$.

Three major roadblocks stand in the way of calculating this estimator. First, evaluating $\psi_\lambda(f^{(s)})$ is problematic, as identifying connected components of level sets of $f^{(s)}$ is extremely costly if the data are in even a moderately high-dimensional space. Instead, we will use a surrogate clustering function $\tilde{\psi}_\lambda$, which approximates the true clustering function and is more tractable. We will discuss this in more detail in Section 2.4.

The second roadblock is the fact that we must design an appropriate loss function D to use in estimating the level set clustering. Since these objects are sub-partitions, usual loss functions on partitions that are employed in model-based clustering will be inappropriate. We will discuss the issue further and introduce an appropriate loss in Section 2.5.

Finally, optimizing the risk function over the space of all sub-partitions, as shown in Equation (1), will be computationally intractable, since the number of elements in $\mathcal{P}(\mathcal{X}_n)$ is immense. However, leveraging on the current Bayesian clustering literature, we adapt the discrete optimization algorithm of Dahl et al. (2022) to handle our case of sub-partitions.

Having addressed these issues, we refer to the resulting class $\{\hat{\psi}_{\lambda,M}\}$ as Bayesian level set (BALLET) estimators. In Section 4 we show that, under suitable models M for density f , the BALLET estimator $\hat{\psi}_{\lambda,M}$ consistently estimates the level- λ clustering based on f_0 .

2.4 Surrogate Clustering Function

Computing the clustering function $\psi_\lambda(f)$ based on the level set $S_{\lambda,f} = \{x \in \mathcal{X} : f(x) \geq \lambda\}$ involves two steps. The first identifies the subset of observations $A_{\lambda,f} = S_{\lambda,f} \cap \mathcal{X}_n$, called the active points for f , and the second separates the active points according to the topologically connected components of $S_{\lambda,f}$. The first step is no more difficult than evaluating f at each of the n observations and checking whether $f(x_i) \geq \lambda$ for $i \in \{1, \dots, n\}$. However, identifying the connected components of $S_{\lambda,f}$ can be computationally intractable unless $\mathcal{X}$ is one-dimensional. This is a familiar challenge in algorithmic level set clustering (Campello et al., 2019).

A common approach with theoretical support (Devroye and Wise, 1980; Rinaldo and Wasserman, 2010; Sriperumbudur and Steinwart, 2012) is to approximate the level set $S_{\lambda,f}$

with a tube of diameter $\delta > 0$ around the active points: $T_\delta(A) = \cup_{x_i \in A} B(x_i, \delta/2)$, where $B(x, \delta/2)$ is the open ball of radius $\delta/2$ around x and $A = A_{\lambda, f}$ denotes the active points. Calculating the connected components of $T_\delta(A)$ is straightforward. If we define $G_\delta(A)$ as the δ -neighborhood graph with vertices A and edges $\{(x, x') \in A \times A \mid \|x - x'\| < \delta\}$, then two points $x, x' \in \mathcal{X}$ lie in the same connected component of $T_\delta(A)$ if and only if there exist active points $x_i, x_j \in A$ such that $\|x - x_i\| < \frac{\delta}{2}$, $\|x' - x_j\| < \frac{\delta}{2}$ and x_i, x_j are connected by a path in $G_\delta(A)$. The problem is further simplified since we only need to focus on the active points: Any $x_i, x_j \in A$ lies in the same connected component of $T_\delta(A)$ if and only if x_i, x_j are connected by a path in $G_\delta(A)$. Lemma 13 in Section I.3 provides more details.

Hence, we define a computationally-tractable surrogate clustering function

$$\tilde{\psi}_{\delta, \lambda}(f) = \text{CC}\{G_\delta(A_{\lambda, f})\} \quad (2)$$

where the dependence on the density f and level λ enter through the active points $A_{\lambda, f} = \{x \in \mathcal{X}_n \mid f(x) \geq \lambda\}$, and CC is the function that maps graphs to the *graph-theoretic* connected components of their vertices (Dasgupta et al., 2008, Chapter 3).

In Section C, we discuss how the DBSCAN clustering algorithm (Ester et al., 1996; Schubert et al., 2017) essentially corresponds to evaluating $\psi_{\delta, \lambda}(\hat{f})$ for a certain density estimator $\hat{f}$ of f_0 . In fact, for a general f , the computational complexity of evaluating $\tilde{\psi}_{\delta, \lambda}(f)$ is comparable to that of DBSCAN with the additional cost of evaluating f at the data points $\mathcal{X}_n$.

Compared to the clustering point estimate $\tilde{\psi}_{\delta, \lambda}(\hat{f})$ obtained by inserting a density estimator $\hat{f}$ based on $\mathcal{X}_n$, the main motivation behind our Bayesian clustering machinery of eq. (1) is to account for the variability of $\tilde{\psi}_{\delta, \lambda}(f)$ in the posterior distribution of f . We expect our Bayesian point estimate of eq. (1) to be more reliable than $\tilde{\psi}_{\delta, \lambda}(\hat{f})$ in difficult level set clustering problems involving substantial uncertainty in density estimation.

Our clustering $\psi_{\delta, \lambda}(f)$ depends on the choice of the parameter $\delta > 0$. For some $k \in \mathbb{N}$, $\gamma \in [0, 1)$, and an estimate $\hat{f}$ of f_0 , we suggest the data adaptive value of

$$\hat{\delta} = q_{1-\gamma}\{\delta_k(x_i) : x_i \in A_{\lambda, \hat{f}}\}, \quad (3)$$

the $1 - \gamma$ quantile of the k -nearest neighbor distance $\delta_k(x)$ among the estimated active data points $A_{\lambda, \hat{f}}$, with our default choice of $\gamma = 0.01$. The intuition here is that the value $\hat{\delta}$ will be smaller than the required distance between disjoint level λ clusters of f_0 if the k -closest data points to most ($> 99\%$) of the active points are known to belong to the same cluster as the initial point. The choice of k here also needs to be large enough to ensure that the level λ cluster of f_0 is not disconnected by the skeleton graph $G_{\hat{\delta}}(A_{\lambda, \hat{f}})$. Noting that the performance of BALLET clustering was not sensitive to our choice of k (e.g. Figure 19), we use the default value of $k = \lceil \log n \rceil$ in our analysis. In Section 4.2.2, we theoretically study the accuracy of approximating $\psi_\lambda(f_0)$ by $\tilde{\psi}_{\delta, \lambda}(\hat{f})$. For suitably large $C > 0$, as long as $k \in [C \log n, n/C]$ and $\gamma < 1$, using $\delta = \hat{\delta}$ from (3) will lead to consistent BALLET clustering with high probability (Theorem 4 and Lemma 5).

2.5 Loss Function for Comparing Sub-partitions

In order for (1) to have the interpretation of a posterior Fréchet mean, $D : \mathcal{P}(\mathcal{X}_n) \times \mathcal{P}(\mathcal{X}_n) \rightarrow [0, \infty]$ must be chosen to be a metric on the space of sub-partitions $\mathcal{P}(\mathcal{X}_n)$.

While any standard loss function on partitions (Dahl et al., 2022) has a natural extension to sub-partitions, this does not result in a metric on $\mathcal{P}(\mathcal{X}_n)$. For example, consider the popular Binder’s loss L_{Binder} which is a metric on the space of partitions (Binder, 1978; Wade and Ghahramani, 2018). Given a subset $C \subseteq \mathcal{X}_n$ and its complement $C' = \mathcal{X}_n \setminus C$, what should be the resulting loss between the sub-partitions $\mathcal{C} = \{C\}$ and $\mathcal{C}' = \{C'\}$? While $\mathcal{C}$ and $\mathcal{C}'$ are incredibly different when considered as level set clustering, the induced partitions are the same resulting in the loss $L_{\text{Binder}}(\{C, \mathcal{X}_n \setminus C\}, \{C', \mathcal{X}_n \setminus C'\}) = 0$.

Instead we now propose a modification of the Binder’s loss, which will be a metric on the space of sub-partitions $\mathcal{P}(\mathcal{X}_n)$. Our Inactive/Active (IA) Binder’s loss takes the form of Binder’s loss for data points that are active in both partitions, with a penalty for points active in one partition and inactive in the other. We represent any sub-partition $\mathcal{C} = \{C_1, \dots, C_k\} \in \mathcal{P}(\mathcal{X}_n)$ with a length n allocation vector $\vec{c} = (c_1, \dots, c_n) \in \{0, 1, \dots, k\}^n$ such that $c_i = h$ if $x_i \in C_h$ and $c_i = 0$ if $x_i \in \mathcal{X}_n \setminus \cup_{h=1}^k C_h$. Given two partitions $\mathcal{C}, \mathcal{C}'$ with active sets $A, A' \subseteq \mathcal{X}_n$ and allocation vectors $\vec{c}, \vec{c}'$, the loss between them is defined as

$$\begin{aligned} L_{\text{IA-Binder}}(\mathcal{C}, \mathcal{C}') &= (n-1)(m_{ai} |A \cap I'| + m_{ia} |I \cap A'|) + \sum_{\substack{1 \leq i < j \leq n \\ x_i, x_j \in A \cap A'}} a \mathbb{1}_{(c_i=c_j; c'_i \neq c'_j)} + b \mathbb{1}_{(c_i \neq c_j; c'_i=c'_j)}, \end{aligned} \quad (4)$$

where $I = \mathcal{X}_n \setminus A$ and $I' = \mathcal{X}_n \setminus A'$ denote the inactive sets of $\mathcal{C}$ and $\mathcal{C}'$. The loss is a well-defined function of $\mathcal{C}$ and $\mathcal{C}'$ since the right-hand side is invariant to any permutation of the active labels in $\vec{c}$ and $\vec{c}'$. The summation term is the Binder’s loss with parameters $a, b > 0$ restricted to points active in both sub-partitions. The first two terms, based on parameters $m_{ai}, m_{ia} > 0$, correspond to a loss of $(n-1)m_{ai}$ and $(n-1)m_{ia}$ incurred by points that are active in $\mathcal{C}$ but inactive in $\mathcal{C}'$ and vice versa. We focus mainly on the setting where $a = b$ and $m_{ai} = m_{ia} = m \geq a/2$ with our default choice of $a = b = 1$ and $m = 1/2$ used throughout our analysis. Under these conditions Theorem 2 in Section 4.2 shows that $L_{\text{IA-Binder}}$ is a metric on $\mathcal{P}(\mathcal{X}_n)$. Our starting point is Remark 3, which provides an alternate representation of this loss.

Given *any* distribution on $\mathcal{C}$, we can compute the Bayes risk for an estimate $\mathcal{C}'$ as the posterior expectation of the IA-Binder’s loss:

$$\begin{aligned} R_{\text{IA-Binder}}(\mathcal{C}') &= E\{L_{\text{IA-Binder}}(\mathcal{C}, \mathcal{C}')\} \\ &= (n-1) \left\{ m_{ai} \sum_{i=1}^n \Pr(x_i \in A) \mathbb{1}_{(x_i \in I')} + m_{ia} \sum_{i=1}^n \Pr(x_i \in I) \mathbb{1}_{(x_i \in A')} \right\} + \\ &\quad \sum_{1 \leq i < j \leq n} \mathbb{1}_{(x_i \in A', x_j \in A')} \left\{ a \Pr(x_i \in A, x_j \in A, c_i = c_j) \mathbb{1}_{(c'_i \neq c'_j)} + \right. \\ &\quad \left. b \Pr(x_i \in A, x_j \in A, c_i \neq c_j) \mathbb{1}_{(c'_i = c'_j)} \right\}. \end{aligned} \quad (5)$$

The probabilities are computed based on the random clustering $\mathcal{C} = \tilde{\psi}_{\delta,\lambda}(f)$, where f is drawn from the posterior $P_M(\cdot|\mathcal{X}_n)$. Our **BALLET** estimator for level- λ clustering is then

$$\begin{aligned}\hat{\psi}_{\delta,\lambda,M}(\mathcal{X}_n) &\doteq \arg \min_{\mathcal{C}' \in \mathcal{P}(\mathcal{X}_n)} E_{f \sim P_M(\cdot|\mathcal{X}_n)} [L_{\text{IA-Binder}}\{\tilde{\psi}_{\delta,\lambda}(f), \mathcal{C}'\}] \\ &\approx \arg \min_{\mathcal{C}' \in \mathcal{P}(\mathcal{X}_n)} \sum_{s=1}^S L_{\text{IA-Binder}}\{\tilde{\psi}_{\delta,\lambda}(f^{(s)}), \mathcal{C}'\},\end{aligned}\tag{6}$$

where the dependence of the estimator on the data is mediated by the posterior distribution $P_M(\cdot|\mathcal{X}_n)$ from which we generate samples $f^{(1)}, \dots, f^{(S)}$. We precompute Monte Carlo estimates of the probabilities appearing in equation (5). Then, estimating $\hat{\psi}_{\delta,\lambda,M}(\mathcal{X}_n)$ is based on optimizing the objective function. We rely on a modification of the algorithm of Dahl et al. (2022) described in Section D.

When the posterior uncertainty in f is small, one may use a heuristic **BALLET plugin** estimate $\hat{\mathcal{C}} = \tilde{\psi}_{\delta,\lambda}(\hat{f})$ that avoids the expensive optimization in (6) by directly computing the level set clusters of the posterior mean density $\hat{f}(x) \approx \frac{1}{S} \sum_{s=1}^S f^{(s)}(x)$. While in many instances we found the **BALLET** plugin estimate to have similar performance to our **BALLET** estimator (6) (e.g. Tables 1 to 3), the two estimates can be different (Figure 8). Although Bayes estimators that directly target the quantity of interest are more principled (Section D.1), the two-stage plugin approach represents a practical alternative for point estimation.

3. Credible Bounds

In addition to a clustering point estimate, we characterize the uncertainty. One popular strategy in Bayesian clustering is to examine the $n \times n$ posterior similarity matrix, whose i, j th entry contains the co-clustering probability $\Pr(c_i = c_j|\mathcal{X}_n)$. Such summaries are complicated in our case by the fact that the entry i and/or j may be inactive. An appealing alternative is to adapt the method of Wade and Ghahramani (2018) to compute credible balls for level set sub-partitions.

To find a credible ball around the point estimate $\hat{\mathcal{C}}$ with credible level $1 - \alpha$ for $\alpha \in [0, 1]$, we first find

$$\epsilon^* \doteq \arg \min_{\epsilon > 0} P_M\{\tilde{\psi}_{\delta,\lambda}(f) \in B_\epsilon(\hat{\mathcal{C}})|\mathcal{X}_n\} \geq 1 - \alpha,\tag{7}$$

the smallest radius $\epsilon = \epsilon^*$ such that the ball $B_\epsilon(\hat{\mathcal{C}}) = \{\mathcal{C}' \in \mathcal{P}(\mathcal{X}_n) : L_{\text{IA-Binder}}(\hat{\mathcal{C}}, \mathcal{C}') \leq \epsilon\}$ of radius ϵ around $\hat{\mathcal{C}}$ has a posterior coverage probability of at least $1 - \alpha$. Then, the posterior distribution will assign a posterior probability close to $1 - \alpha$ to the event that $B_{\epsilon^*}(\hat{\mathcal{C}})$ contains $\mathcal{C} = \tilde{\psi}_{\delta,\lambda}(f)$, the unknown level set sub-partition.

The $1 - \alpha$ coverage credible ball $B_{\epsilon^*}(\hat{\mathcal{C}})$ typically contains a large number of possible sub-partitions. To summarize credible balls in the space of data partitions, Wade and Ghahramani (2018) recommend identifying *vertical* and *horizontal* bounds based on the partial ordering of partitions associated with a Hasse diagram. The vertical upper bounds were defined as the partitions in $B_{\epsilon^*}(\hat{\mathcal{C}})$ that contained the smallest number of sets; vertical lower bounds, accordingly, were the partitions in $B_{\epsilon^*}(\hat{\mathcal{C}})$ that contained the largest number of sets; horizontal bounds were the partitions in $B_{\epsilon^*}(\hat{\mathcal{C}})$ that were the *farthest* from $\hat{\mathcal{C}}$ at distance $L_{\text{IA-Binder}}$.

In our setting, in addition to similarity of sub-partitions in terms of their clustering structure, we must also compare inclusion or exclusion of observations from the active set. Uncertainty in the clustering structure will be partly attributable to uncertainty in which points are active. Fortunately, the space of sub-partitions is a lattice with an associated Hasse diagram (Section B). We can move *down* the sub-partition lattice by splitting clusters or removing items from the active set, while we can move *up* the lattice of sub-partitions by merging clusters or absorbing noise points into the active set.

We propose the following computationally efficient algorithm for computing upper and lower bounds for the credible ball. Suppose we know our credible ball radius ϵ^* from Equation (7) needed to achieve the desired coverage. We seek our upper bound by starting at the point estimate and greedily adding to the active set, one at a time, the item from the inactive set that has the greatest posterior probability of being active and reexamining the resulting connected components; this continues until we find a sub-partition that is farther than ϵ^* from the point estimate. To find a lower bound, we perform the analogous greedy removal process. The resulting bounds from applying this algorithm can be seen in Figures 5, 13, 14 and 21.

4. Consistency theory

In Section 4.1 we develop a general consistency theorem for Bayesian density-based clustering under three intuitive assumptions. Next in Section 4.2, we carefully apply this result to our **BALLET** estimator $\hat{\psi}_{\delta,\lambda,M}(\mathcal{X}_n)$ from (6) and derive mild conditions under which our method will be consistent. In the process, we provide theoretical guarantees on the accuracy of our surrogate clustering function from Section 2.4, indicating the choices of the parameter δ that lead to consistent estimation of level- λ clusters. Indeed, our data adaptive choice of $\hat{\delta}$ in (3) will be seen to satisfy this condition under suitable assumptions.

4.1 A general consistency result for Bayesian density-based clustering

In this section we show asymptotic consistency of a generic Bayesian density-based clustering estimator of the form

$$\hat{\psi}_M(\mathcal{X}_n) = \arg \min_{\mathcal{C} \in \mathcal{P}(\mathcal{X}_n)} E_{f \sim P_M(\cdot|\mathcal{X}_n)}[D\{\tilde{\psi}(f), \mathcal{C}\}], \quad (8)$$

where D is a loss on the space $\mathcal{P}(\mathcal{X}_n)$ of data sub-partitions and $\tilde{\psi} : \mathcal{D}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{X}_n)$ is an easy-to-compute surrogate that approximates the target density-based clustering function $\psi : \mathcal{D}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{X}_n)$. Similar to previous sections, we omit notation for the implicit dependence of D , $\tilde{\psi}$, and ψ on $\mathcal{X}_n$ and n . We will assume that the loss D is a metric that takes values in $[0, 1]$. We state our consistency result in terms of convergence in probability. Recall that a sequence of random variables $\{X_n\}_{n \geq 1}$ converges to zero in probability, denoted by $X_n \xrightarrow{P} 0$ as $n \rightarrow \infty$, if $\lim_{n \rightarrow \infty} \Pr(|X_n| > \epsilon) = 0$ for every fixed $\epsilon > 0$.

Under some mild assumptions stated later, the following theorem establishes consistency of the estimator (8). In particular, when data $\mathcal{X}_n$ are generated independently from f_0 , it states that the Bayesian density-based clustering estimator defined in (8) will be close to the target clustering $\psi(f_0)$ in terms of loss D for large values of n .

Theorem 1. (*Consistency of density-based clustering*) Suppose that Assumptions 1 to 3 stated below hold, and $\mathcal{X}_n = \{x_1, \dots, x_n\} \stackrel{i.i.d.}{\sim} f_0$. Then

$$0 \leq D\{\hat{\psi}_M(\mathcal{X}_n), \psi(f_0)\} \leq 2\tau_1(\mathcal{X}_n) + 2\tau_2(\mathcal{X}_n) \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty,$$

where $\hat{\psi}_M(\mathcal{X}_n)$ is the density-based clustering estimate (8) and the error terms τ_1 and τ_2 are as defined in Assumptions 2 and 3.

We now discuss the assumptions underlying Theorem 1. The proof of Theorem 1, provided in Section I.1, captures the intuition that as long as the posterior distribution of f is concentrated around f_0 in terms of a metric ρ on $\mathcal{D}(\mathcal{X})$ (Assumption 2) that can guarantee that the two clusterings $\hat{\psi}(f)$ and $\psi(f_0)$ are close (Assumption 3), then our Bayesian density-based clustering estimator $\hat{\psi}_M(\mathcal{X}_n)$ will also be close to $\psi(f_0)$ by using the triangle inequality for D (Assumption 1).

Assumption 1. Suppose that $D : \mathcal{P}(\mathcal{X}_n) \times \mathcal{P}(\mathcal{X}_n) \rightarrow [0, 1]$ is a metric.

Next, we assume that the Bayesian model M for the unknown density f is such that its posterior distribution $P_M(\cdot | \mathcal{X}_n)$, under samples $\mathcal{X}_n = \{x_1, \dots, x_n\}$ drawn independently from f_0 , contracts at rate ϵ_n to f_0 in some metric ρ on the space of densities $\mathcal{D}(\mathcal{X})$.

Assumption 2 (Posterior contraction). If $\mathcal{X}_n = \{x_1, \dots, x_n\}$ are drawn independently from f_0 , then there is a metric ρ on $\mathcal{D}(\mathcal{X})$ and there is a non-negative sequence of numbers $\{\epsilon_n\}_{n \geq 1}$ converging to zero such that

$$\tau_1(\mathcal{X}_n) \doteq P_M(f : \rho(f, f_0) \geq \epsilon_n K_n | \mathcal{X}_n) \xrightarrow{P} 0 \text{ as } n \rightarrow \infty,$$

for every non-negative sequence $\{K_n\}_{n \in \mathbb{N}}$ that diverges to infinity.

Assumption 3. There is a non-negative sequence $\{K_n\}_{n \in \mathbb{N}}$ that diverges to infinity such that $\tau_2(\mathcal{X}_n) \doteq \sup_{f \in \mathcal{D}(\mathcal{X}) : \rho(f, f_0) \leq K_n \epsilon_n} D\{\tilde{\psi}(f), \psi(f_0)\} \xrightarrow{P} 0$ as $n \rightarrow \infty$, where ρ and ϵ_n are as given in Assumption 2.

Assumptions 2 and 3 are related in that we need a common sequence $\{(\epsilon_n, K_n)\}_{n \geq 1}$ and the same metric ρ on $\mathcal{D}(\mathcal{X})$ such that both Assumptions 2 and 3 hold. Standard posterior contraction results (e.g. Ghosal and van der Vaart, 2017, Chapter 9) can establish the condition in Assumption 2 for various models M and suitable rates $\epsilon_n \rightarrow 0$ when ρ is the Hellinger or total-variation metric on $\mathcal{D}(\mathcal{X})$. However, here one may need contraction in a stronger metric ρ on $\mathcal{D}(\mathcal{X})$ to ensure continuity of the clustering functional $\psi : \mathcal{D}(\mathcal{X}) \rightarrow \mathcal{P}(\mathcal{X}_n)$ to guarantee Assumption 3 even when $\tilde{\psi} = \psi$. For example, for our application to level set clustering we will use the L^∞ metric $\rho(f, g) = \|f - g\|_\infty \doteq \sup_{x \in \mathcal{X}} |f(x) - g(x)|$ in Section 4.2. Similarly, we expect to use a metric ρ that captures uniform convergence of both the density f and its derivatives to satisfy Assumption 3 when ψ describes modal clustering (see the introduction of Shen and Ghosal, 2017). Thus establishing posterior contraction results in stronger metrics ρ than the standard Hellinger distance is a promising active area of research (Giné and Nickl, 2011; Castillo, 2014, 2017; Naulet, 2022; Shen and Ghosal, 2017; Li and Ghosal, 2021) that can help establish consistency of Bayesian density-based clustering.

4.2 Application to level set clustering

Note that (6) represents a special case of (8), when $\tilde{\psi} = \tilde{\psi}_{\delta,\lambda}$ is the surrogate clustering function defined in (2), $\psi = \psi_\lambda$ is the level- λ clustering function defined in Section 2.3, and $D = \binom{n}{2}^{-1} L_{\text{IA-Binder}}$ is a rescaled version of the Inactive-Active Binder loss (4). We will fix this choice of $\psi, \tilde{\psi}$ and D throughout this section. We show that Assumptions 1 to 3 are satisfied for suitable choices of the parameter $\delta > 0$ and suitable conditions on the density f_0 , level $\lambda > 0$, and model M . Following the existing level set clustering theory (e.g. Sriperumbudur and Steinwart, 2012; Rinaldo and Wasserman, 2010; Jiang, 2017), in this section we will use the L^∞ metric $\rho(f, g) = \|f - g\|_\infty = \sup_{x \in \mathcal{X}} |f(x) - g(x)|$ on $\mathcal{D}(\mathcal{X})$ in Assumptions 2 and 3. While posterior consistency of the density f in the L^∞ metric is a strong requirement, Remark 7 briefly discusses how this requirement might be weakened.

4.2.1 PROPERTIES OF IA-BINDER'S LOSS

To establish the validity of Assumption 1 we study the properties of our Inactive-Active Binder loss (4). The following theorem proved in Section I.2 shows that Assumption 1 is satisfied for suitable choices of constants in our Inactive-Active Binder loss (4).

Theorem 2. *Suppose $0 < a = b \leq 1$, $m = m_{ia} = m_{ai} \leq 1$, and $a \leq 2m$. Then $D = \binom{n}{2}^{-1} L_{\text{IA-Binder}}$ is a metric on $\mathcal{P}(\mathcal{X}_n)$ that is bounded above by 1.*

The following remark, which will be useful to interpret the conclusion of Theorem 1, describes when the distance D between two sub-partitions $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{P}(\mathcal{X}_n)$ will be small.

Remark 3. *We say that a pair of distinct points $x_i, x_j \in \mathcal{X}_n$ is clustered differently by $\mathcal{C}_1$ and $\mathcal{C}_2$ if the activity status of either x_i or x_j is different across $\mathcal{C}_1$ and $\mathcal{C}_2$, or else both x_i and x_j are active in both $\mathcal{C}_1$ and $\mathcal{C}_2$ but the two points belong to the same cluster in $\mathcal{C}_1$ (or $\mathcal{C}_2$) but to different clusters in $\mathcal{C}_2$ (or $\mathcal{C}_1$). Importantly, $L_{\text{IA-Binder}}$ can be expressed as a sum of non-negative penalties over distinct pairs of points from $\mathcal{X}_n$*

$$L_{\text{IA-Binder}}(\mathcal{C}_1, \mathcal{C}_2) = \sum_{1 \leq i < j \leq n}^n \phi_{i,j},$$

where the penalty $\phi_{i,j} \in \{0, a, m, 2m\}$ takes a positive value of at least $\min(a, m)$ when the pair x_i, x_j is clustered differently by $\mathcal{C}_1$ and $\mathcal{C}_2$. (See (11) in Section I.2 for exact details.) Thus for the choice of $a, m \in [1/2, 1]$ and any $\epsilon \in (0, 1/2)$, if the rescaled loss $D(\mathcal{C}_1, \mathcal{C}_2) = \binom{n}{2}^{-1} L_{\text{IA-Binder}}(\mathcal{C}_1, \mathcal{C}_2)$ is less than ϵ then at most 2ϵ fraction of all pairs of points from $\mathcal{X}_n$ will be clustered differently by $\mathcal{C}_1$ and $\mathcal{C}_2$. Conversely, if at most ϵ fraction of all pairs of points from $\mathcal{X}_n$ are clustered differently by $\mathcal{C}_1$ and $\mathcal{C}_2$ then $D(\mathcal{C}_1, \mathcal{C}_2) < 2\epsilon$.

4.2.2 ACCURACY OF OUR LEVEL-SET CLUSTERING SURROGATE

We now examine Assumption 3 here, while Assumption 2 will be examined in Section 4.2.3.

The following result, proved in Section I.3, demonstrates that Assumption 3 will be satisfied as long as the density f_0 satisfies some mild conditions and $\gamma = K_n \epsilon_n \rightarrow 0$. Generally speaking, we require that $f_0 : \mathbb{R}^d \rightarrow [0, \infty)$ is continuous and vanishing in the tails (Assumption 4), is not flat around the level λ (Assumption 5), and has a level- λ clustering that

is stable with respect to small perturbations in λ (Assumption 6). Under these conditions, with high-probability our surrogate clustering estimator $\tilde{\psi}_{\delta,\lambda}(f)$ from Section 2.4 will be close to the true clustering $\psi_\lambda(f_0)$ in terms of our distance D as long as f is close to f_0 in the L^∞ metric and δ lies in a suitable range.

Theorem 4. *Suppose $\mathcal{X} = \mathbb{R}^d$ and the density f_0 and the level $\lambda > 0$ satisfy Assumptions 4 to 6 in Section I.3. Suppose further that f_0 is α -Hölder continuous for some $\alpha \in (0, 1]$, the dataset $\mathcal{X}_n = \{x_1, \dots, x_n\}$ is drawn independently from f_0 with $n \geq 16$, and D is the re-scaled loss in Theorem 2. Then, depending on f_0 , there are finite constants $C_0, \bar{\delta}, \bar{\gamma} > 0$ such that*

$$\sup_{f: \|f - f_0\|_\infty \leq \gamma} D\{\tilde{\psi}_{\delta,\lambda}(f), \psi_\lambda(f_0)\} \leq C_0 \left\{ \max(\gamma, \delta^\alpha) + \sqrt{\frac{\ln n}{n}} \right\}$$

holds uniformly over all $\delta \in [r_{n,\lambda,d}, \bar{\delta})$ and $\gamma \in (0, \bar{\gamma})$ with probability at least $1 - \frac{1+n}{n^2}$. Here $r_{n,\lambda,d} \doteq 2 \left(\frac{16d \ln n}{nv_d \lambda} \right)^{1/d}$ where v_d is the volume of the unit Euclidean ball in d dimensions.

The constraint $\delta \geq r_{n,\lambda,d}$ in Theorem 4 ensures that, with high probability, every open ball $B(x, \delta/2)$ contained in S_{λ,f_0} will also contain at least one data point $x_i \in \mathcal{X}_n \cap B(x, \delta/2)$. This key result is used in Lemma 10 to show that the level set estimator $T_\delta(A_{f,\lambda})$ from Section 2.4 will be suitably close to S_{λ,f_0} when $\|f - f_0\|_\infty$ and δ are small (and $\delta \geq r_{n,\lambda,d}$). The following lemma proved in Section I.4 shows that our data adaptive choice of $\hat{\delta}$ in (3) will satisfy conditions of Theorem 4 with high probability if $\log n \ll k \ll n$ as $n \rightarrow \infty$.

Lemma 5. *Suppose the assumptions of Theorem 4 are satisfied and the density estimator $\hat{f}$ satisfies $\|\hat{f} - f\|_\infty \leq \lambda/2$. Then there is a finite constant $L > 0$ depending on f_0 and λ such that if $k \in [L \ln n, n/L]$ then $\hat{\delta} \in [r_{n,\lambda,d}, \bar{\delta})$ with probability at least $1 - 2e^{-\frac{1}{32}\sqrt{\frac{k}{d \ln n}}}$.*

4.2.3 CONSISTENCY OF LEVEL SET CLUSTERING

Assumption 2 requires posterior contraction around f_0 in the L^∞ norm. While such contraction results can be obtained when the model M is based on a parametric family that contains f_0 , the search for such results when M is a non-parametric model is currently an active area of research. For univariate density estimation on $\mathcal{X} = [0, 1]$, such contraction rates have been established for kernel mixture models, random histogram priors, Pólya trees, Gaussian process and wavelet series priors on the log density (Giné and Nickl, 2011; Castillo, 2014, 2017; Naulet, 2022). For multivariate density estimation on $\mathcal{X} = [0, 1]^d$, refer to Li and Ghosal (2021) and references therein.

Combining all the results in this section leads to the following corollary of Theorem 1.

Corollary 6. *Suppose $\mathcal{X} = \mathbb{R}^d$, density $f_0 \in \mathcal{D}(\mathcal{X})$ and level $\lambda > 0$ satisfy Assumptions 4 to 6 in Section I.3, and data $\mathcal{X}_n = \{x_1, \dots, x_n\}$ are drawn independently from f_0 . Recall the **BALLET** estimator $\hat{\psi}_{\delta,\lambda,M}(\mathcal{X}_n)$ from (6) based on:*

1. *the loss $L_{IA\text{-}Binder}$ with parameters $0 < a = b \leq 1$, $m = m_{ia} = m_{ai} \leq 1$, and $a \leq 2m$,*
2. *a model M that satisfies Assumption 2, and*

3. a non-random $\delta \in [2 \left(\frac{16d \ln n}{nv_d \lambda} \right)^{1/d}, \bar{\delta}]$ or the data adaptive choice of $\delta = \hat{\delta}$ from (3) with $\gamma < 1$ and $\log n \ll k \ll n$ as $n \rightarrow \infty$,

where $\bar{\delta}$ is a positive constant that depends on f_0 and λ . Then

$$\binom{n}{2}^{-1} L_{IA\text{-Binder}}\{\hat{\psi}_{\delta,\lambda,M}(\mathcal{X}_n), \psi_\lambda(f_0)\} \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty.$$

By Remark 3, the corollary implies that only a vanishingly small fraction of pairs of distinct points from $\mathcal{X}_n$ will be clustered differently by our **BALLET** estimator $\hat{\psi}_{\delta,\lambda,M}(\mathcal{X}_n)$ and the associated true level set clustering $\psi_\lambda(f_0)$ as $n \rightarrow \infty$.

Remark 7. Assumption 2 with $\rho(f, g) = \|f - g\|_\infty$ seems stronger than necessary to establish the consistency of our **BALLET** estimator (6), which depends on the model M only through the distribution of the level set $S_{\lambda,f} = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}$ under the posterior draw $f \sim P_M(\cdot | \mathcal{X}_n)$. One might thus hope to leverage existing posterior contraction results (Gayraud and Rousseau, 2005, 2007; Li and Ghosal, 2021) for level sets that show

$$P_M [\tilde{\rho}(S_{\lambda,f}, S_{\lambda,f_0}) > \epsilon | \mathcal{X}_n] \xrightarrow{P} 0 \text{ as } n \rightarrow \infty, \text{ for each } \epsilon > 0,$$

where $\tilde{\rho}(A, B) = \text{Leb}(A \Delta B)$ is typically the Lebesgue measure of the symmetric difference between (measurable) subsets $A, B \subseteq \mathcal{X}$. Consistency of **BALLET** then essentially reduces to establishing a ‘continuity’ result similar to Theorem 4 that will bound the distance $D\{\tilde{\psi}_{\delta,\lambda}(f), \psi_\lambda(f_0)\}$ between clusterings whenever the distance $\tilde{\rho}(S_{\lambda,f}, S_{\lambda,f_0})$ between the corresponding level sets is small. This approach seems more feasible if $\tilde{\rho}$ can be taken to be a stronger metric like the Hausdorff metric (Li and Ghosal, 2021; Chen et al., 2017).

5. Illustrative Challenge Datasets

To highlight some of the appealing properties of the **BALLET** estimator, we analyze two illustrative clustering datasets: a simulated example of the classic two moon problem and an RNA sequencing dataset (<https://www.reneshbedre.com/blog/tsne.html>).

For each dataset, we model the observations as iid draws from density f and f as a draw from a Dirichlet process mixture of normal distributions with a multivariate normal-inverse Wishart base measure (DPMM). We generate samples $f^{(1)}, \dots, f^{(S)}$ from the posterior $f | \mathcal{X}_n$ using the `dirichletprocess` package, available on CRAN.

We then use these posterior samples to compute **BALLET** clustering point estimates. For the two-moon problem, we choose the target density level λ at the 10th percentile of the estimated observation densities $\{\hat{f}(x_i) : x_i \in \mathcal{X}_n\}$ such that 90% of the observations are assigned to clusters and 10% are labeled as noise. For the RNA-seq data, we set λ at the 15th percentile. These results are visualized in the right column of Figure 3.

In the center column of Figure 3 we visualize the clustering estimate obtained from a traditional mixture component allocation approach to Bayesian clustering and summarized using Dahl et al. (2022). The same DPMM posterior was used for both sets of clustering estimates; the associated density point estimates, $\hat{f}$, are visualized in the left column.

Additional analyses of these and one other simulated data set are collected in Section E. In particular, we show credible bounds (Figure 13), highlight the robustness of **BALLET** to alternative models for f (Figure 11), and present results over a range of values for λ (Figure 12, Figure 14). A discussion of how we chose the level λ can be found in Section J.

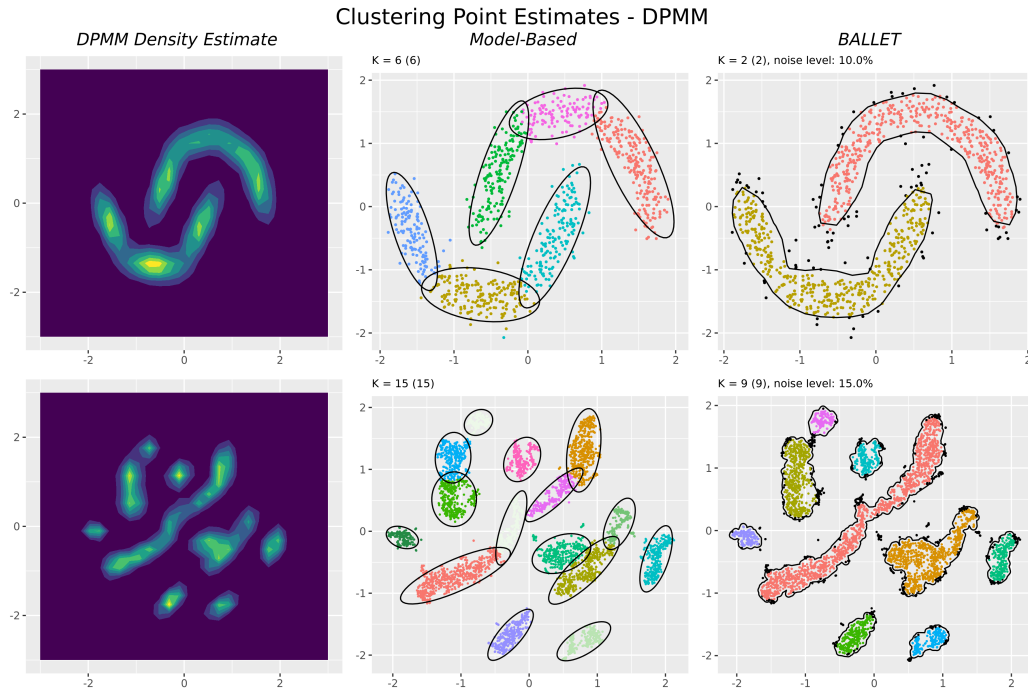


Figure 3: Analysis of the two moons and RNA-seq datasets. The first column shows a heatmap of $E(f|\mathcal{X}_n)$ for the DPMM model. The center column shows the cluster estimate obtained from the traditional mixture-component allocation approach, and the third shows our **BALLET** point estimates. The number of clusters identified in each point estimate is shown at the top of each subplot, with the number of non-singleton clusters listed in parentheses. For the **BALLET** subplots we also note the target level of noise-points used to set λ .

6. Analysis of Astronomical Sky Survey Data

Astronomical sky surveys document the locations and redshifts of galaxies in the cosmos (Nichol et al., 1992). One aim in collecting the data is to analyze the spatial distribution of galaxies, as the size and distribution of high-density regions can help us estimate certain parameters of cosmological models, as described by Jang (2006) in their non-Bayesian analysis of this level set clustering problem. Here, we perform a parallel analysis using **BALLET**, which offers us the benefits of more stable Bayesian nonparametric density estimation and Bayesian uncertainty quantification.

The data $\mathcal{X}_n$ are a cleaned subset of the Edinburgh-Durham Southern Galaxy Catalogue (Nichol et al., 1992) consisting of $n \approx 41,000$ observations in a square region $\mathcal{X} \subseteq \mathbb{R}^2$ and come with two catalogs of *suspected* cluster locations: the Abell catalog (Abell et al., 1989) and the Edinburgh/Durham Cluster Catalog I (EDCCI) (Lumsden et al., 1992). The

former was created by visual inspection of the data by domain experts, while the EDCCI was produced by a custom-built cluster identification algorithm. Figure 22 visualizes the locations from these two catalogs overlying our posterior density estimate (Section 6.1). Here we aim to estimate level set clusters and their uncertainty, and compare the results to locations in the two catalogs, which will serve as our imperfect ground truth.

We first conduct a simulation study, generating one hundred synthetic datasets designed to resemble the Edinburgh-Durham Southern Galaxy Catalogue data, analyzing them by the same **BALLET** methodology we will use for the real data, and computing sensitivity and specificity in detecting regions with excess density. To accommodate the fact that target clusters are described only by their central point (corresponding to a *simulated* catalog location) henceforth called a *target point*, we evaluate sensitivity and specificity based on small ellipses enclosing each estimated cluster: sensitivity is measured as the proportion of target points contained in at least one ellipse, while specificity is measured as the proportion of ellipses that contain a target point. Since sensitivity and specificity will both be equal to one if all the data points are assigned to a single cluster, we also compute a metric called *exact match*, defined as the fraction of ellipses that have exactly one target point. As a competitor, we apply DBSCAN (Ester et al., 1996).

6.1 Density Model and Choice of Parameters

In both the simulation study and real data analysis, we model the density f with a simple mixture of random histograms: $f(x) = \sum_{k=1}^K \pi_k H_k(x; \mathcal{B}_k, \vec{\rho}_k)$, where $H_k(x; \mathcal{B}_k, \vec{\rho}_k) = \sum_{m=1}^M \mathbb{1}_{(x \in B_{km})} \rho_{km}$ is a histogram density with bins $\mathcal{B}_k = (B_{k1}, \dots, B_{kM})$ and weights $\vec{\rho}_k = (\rho_{k1}, \dots, \rho_{kM})$. We provide more details on our prior along with a fast approximation to sample from the posterior of f in Section F.

Cosmological theory (see Jang, 2006) suggests the use of the level $\lambda = (1+c)\bar{f}$, where the constant c is approximately one and $\bar{f} = \frac{\int_{\mathcal{X}} f(x) dx}{\text{Vol}(\mathcal{X})} = 1/\text{Vol}(\mathcal{X})$ denotes the average value of f . We chose the value $c = 1$ for our analysis of the real data. This corresponded to declaring the fraction $\nu = 0.927$ of data points as noise. In the simulation study, we fix the fraction of noise points which are not assigned to a cluster at $\nu = 0.9$ and set $\delta = \hat{\delta}$ from (3). The analogous parameter settings for DBSCAN are $\text{MinPts} = k$ and $\text{Eps} = q_{1-\nu}[\{\delta_k(x_i) : x_i \in \mathcal{X}_n\}]$ (Ester et al., 1996), where $\delta_k(x)$ is the distance from x to the k th nearest point in the dataset $\mathcal{X}_n$ and q_α is the quantile function corresponding to $\alpha \in (0, 1)$. Unlike for **BALLET**, the performance of DBSCAN in our simulation study was sensitive to the choice of k (see Figure 19). We also present results from DBSCAN in Sections G and H with $\text{MinPts} = 60$ which was chosen via grid-search to optimize performance. The results were comparable to those of **BALLET** using the default parameter values.

6.2 Simulation Study

The simulation data were drawn from a mixture distribution that placed $\nu = 90\%$ of its mass in a uniform distribution over the unit square and divided the remaining 10% between 42 bivariate isotropic Gaussian components, with relative weights determined by a draw from a uniform distribution over the probability simplex. The component means are sampled uniformly from the unit square, and the variances were drawn from a diffuse inverse gamma distribution. We randomly generated one hundred such mixture distributions and

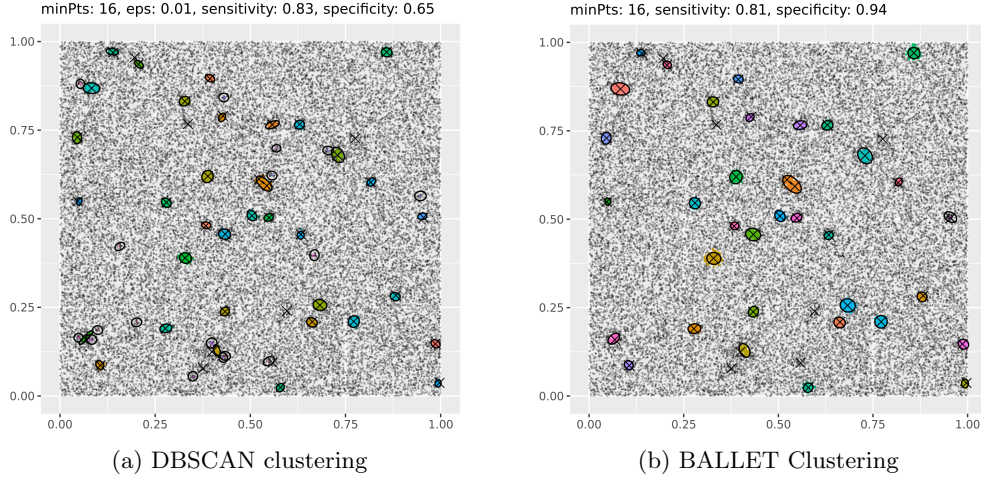


Figure 4: Clusters estimated by DBSCAN and BALLET for a representative synthetic sky survey dataset from our simulation study. We see an apparent preference of DBSCAN for detecting a large number of singleton or near-singleton clusters.

drew $n = 40000$ independent and identically distributed observations from each mixture distribution, dropping any observations that fell outside the unit square. We plot a typical synthetic data set in Figure 17 and display the associated true and estimated high-density regions in Figure 18.

In Figure 4, we show the result of applying DBSCAN and BALLET to the typical synthetic dataset, highlighting DBSCAN’s apparent preference for detecting a large number of singleton or near-singleton clusters given our default choice of $\text{MinPts} = k_0 = \lceil \log_2(n) \rceil = 16$ and the known fraction of noise points $\nu = 90\%$. The average performance of DBSCAN and BALLET clustering (point estimate and upper and lower bounds) in all the hundred datasets is shown in Table 2. DBSCAN achieved an average sensitivity of 0.86, but suffered substantial false positives with an average specificity of 0.49 (exact match = 0.45). BALLET achieved an average sensitivity of 0.78 while maintaining a nearly perfect average specificity at 0.99 (exact match = 0.88). The BALLET lower and upper bounds performed more and less conservatively, respectively, than the point estimate. In particular, on average, the BALLET lower bound had less sensitivity (.62) but more specificity (.99) and exact matches (.9), while the BALLET upper bound had more sensitivity (.89) but less specificity (.96) and exact matches (.83).

The performance of DBSCAN improved to match that of BALLET when $\text{MinPts} = k = 60$ was chosen to maximize the sum of the sensitivity and specificity values (Table 2). The performance of BALLET remained insensitive to the choice of k (Figure 19). Thus while carefully tuning hyper-parameters based on the ground truth was necessary for DBSCAN to match the performance of BALLET, the performance of BALLET seems more robust to loss parameters. This may be because BALLET separates careful density modeling from the task of inferring level set clusters (see Remark 8).

6.3 Sky Survey Data Analysis

We applied DBSCAN and BALLET to the Edinburgh-Durham Southern Galaxy Catalogue data as described above, choosing $\text{MinPts} = k_0$ based on our default value of $k_0 = \lceil \log_2(n) \rceil = 16$ or $\text{MinPts} = 60$, the value optimized in our simulation study. Clustering results are shown in Figures 23 to 25.

	DBSCAN	DBSCAN ¹	BALLET Lower	BALLET Est.	BALLET Upper	BALLET Plugin
Sensitivity	0.79	0.69	0.29	0.67	0.86	0.67
Specificity	0.20	0.65	0.87	0.69	0.42	0.69
Exact Match	0.17	0.46	0.67	0.51	0.32	0.53

Table 1: DBSCAN and BALLET coverage of suspected galaxy clusters in the EDCCI catalog. Column DBSCAN reports performance with our default tuning parameter choice $\text{MinPts} = 16$, while DBSCAN¹ shows performance with $\text{MinPts} = 60$ based on our simulation study.

Table 1 compares inferred clusters to the EDCCI catalog of suspected galaxy clusters. While DBSCAN with heuristic parameter choice detected 79 percent of the EDCCI clusters, the method only had a specificity of 20 percent. DBSCAN with the parameter optimized in our simulation study found 69 percent of the EDCCI clusters with a specificity of 65 percent. BALLET recovered 67 percent of the EDCCI clusters and had a specificity of 69 percent. DBSCAN and BALLET detected only 40 percent of the Abell catalog clusters (Table 3), but performed better at recovering suspected galaxy clusters in the EDCCI, which is considered more reliable (Jang, 2006).

Figure 5 visualizes BALLET clustering uncertainty (Section 3) via upper and lower bounds for a 95 percent credible ball. The lower bound has fewer and smaller clusters and tends to include locations that the EDCCI and Abell catalogs agree on. In contrast, the upper bound has larger and more numerous clusters, and tends to include many of the suspected cluster locations from both the catalogs. Based on Tables 1 and 3, one may suspect that the 14 percent EDCCI locations and 44 percent Abell locations that were not discovered by the BALLET upper bound may be erroneous. On the other hand, we may have high confidence in the 29 percent locations in EDCCI and 21 percent locations in Abell which were discovered by the BALLET lower bound.

7. Discussion

In this article, we developed a Bayesian approach to density-based clustering, focusing on level set clustering as an important special case. Our key idea is to use Bayesian decision theory (Berger, 2013) to separate the tasks of modeling the data density and inferring clusters. This provides a general new paradigm for inferring clusters, while representing uncertainty in clustering. A decision-theoretic decoupling approach has proved useful in various problem settings like interpretable modeling (Gutiérrez-Peña and Walker, 2005; Afrabandpey et al., 2020; Woody et al., 2021), variable selection in regression (Kowal, 2022a; Hahn and Carvalho, 2015), factor analysis (Bolfarine et al., 2024), structured covariance estimation (Bashir et al., 2019), and analysis of functional data (Kowal and Bourgeois, 2020;

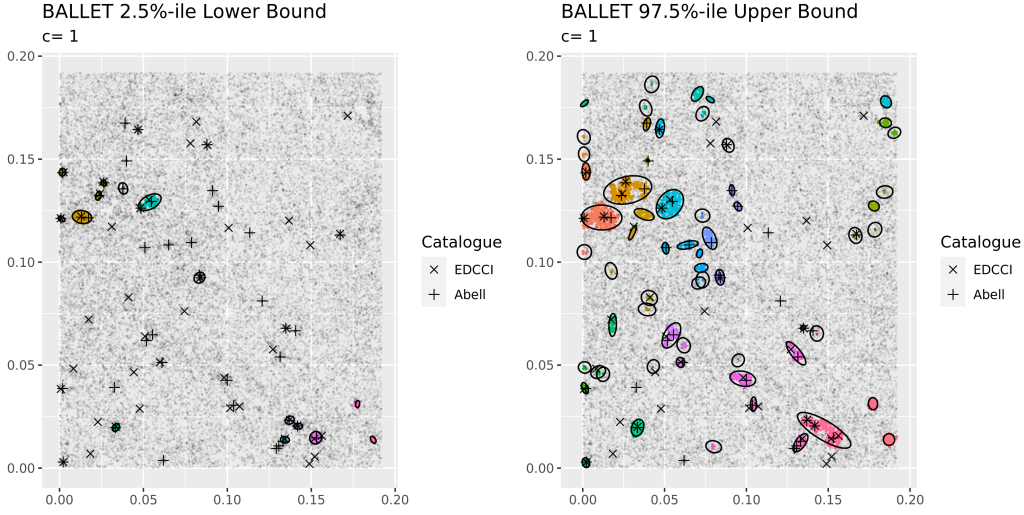


Figure 5: Upper and lower bounds of the 95% credible ball centered at our BALLET estimate of the galaxy clusters in the Edinburgh-Durham Southern Galaxy Catalogue data.

Kowal, 2022b). Our approach is also a case of this posterior decoupling methodology where we establish necessary conditions for consistency (Theorem 1).

A crucial and implicit part of our methodology is the model M on the space of densities. In any application, the problem of coming up with a good model M is of course an issue that pervades Bayesian statistics. As we note in Section 4, if the posterior $P_M(\cdot|\mathcal{X}_n)$ is consistent, the choice of the density model M will not substantially affect the discovery of the true clustering $\psi_\lambda(f_0)$ for large sample sizes. Figures 10 to 12 in Section E demonstrate this effect. For smaller sample sizes, a thoughtful choice for M (e.g. a parametric mixture model with few components) can be used with our methodology to ensure that there is enough signal to detect true clusters. For high dimensional problems, leveraging Chandra et al. (2023), one can use BALLET to find the level set clusters for a low-dimensional latent representation of the data.

While level set clustering is a popular and conceptually appealing framework, a key practical challenge is the choice of the level $\lambda > 0$ (Campello et al., 2019). Indeed, based on visualizing the density estimate for our sky survey data (Figure 6), we expect our clusters to be sensitive to the exact value of the scientific constant c . To reduce sensitivity to λ , we describe a *persistent* clustering approach in Section K that computes BALLET clusters for values of $c \in [.8, 1.2]$, visualizing these clusters with a cluster tree (Zappia and Oshlack, 2018). This tree is then processed to infer clusters that remained active or *persistent* across all the levels in the tree. This approach improved our specificity in detecting the two catalogs without losing sensitivity.

While we have focused on level set clustering, our Bayesian density-based clustering framework is broad and motivates multiple directions for future work. One possibility is to avoid focusing on a single threshold λ , but instead estimate a cluster tree obtained by varying the threshold. Loss functions introduced by Fowlkes and Mallows (1983) provide a relevant starting point. An alternative is to target a single clustering, but vary the threshold

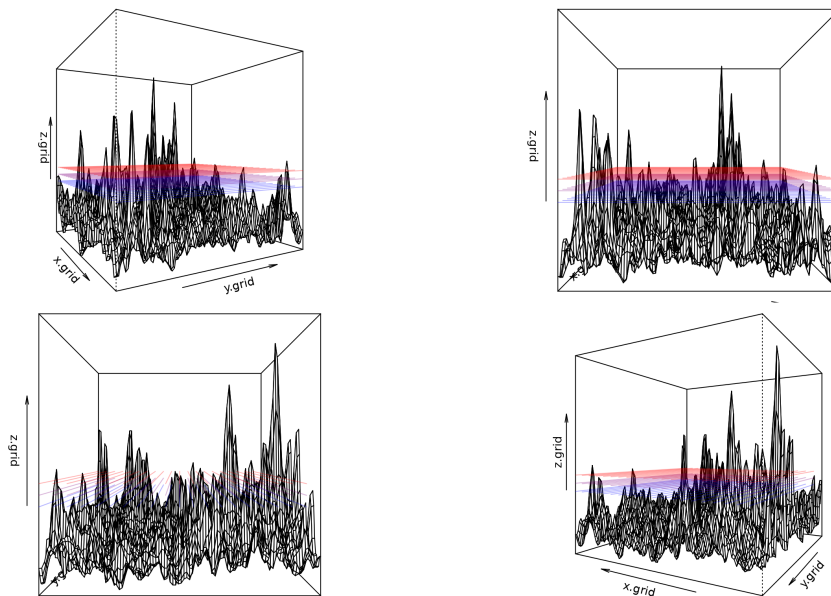


Figure 6: Visualizing our density estimate (plotted on the z -axis) for the Edinburgh-Durham Southern Galaxy Catalogue data. The colored lines mark the choice of different levels corresponding to the values of $c \in \{.8, 1, 1.2\}$.

λ over the observation space in a data-adaptive manner (Campello et al., 2015). Varying λ is important in uncovering distinct cluster structures at varying levels of the density; refer, for example, to the illustrative example in Figure 31.

Finally note that for a general non-parametric density f , it is hard to find a single notion of clustering that will be universally appropriate across all applications. However, a natural notion at least when f is sufficiently regular, may be that of modal clustering (Chacón, 2015; Menardi, 2015) that associates clusters with the domain of attraction of the modes of f . Interestingly, as recently argued in Arias-Castro and Qiao (2023), both level set clustering and modal clustering may fundamentally be the same approach.

Acknowledgments

This work was partially funded by grants R01-ES028804 and R01-ES035625 from the United States National Institutes of Health and N00014-21-1-2510 from the Office of Naval Research. The authors would like to thank Dr. Woncheol Jang for kindly providing the data for our case study, and Dr. Cliburn Chan for suggesting applications in cosmology.

Appendix A. Literature on Bayesian Clustering

The last two decades have witnessed a significant maturation of the Bayesian clustering literature (Medvedovic and Sivaganesan, 2002; Fritsch and Ickstadt, 2009; Wade and Ghahramani, 2018; Rastelli and Friel, 2018; Dahl et al., 2022). By designing and characterizing loss functions on partitions and developing search algorithms to identify partitions which minimize Bayes risk, these articles and others have established a sound framework for Bayesian decision-theoretic clustering. This literature acknowledges the *cluster-splitting* problem alluded to in our preceding discussion, with Wade and Ghahramani (2018) and Dahl et al. (2022) finding that clustering point estimates obtained by minimizing Bayes risk under certain parsimony-encouraging loss functions are less prone to cluster-splitting.

However, these loss functions cannot completely eliminate the problem. Guha et al. (2021) shows that a fundamental cause of cluster splitting is that Bayesian mixture models converge to the mixture that has minimum Kullback-Leibler divergence to the true density. When the components of the mixture are not specified correctly, it may require infinitely many parametric components to recapitulate the true data-generating density. Thus, as data accumulate, it would seem futile to attempt to overcome the cluster-splitting problem merely by encouraging parsimony in the loss function. If the components are at all misspecified as data accumulate, eventually the preponderance of evidence will insist on splitting the clusters to reflect the multiplicity of parametric components. Indeed, in our illustrative example in Figure 1 (a) we used the parsimony-encouraging Variation of Information (VI) loss to obtain the Gaussian mixture model-based clustering point estimate.

One response to this problem is the coarsened Bayes methodology of Miller and Dunson (2019), which only assumes the mixture model to be *approximately* correctly specified. Another approach to mitigate the problem is to expand the class of mixture components (Frühwirth-Schnatter and Pyne, 2010; Malsiner-Walli et al., 2017; Stephenson et al., 2020). As we have claimed above, naive applications of this strategy can lead to loss of practical identifiability and computational challenges, although Dombowsky and Dunson (2025) have had some success increasing component flexibility *indirectly* by merging nearby less flexible mixture components in a post-processing step. The generalized Bayes paradigm, introduced by Bissiri et al. (2016), also provides an answer to the cluster splitting problem via a loss-function-based Gibbs posterior for clustering (Rigon et al., 2023).

The idea of defining Bayesian clustering as a problem of computing a risk-minimizing summary ψ of the posterior distribution on density f can be viewed as related to the existing literature on decision-theoretic summaries of posterior distributions (Woody et al., 2021; Afrabandpey et al., 2020; Ribeiro et al., 2018), though this literature has focused largely on extracting interpretable conclusions from posterior distributions on regression surfaces. In contrast, clustering in the manner we have proposed extracts an interpretable summary from a posterior distribution on the data-generating density. In addition, while the authors of that literature focus on the interpretability of summary functions ψ , we use the clustering example to emphasize that ideally ψ should also be robust, in the sense that $\psi(f^*)$ will be close to $\psi(f)$ when f^* is close to f , since this would suggest that small amounts of prior bias or model misspecification would not lead to large estimation errors.

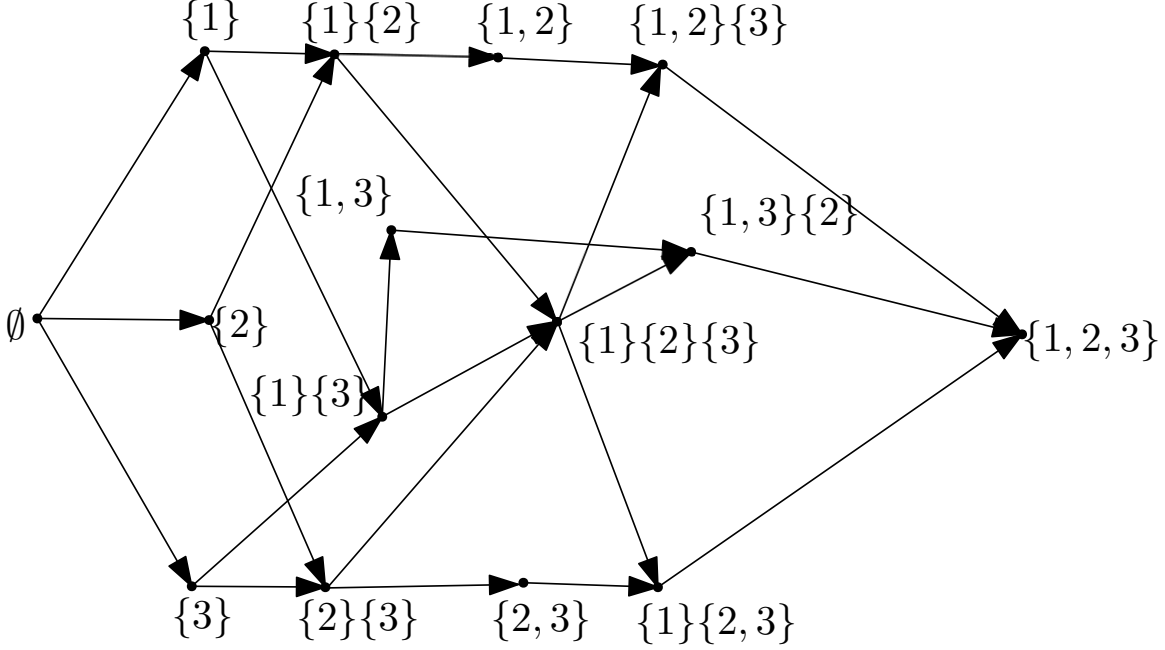


Figure 7: Hasse diagram for the lattice of sub-partitions $\mathcal{P}(\mathcal{X})$ of the space $\mathcal{X} = \{1, 2, 3\}$. This diagram has the property that $\mathcal{C} \prec \mathcal{C}'$ if and only if there is a path from $\mathcal{C}$ to $\mathcal{C}'$.

Appendix B. The lattice of sub-partitions

The space of sub-partitions $\mathcal{P}(\mathcal{X})$ forms a lattice under the partial order given by $\mathcal{C} \preceq \mathcal{C}'$ defined by the existence of a map $\phi : \mathcal{C} \rightarrow \mathcal{C}'$ such that $C \subseteq \phi(C)$ for each $C \in \mathcal{C}$. One can check that $(\mathcal{P}(\mathcal{X}), \preceq)$ with join $\mathcal{C} \vee \mathcal{C}' \doteq \{C \cup C' \mid C \in \mathcal{C}, C' \in \mathcal{C}', C \cap C' = \emptyset\}$ and meet $\mathcal{C} \wedge \mathcal{C}' = \{C \cap C' \mid C \in \mathcal{C}, C' \in \mathcal{C}', C \cap C' = \emptyset\}$ is a lattice.

We denote $\mathcal{C} \prec \mathcal{C}'$ if $\mathcal{C} \preceq \mathcal{C}'$ but it is not the case that $\mathcal{C}' \preceq \mathcal{C}$. We can define a Hasse diagram for this lattice based on the relation $\mathcal{C} \rightarrow \mathcal{C}'$ if $\mathcal{C} \prec \mathcal{C}'$ but there is no $\mathcal{C}'' \in \mathcal{P}(\mathcal{X})$ such that $\mathcal{C} \prec \mathcal{C}'' \prec \mathcal{C}'$. One can show that $\mathcal{C} \rightarrow \mathcal{C}'$ if and only if one of the following conditions hold:

- $\mathcal{C}'$ is obtained by merging two active clusters in $\mathcal{C}$. That is, after suitable reordering: $\mathcal{C} = \{C_1, \dots, C_k\}$ and $\mathcal{C}' = \{C_1 \cup C_2\} \cup \{C_r : r \in \{3, \dots, k\}\}$.
- $\mathcal{C}'$ is obtained by adding a noise point to its own cluster: i.e., $\mathcal{C}' = \mathcal{C} \cup \{n\}$ for some $n \in \mathcal{X}$ that is not active in $\mathcal{C}$.

This relation allows us to construct a Hasse diagram: a directed acyclic graph with nodes $\mathcal{P}(\mathcal{X})$ and edges given by the relation $\rightarrow$. This diagram has the property that $\mathcal{C} \prec \mathcal{C}'$ if and only if there is a path from $\mathcal{C}$ to $\mathcal{C}'$. The Hasse diagram for the lattice of sub-partitions of $\mathcal{X} = \{1, 2, 3\}$ is shown in Figure 7.

Appendix C. DBSCAN and other level set clustering methods

Starting from works like Hartigan (1975), the topic of level set clustering has been extensively studied from the perspective of algorithms (Bhattacharjee and Mitra, 2020; Campello et al., 2019), statistical methodology (Cuevas et al., 2000, 2001; Stuetzle and Nugent, 2010; Scrucça, 2016), and statistical theory (Menardi, 2015; Wang et al., 2019; Steinwart et al., 2023). Interestingly, while the popular DBSCAN algorithm (Ester et al., 1996; Schubert et al., 2017) has been around for a while, tools for its theoretical study are more recent (Sriperumbudur and Steinwart, 2012; Jiang, 2017; Wang et al., 2019). Here we describe the DBSCAN algorithm and relate it to our surrogate clustering function $\tilde{\psi}_{\delta,\lambda}(f)$, which we described in Section 2.4 motivated by statistical theory.

The DBSCAN algorithm finds arbitrarily shaped clusters of related data points in large spatial databases (Ester et al., 1996). The DBSCAN *cluster model* (Schubert et al., 2017, Section 2.1) is not explicitly described in terms of the data density f_0 , but rather in terms of a notion of distance $\text{dist}(x_i, x_j)$ measuring relatedness between observations $x_i, x_j \in \mathcal{X}_n$ and two free parameters $\text{Eps} > 0$ and $\text{MinPts} \in \mathbb{N}$. A data point $x \in \mathcal{X}_n$ is called a *core* point if it has at least MinPts many neighbors $N_{\text{Eps}}(x) \doteq \{y \in \mathcal{X}_n : \text{dist}(x, y) \leq \text{Eps}\}$ that are within a distance Eps of it (i.e. $|N_{\text{Eps}}(x)| \geq \text{MinPts}$). The set of all *core* points $\mathcal{A} = \{x \in \mathcal{X}_n : |N_{\text{Eps}}(x)| \geq \text{MinPts}\}$ are then clustered based on the partition induced by the transitive closure of the relation $\{(x, y) \in \mathcal{A} \times \mathcal{A} : \text{dist}(x, y) \leq \text{Eps}\}$. In words, the DBSCAN clustering of $\mathcal{A}$ is the finest partition of $\mathcal{A}$ where each pair of points $x, y \in \mathcal{A}$ satisfying $\text{dist}(x, y) \leq \text{Eps}$ are clustered together. While the DBSCAN algorithm goes on further to add some of the *non-core* points (called *border* points) that lie within a neighborhood $N_{\text{Eps}}(x)$ of some core point $x \in \mathcal{A}$ to a corresponding cluster, for consistency with level set clustering, this step is avoided by a variant of the algorithm called DBSCAN* (Campello et al., 2015).

When $\mathcal{X} = \mathbb{R}^d$ and $\text{dist}(x, y) = \|x - y\|$ is Euclidean distance, the notion of *core points* from DBSCAN is seen to be related to the notion of *core* or *active* points that we introduced in Section 2.1. In fact, and as indicated in Sriperumbudur and Steinwart (2012); Jiang (2017); Campello et al. (2015), the clustering from DBSCAN* is the same as our surrogate clustering $\tilde{\psi}_{\delta,\lambda}(\hat{f}_\delta) \in \mathcal{P}(\mathcal{X}_n)$ where $\hat{f}_\delta(x) = n^{-1} \sum_{x_i \in \mathcal{X}_n} \kappa_\delta(x_i - x)$ is the kernel density estimate based on the uniform kernel $\kappa_\delta(z) = \mathbb{I}\{\|z\| \leq \delta\} / (v_d \delta^d)$ and $v_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ is the volume of the d -dimensional unit Euclidean ball. Here $\delta = \text{Eps}$ and $\lambda = \text{MinPts} / (nv_d \delta^d)$ can be expressed in terms of the original DBSCAN parameters $\text{Eps} > 0$ and $\text{MinPts} \in \mathbb{N}$. In fact, as noted in Campello et al. (2019), there is also another representation of DBSCAN* as $\tilde{\psi}_{\delta,\lambda}(\hat{f}_k)$ where $\hat{f}_k(x) = \frac{k}{nv_d} \delta_k(x)^{-d}$ is the k -nearest neighbor density estimator (Biau and Devroye, 2015) with $\lambda = \frac{k}{nv_d} \delta^{-d}$, $\delta = \text{Eps}$, and $k = \text{MinPts}$.

Remark 8. From the first formulation $\tilde{\psi}_{\delta,\lambda}(\hat{f}_\delta)$ the parameter $\text{Eps} = \delta$ for DBSCAN simultaneously controls both the regularity of the kernel density estimator $\hat{f}_\delta$ used to discover core points $\mathcal{A} = \mathcal{A}_{\lambda, \hat{f}_\delta}$ and also the connectivity of resulting clusters based on the connectivity of the graph $G_\delta(\mathcal{A})$. This is in contrast to *BALLET* where the parameter δ only controls the connectivity of the clusters, and may explain why *BALLET* clustering was seen to be less sensitive to the choice of this parameter in Figure 19.

C.1 Time complexity of evaluating surrogate function

The time complexity of evaluating $\tilde{\psi}_{\delta,\lambda}(f)$ at any fixed f is comparable to that of the DBSCAN algorithm and an additional time complexity κ_n of evaluating f at all of the points in $\mathcal{X}_n$. To see this, suppose first that the δ neighborhood graph for all the data points $G_\delta(\mathcal{X}_n)$ can be pre-computed and stored for future use in an adjacency list representation (Dasgupta et al., 2008, Chapter 3). In order to evaluate $\tilde{\psi}_{\delta,\lambda}(f)$, one can then (i) calculate the set of active nodes $A_{\lambda,f} \subseteq \mathcal{X}_n$ by evaluating f at all the data points, (ii) extract the subgraph $G_\delta(A_{\lambda,f})$ of $G_\delta(\mathcal{X}_n)$ by scanning the precomputed adjacency list, and (iii) compute the connected components of $G_\delta(A_{\lambda,f})$ by using the standard breadth (or depth) first search algorithm (Dasgupta et al., 2008, Chapter 3). Thus, given our precomputed adjacency list representation of $G_\delta(\mathcal{X}_n)$, the time complexity to evaluate $\tilde{\psi}_{\delta,\lambda}(f)$ is $O(\kappa_n + |G_\delta(\mathcal{X}_n)|)$ where $|G_\delta(\mathcal{X}_n)|$ is the sum of the number of edges and vertices in $G_\delta(\mathcal{X}_n)$. The time complexity of pre-computing the graph $G_\delta(\mathcal{X}_n)$ is at most that of running the DBSCAN algorithm up to constant multiples. Indeed, $G_\delta(\mathcal{X}_n)$ can be constructed by performing a range query for each point $x_i \in \mathcal{X}_n$ to discover the set of points $B(x_i, \delta) \cap \mathcal{X}_n$; however, this sequence of range queries is also an essential part of the DBSCAN algorithm (see Schubert et al., 2017) which would thus also require as many steps.

Appendix D. The BALLET optimization algorithm

For any sub-partition $\mathcal{C} = \{C_1, \dots, C_k\}$ of $\{x_1, \dots, x_t\}$, we use an equivalent allocation vector representation $\vec{c} = (c_1, \dots, c_t) \in \{0, 1, \dots, k\}^t$ given by $c_i = h$ if the point x_i belongs to the cluster h , i.e. $x_i \in C_h$, and $c_i = 0$ if the point x_i is classified as noise under this sub-partition, i.e. $x_i \in \{x_1, \dots, x_t\} \setminus \bigcup_{h=1}^k C_h$.

Given Monte Carlo samples $\{f^{(s)}\}_{s=1}^S$ from the posterior distribution $P_M(\cdot|\mathcal{X}_n)$, we first compute the clusterings $\mathcal{C}^{(s)} = \tilde{\psi}_{\delta,\lambda}(f^{(s)}) \in \mathcal{P}(\mathcal{X}_n)$ and their allocation vectors $\vec{c}^{(s)} = (c_1^{(s)}, \dots, c_n^{(s)})$ for each $s \in \{1 \dots S\}$. Next, these allocation vectors are used to precompute the probability estimates in (5), namely

$$\begin{aligned} \hat{\pi}_{i,j}^{(1)} &\doteq S^{-1} \sum_{s=1}^S \mathbb{1}_{(c_i^{(s)} \neq 0, c_j^{(s)} \neq 0, c_i^{(s)} = c_j^{(s)})}, & \hat{\pi}_{i,j}^{(2)} &\doteq S^{-1} \sum_{s=1}^S \mathbb{1}_{(c_i^{(s)} \neq 0, c_j^{(s)} \neq 0, c_i^{(s)} \neq c_j^{(s)})}, \\ \hat{\alpha}_i &\doteq S^{-1} \sum_{s=1}^S \mathbb{1}_{(c_i^{(s)} \neq 0)} \end{aligned}$$

for each $i \neq j \in \{1, \dots, n\}$. With this, the optimization problem in (6) reduces to minimizing the risk

$$\begin{aligned} R(\vec{c}') &= (n-1) \left\{ m_{ai} \sum_{i=1}^n \mathbb{1}_{(c'_i=0)} \hat{\alpha}_i + m_{ia} \sum_{i=1}^n \mathbb{1}_{(c'_i \neq 0)} (1 - \hat{\alpha}_i) \right\} \\ &\quad + \sum_{1 \leq i < j \leq n} \mathbb{1}_{(c'_i \neq 0, c'_j \neq 0)} \{ a \hat{\pi}_{i,j}^{(1)} \mathbb{1}_{(c'_i \neq c'_j)} + b \hat{\pi}_{i,j}^{(2)} \mathbb{1}_{(c'_i = c'_j)} \} \end{aligned} \tag{9}$$

over all allocation vectors $\vec{c}' = (c'_1, \dots, c'_n)$ corresponding to sub-partitions $\mathcal{C}' \in \mathcal{P}(\mathcal{X}_n)$.

Although exact minimization over the combinatorial space $\mathcal{P}(\mathcal{X}_n)$ of sub-partitions is computationally intractable, we can adapt heuristic algorithms for approximate minimization over the related space of partitions of $\mathcal{X}_n$ (e.g., Fritsch and Ickstadt, 2009; Rastelli and Friel, 2018). Particularly, we consider the algorithm of Dahl et al. (2022) that, given a candidate partition of $\mathcal{X}_n$, provides two important ways to compute a candidate set of partitions that may have a smaller objective value: (i) a series of incremental update steps called the *sweetening phase* that reassigns each data point x_i (chosen in a random order) to a different cluster if doing so will decrease the objective, and (ii) a series of major update steps called the *zealous update phase* that repeatedly destroys a randomly chosen cluster and then incrementally reallocates the data points if doing so decreases the objective. Starting from an initial partition that is either selected at random or is built incrementally to have a small objective value, the algorithm of Dahl et al. (2022) improves the initial partition using the *sweetening phase* followed by the *zealous update phase*. This entire process is repeated (in parallel) many times, and the partition with the least objective value among all the explored partitions is reported.

The main primitive operation needed to implement the above algorithm is to incrementally find a low-risk partition including a new data point (say x_{t+1} for $t \in \{1, \dots, n-1\}$) that respects a given low-risk partition $\{C_1, \dots, C_k\}$ of some existing set of data points, say $\{x_1, \dots, x_t\}$. Indeed, the following two kinds of such partitions of $\{x_1, \dots, x_{t+1}\}$ are possible: (a) the new point x_{t+1} is added to its own cluster; this is the partition $\{C_1, \dots, C_k, \{x_{t+1}\}\}$, or (b) the new point is added to one of the existing clusters (say C_h); this is the partition $\{C_1, \dots, C'_h, \dots, C_k\}$, where $C'_h = C_h \cup \{x_{t+1}\}$. For each of these $k+1$ partitions, Dahl et al. (2022) recommend evaluating the objective value restricted only to the data points under consideration (that is, sum only over terms $i, j \in \{1, \dots, t+1\}$ in our empirical risk (9)) and selecting the partition with the smallest risk among the $k+1$ candidates.

The aforementioned primitive operation is easily extended to the case of sub-partitions of $\mathcal{X}_n$. Indeed, suppose $\mathcal{C} = \{C_1, \dots, C_k\}$ is a sub-partition of $\{x_1, \dots, x_t\}$. The sub-partition $\mathcal{C}'$ of $\{x_1, \dots, x_{t+1}\}$ respects $\mathcal{C}$ in the following three possible ways: a) the point x_{t+1} is assigned to the noise cluster; this is just the sub-partition $\mathcal{C}' = \{C_1, \dots, C_k\}$ in our notation, b) the point x_{t+1} is assigned to its own cluster; this is the sub-partition $\mathcal{C}' = \{C_1, \dots, C_k, \{x_{t+1}\}\}$, and (c) the point x_{t+1} is assigned to an existing cluster (say C_h); this is the sub-partition $\mathcal{C}' = \{C_1, \dots, C'_h, \dots, C_k\}$ where $C'_h = C_h \cup \{x_{t+1}\}$. We then evaluate our risk (9) restricted to the indices $i, j \in \{1, \dots, t+1\}$ using the allocation vector $\vec{c}' = (c_1, \dots, c_{t+1})$ corresponding to $\mathcal{C}'$, and select the sub-partition with the smallest risk among the $k+2$ candidates. This primitive operation allows us to implement the initialization, sweetening, and zealous update phases in the Dahl et al. (2022) algorithm to minimize our risk (9) over allocation vectors that correspond to all sub-partitions of $\mathcal{X}_n$. Notably, in the *zealous update phase* the cluster to be destroyed can either be the current noise cluster or one of the current non-noise clusters.

D.1 Avoiding optimization: BALLET decision-theoretic vs plugin estimator?

Recall the heuristic BALLET *plugin* estimate $\hat{\mathcal{C}} = \psi_{\lambda, \delta}(\hat{f})$ that avoids the expensive optimization in (6) by directly computing the level set clusters of the posterior mean density $\hat{f}(x) \approx \frac{1}{S} \sum_{s=1}^S f^{(s)}(x)$. In most cases, the plugin clustering estimate will be similar to

the decision theoretic **BALLET** estimator from (6) when the posterior uncertainty of f , and particularly that of the level set $\{f \geq \lambda\}$, is low. We note this in our results from Section 6 (see Tables 1 to 3).

However, we now illustrate that the two estimators will at times produce different answers because the heuristic plugin estimate does not take into consideration the posterior uncertainty of f , which may be substantial. Indeed, by modifying our simple example from Figure 1, we see differences emerge when the level λ is increased to the point that there is non-trivial posterior uncertainty in the induced level set $\{f \geq \lambda\}$ (Figure 8).

As a general principle, we recommend the use of Bayes estimators that directly target the quantity of interest, rather than a two-stage plugin approach, where a Bayes estimator is computed for an intermediate quantity. Indeed, there are many examples in the literature in which two-stage plugin approaches are suboptimal.

Appendix E. Additional results from analysis of the illustrative challenge datasets

In this section, we present additional results from the analysis of the illustrative challenge datasets. In Figure 9 we visualize the three datasets, and in Figure 10 we show heat maps of the log of the posterior expectation of the data-generating density f under three different models: a Dirichlet process mixture of Gaussian distributions (DPMM), an adaptive Pólya tree model, and a nearest-neighbor Dirichlet mixture model.

In analyzing these datasets, our choice of loss parameters λ for **BALLET** was guided by the discussion in Section J. In particular, we tuned λ to achieve a certain noise level $\nu \in (0, 1)$, and given ν (and thus λ) the parameter δ was automatically chosen using the data adaptive procedure in Section 2.4 with our default choice of $k = \lceil \log n \rceil$. Here, n is the sample size of the dataset under consideration.

We describe the clustering results using **BALLET** for various choices of noise level ν . In Figures 11 and 12 we compare **BALLET** clustering estimates obtained under our three density models for two different noise levels $\nu \in \{5\%, 10\%\}$. The **BALLET** upper and lower bounds for the RNA-seq data corresponding to noise levels $\nu \in \{5\%, 10\%\}$ are shown in Figure 14. The persistent clusters (see Section K) across the noise levels $\nu \in \{5\%, 10\%, 15\%\}$ for the RNA-seq data are shown in Figure 15. We note that the persistent clusters are somewhat qualitatively different across the density models, demonstrating that the choice of prior can have an effect on the nature of clusters that are discovered.

Finally, we also explore an automatic choice of ν for the various datasets and density models using the elbow heuristic mentioned in Section J. The elbow plots describing the selection of ν are shown in Figure 27, while the corresponding clusters are shown in Figure 26.

Appendix F. The mixture of histograms model for densities

This section describes the *mixture of histograms* model that we use to estimate the data-generating density in Section 6. This model can be fit quickly to a large number of data points since the fitting is primarily based on counting the number of observed data points that fall into various bins. Further, in contrast to a standard histogram model, the density

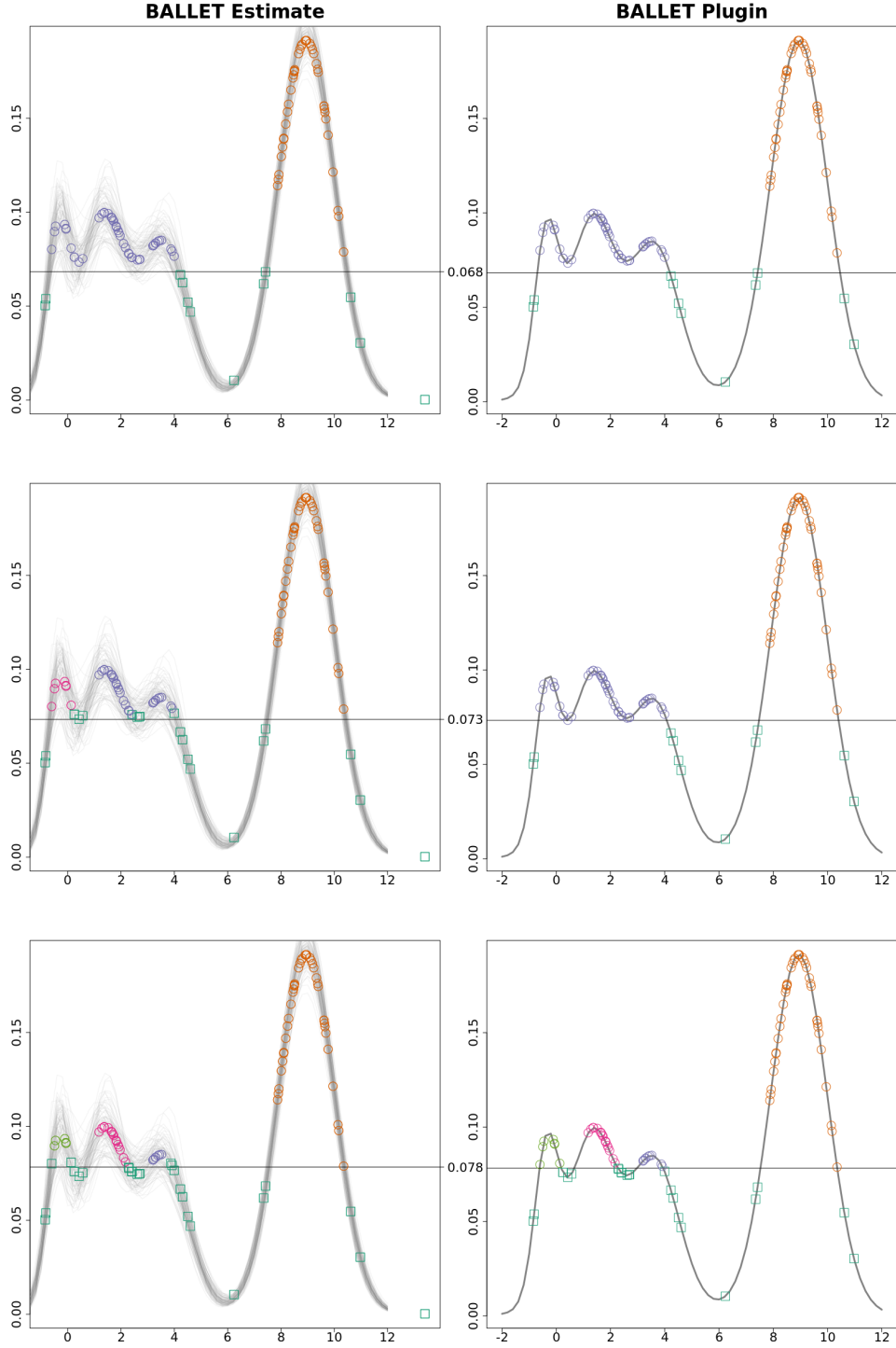


Figure 8: BALLET estimator (6) accounts for the posterior uncertainty of f (left) while the plugin estimator (right) does not. The estimates start to differ when level λ in Figure 1 is increased so that there is non-trivial posterior uncertainty in the level set $\{f \geq \lambda\}$.

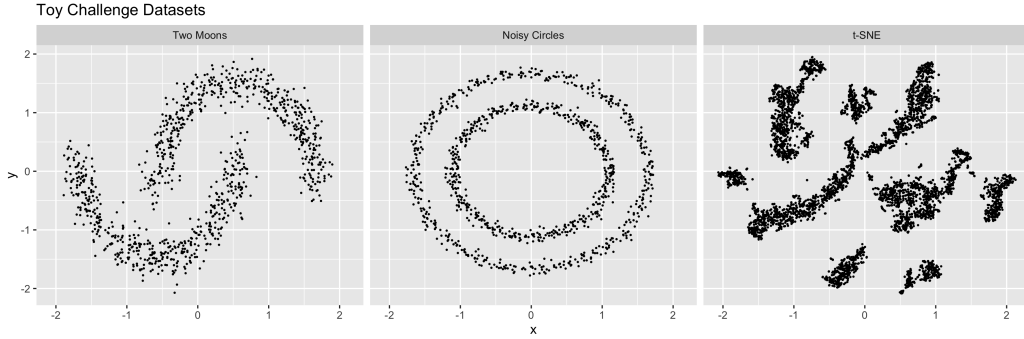


Figure 9: Plots of the three illustrative challenge datasets. From left to right: two moons simulated data, noisy circles simulated data, and a t-SNE embedding of a RNA-seq dataset.

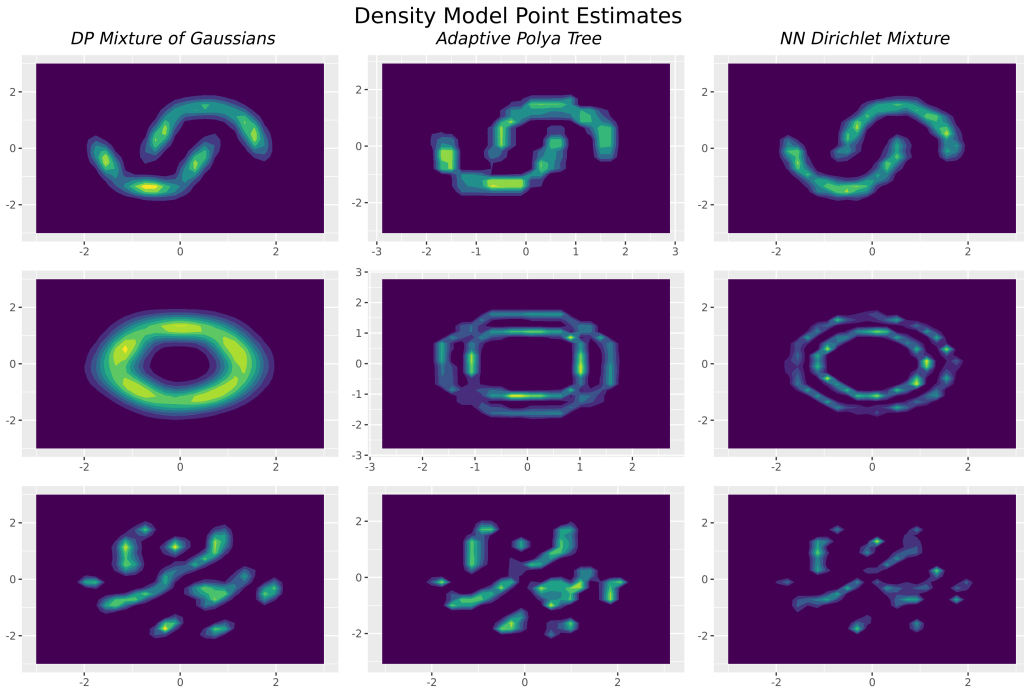


Figure 10: Plots of posterior point estimates of the data-generating densities for each of three illustrative challenge datasets under three different models for the unknown density.

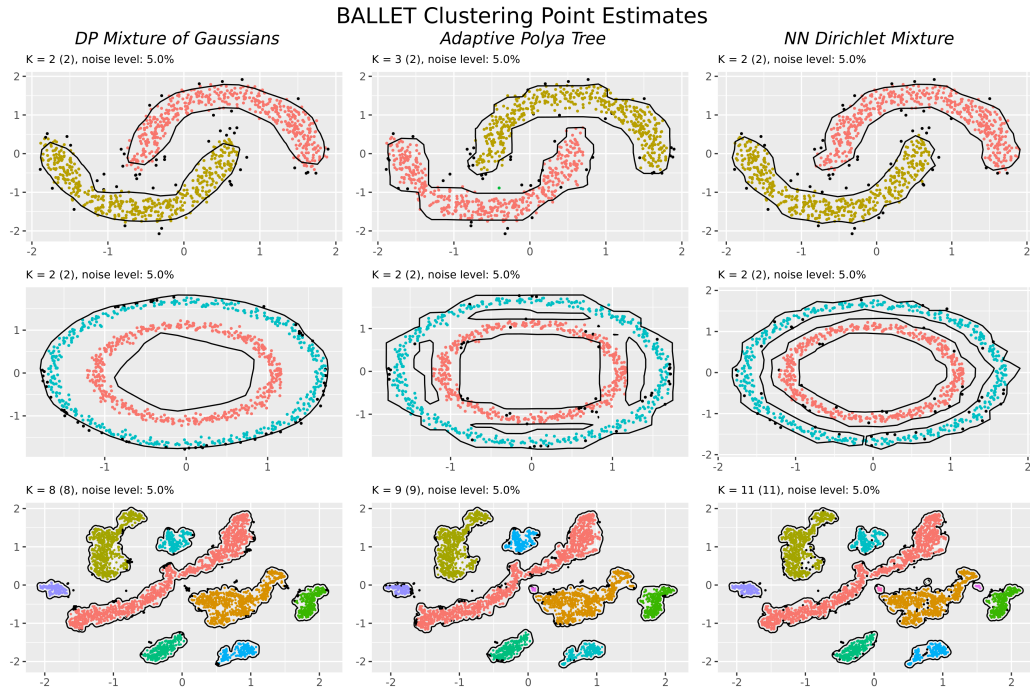


Figure 11: Comparison of BALLET clustering point estimates obtained under the three different density models shown in Figure 10 with $\nu = 5\%$ noise points. The cardinality of the sub partition is displayed in the title of each plot, as $K = X$, and it is followed, in parentheses by the count of clusters with more than 1 observation.

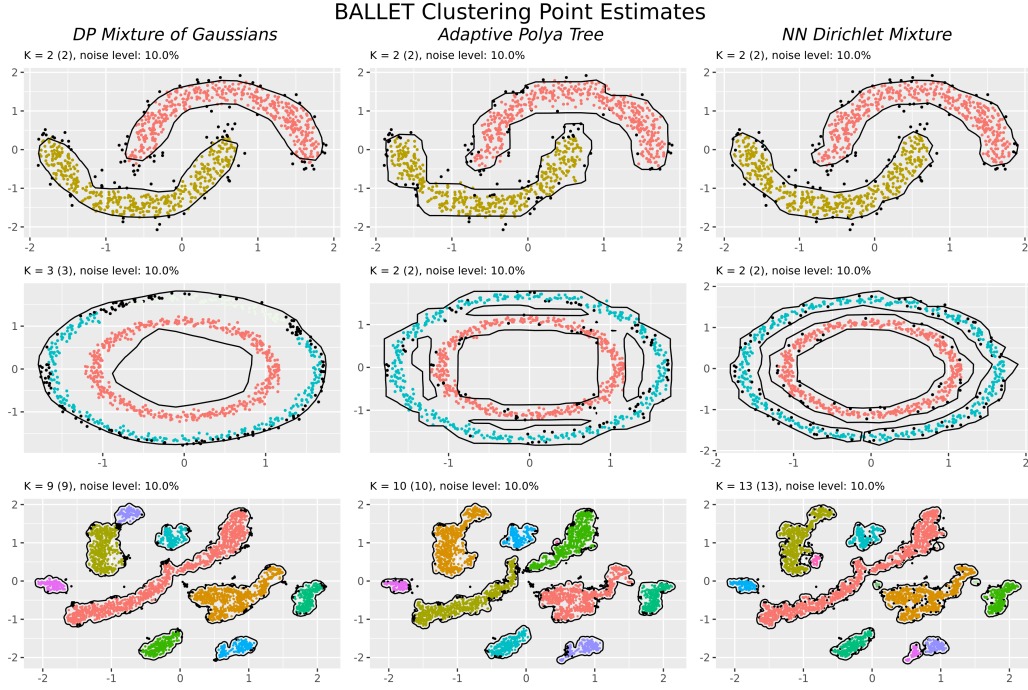


Figure 12: Comparison of BALLET clustering point estimates obtained under the three different density models shown in Figure 10 with $\nu = 10\%$ noise points. Compared to Figure 11, some clusters in second and third rows are seen to split into further clusters based on our choice of the density model. While this may be desirable in the RNA-seq dataset in the last row, increasing the density level does not seem desirable for the Noisy Circles dataset in the second row.

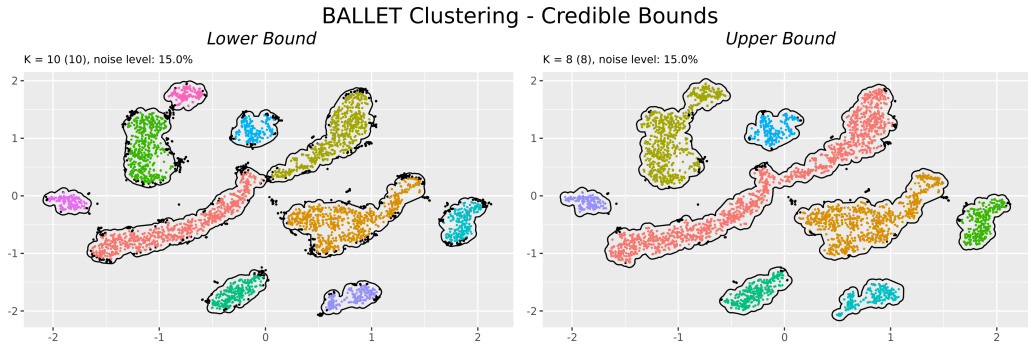


Figure 13: Upper and lower bounds for the 95% credible ball centered at our BALLET clustering estimate for the RNA-seq data, fit with the DPMM model for f . The cardinality of the partition is displayed in the title of each plot, as $K = X$, and it is followed, in parentheses by the count of clusters with more than 1 observation, and the percentage ($\nu = 15\%$) of noise points based on our chosen level λ . Figure 14 in Section E shows additional results for different choices of λ .

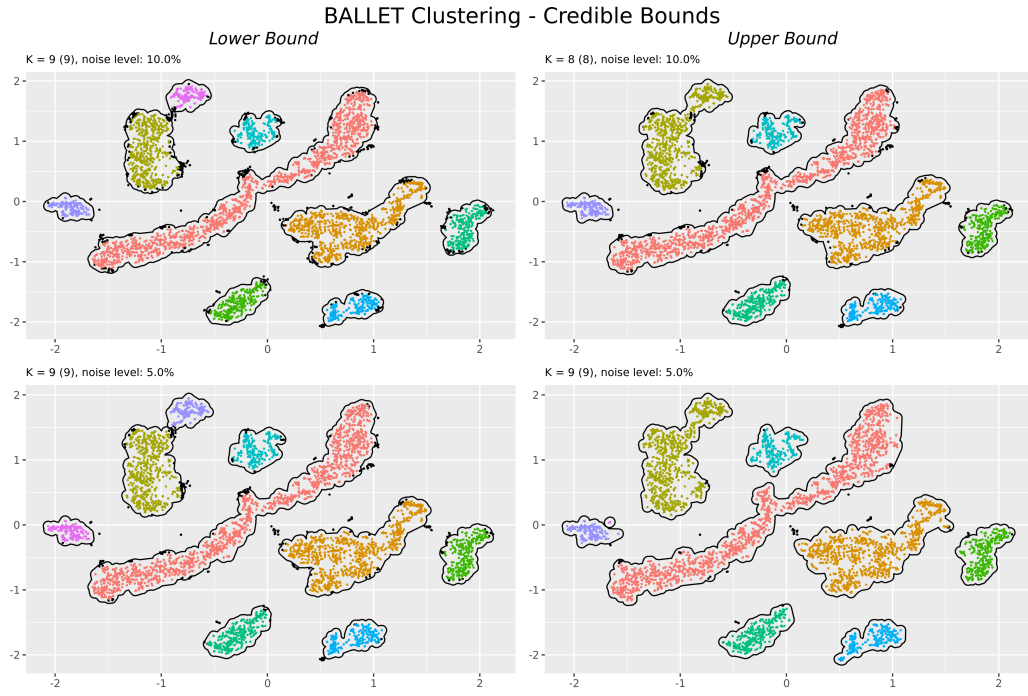


Figure 14: The BALLET upper and lower bounds in Figure 13 for different choices of the level λ , as specified in the subplot titles.

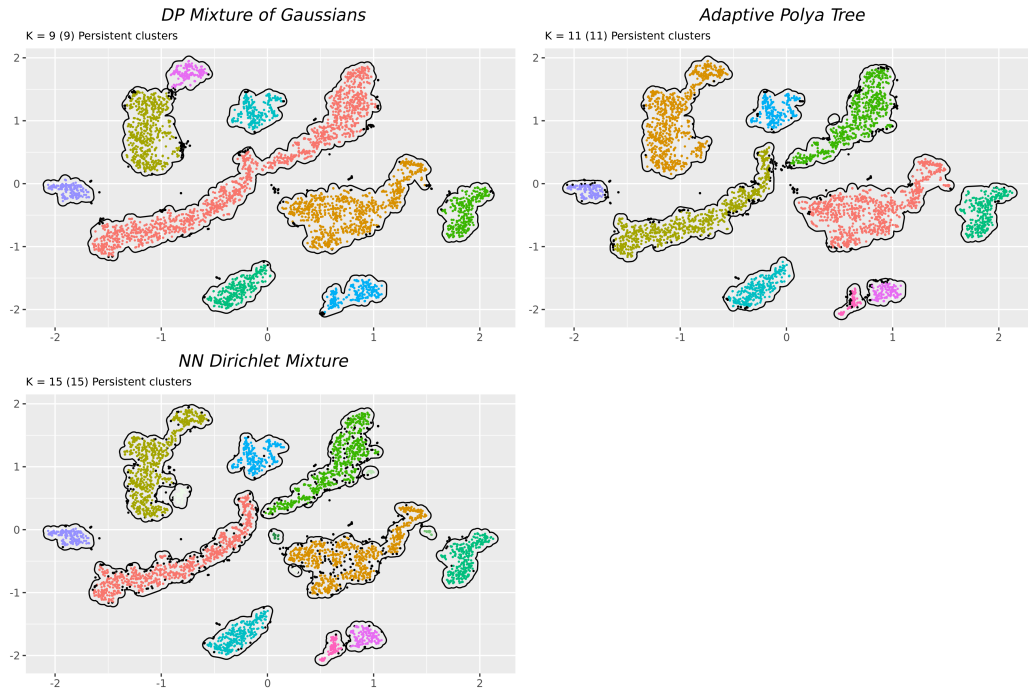


Figure 15: The persistent clusters (see Section K) across the three density models for the RNA-seq data after applying BALLET with noise levels $\nu \in \{5\%, 10\%, 15\%\}$.

function from a mixture of histograms tends to be more regular in the sense of having smaller jumps.

Let us introduce the notation to describe our model. Suppose x_i for $i = 1, \dots, n$ are independent draws from an unknown distribution with density f supported on a compact set $\mathcal{X} \subseteq \mathbb{R}^2$. We assume that f can be represented as a finite mixture $f(x; \vec{\pi}, \vec{\mathcal{B}}, \vec{\rho}) = \sum_{k=1}^K \pi_k H_k(x; \mathcal{B}_k, \vec{\rho}_k)$ of $K \in \mathbb{N}$ histogram densities, where $\vec{\pi} = (\pi_1, \dots, \pi_K)$ is a vector of non-negative weights whose coordinates sum to one. For a given $k \in [K]$, the histogram density $H_k(x; \mathcal{B}_k, \vec{\rho}_k) = \sum_{m=1}^M \mathbb{1}_{(x \in B_{km})} \rho_{km}$ is a step-function based on a partition $\mathcal{B}_k = \{B_{k1}, \dots, B_{kM}\}$ of size M of $\mathcal{X}$ and a set of associated density values $\vec{\rho}_k = (\rho_{km})_{m=1}^M$. For simplicity, we fix $|\mathcal{B}_k| = M$ for all $k = 1, \dots, K$.

It is convenient to view this model in terms of an equivalent augmented-data representation, associating a latent variable Z_i with each observation x_i , so that $f(x_i; Z_i, \vec{\mathcal{B}}, \vec{\rho}) = \sum_{k=1}^K \mathbb{1}_{(Z_i=k)} H_k(x_i; \mathcal{B}_k, \vec{\rho}_k)$ and $\Pr(Z_i = k | \vec{\pi}) = \pi_k$ for each $k \in \{1, \dots, K\}$. We denote the complete set of observations as $\mathcal{D} = \{x_1, \dots, x_n\}$ and the latent histogram allocation variables as $\mathcal{Z} = \{Z_1, \dots, Z_n\}$.

For simplicity, we also assume that $\mathcal{X} = [a, b] \times [c, d]$ and $\mathcal{B}_k$ is a grid-based (or product-based) partition of $\mathcal{X}$. More precisely, we assume that there is a partition $\mathcal{U}_k = \{U_{k1}, \dots, U_{kM'}\}$ of $[a, b]$ and $\mathcal{V}_k = \{V_{k1}, \dots, V_{kM'}\}$ of $[c, d]$ so that $\mathcal{B}_k = \{U \times V | U \in \mathcal{U}_k, V \in \mathcal{V}_k\}$ and $M = M'^2$. We further assume that partitions $\mathcal{U}_k, \mathcal{V}_k$ are constructed based on grid points $\vec{u}_k = \{u_{k0}, \dots, u_{kM'}\}$, $\vec{v}_k = \{v_{k0}, \dots, v_{kM'}\}$ such that $U_{k1} = [u_{k0}, u_{k1}]$, $V_{k1} = [v_{k0}, v_{k1}]$ and $U_{km} = (u_{k,m-1}, u_{k,m}]$ and $V_{km} = (v_{k,m-1}, v_{k,m}]$ for $2 < m \leq M'$.

F.1 Prior distribution on parameters

We now describe our prior distribution for the parameters of the mixture of histograms model. Focusing first on the partition $\mathcal{B}_k$, denote $u_{km} = a + (b - a) \sum_{j=1}^m u'_{kj}$ and $v_{km} = c + (d - c) \sum_{j=1}^m v'_{kj}$ so that $\vec{u}'_k = (u'_{k1}, \dots, u'_{kM'})$ and $\vec{v}'_k = (v'_{k1}, \dots, v'_{kM'})$ lie on the probability simplex. We specify our prior on $\mathcal{U}_k$ and $\mathcal{V}_k$ (and thus $\mathcal{B}_k$) by assuming that $\vec{u}'_k \sim \text{Dirichlet}(\alpha_b \mathbf{1}_{M'})$ and $\vec{v}'_k \sim \text{Dirichlet}(\alpha_b \mathbf{1}_{M'})$ are independent. The parameters M' and α_b can be thought of as controlling the bin resolution and regularity for the histograms, respectively. In our sky survey analysis we set $M' = 50$ ($M = 2500$) and $\alpha_b = 5$.

After specifying our prior for $\mathcal{B}_k$, we complete our prior specification for the histogram H_k by describing our prior for $\vec{\rho}_k$ given $\mathcal{B}_k$. Since H_k is a density that integrates to one, $\vec{\rho}_k$ should satisfy the constraint $\sum_{m=1}^M \rho_{km} A_{km} = 1$ where A_{km} denotes the Lebesgue measure of bin B_{km} . Thus, rather than directly placing a prior on $\vec{\rho}_k$, we place a Dirichlet prior on the parameter $\vec{p}_k = (p_{k1}, \dots, p_{kM})$, where $p_{km} = A_{km} \rho_{km}$ denotes the probability mass assigned to bin B_{km} by the histogram H_k . Thus we suppose $\vec{p}_k | \mathcal{B}_k \sim \text{Dirichlet}(\alpha_d \frac{A_{k1}}{A}, \dots, \alpha_d \frac{A_{kM}}{A})$, choosing $\alpha_d = 1$ as a default.

Finally, we complete our prior specification on the mixture of histograms model for the unknown density f by choosing to treat all parameters $\{\{\mathcal{B}_1, \vec{\rho}_1\}, \dots, \{\mathcal{B}_K, \vec{\rho}_K\}\}$ of the K histograms as *a priori* independent and fixing the weights $\vec{\pi} = \{\frac{1}{K}, \dots, \frac{1}{K}\}$. In our sky survey analysis we set $K = 50$.

F.2 Fast posterior sampling by clipping dependence

We are interested in quickly sampling from the posterior distribution of the density $f | \mathcal{D}$ when the number of observations n is large. Typically, one would draw samples from the joint posterior $\{\{\mathcal{B}_1, \vec{\rho}_1\}, \dots, \{\mathcal{B}_K, \vec{\rho}_K\}\}, \mathcal{Z} | \mathcal{D}$, and then, marginalizing over the uncertainty in $\mathcal{Z}$, use the samples of the histogram parameters to construct a posterior on f . A sampling algorithm designed to converge to this high-dimensional joint posterior object would be extremely computationally intensive, especially given our large sample size, and would likely require an unacceptably large number of samples to converge. Hence, we simplify inferences via a modular Bayes approach similar to that in Liu et al. (2009).

Specifically, to update $\vec{\mathcal{B}} = \{\mathcal{B}_1, \dots, \mathcal{B}_K\}$, we sample from its prior distribution rather than its conditional distribution given the data and other parameters, effectively clipping the dependence of the bin parameters on the other components of the model as described in Liu et al. (2009). Furthermore, we draw only one sample $\vec{\mathcal{B}}^* = \{\mathcal{B}_1^*, \dots, \mathcal{B}_K^*\}$ from the prior distribution on $\vec{\mathcal{B}}$, and reuse this same collection $\vec{\mathcal{B}}^*$ of histogram bins for each round of new samples for the other parameters.

In addition, rather than iterate between sampling $\vec{p}_k$ from its full conditional,

$$\vec{p}_k | \mathcal{D}, \mathcal{Z}, \mathcal{B}_k^* \sim \text{Dirichlet}\left(\sum_{i=1}^n \mathbb{1}_{(x_i \in B_{k1})} \mathbb{1}_{(Z_i=k)} + \alpha_d \frac{A_{k1}}{A}, \dots, \sum_{i=1}^n \mathbb{1}_{(x_i \in B_{kM})} \mathbb{1}_{(Z_i=k)} + \alpha_d \frac{A_{kM}}{A}\right),$$

and alternately sampling $\mathcal{Z}$ from its full conditional, we marginalize the log density of $\vec{p}_k | \mathcal{D}, \mathcal{Z}, \mathcal{B}_k^*$ with respect to the prior distribution on $\mathcal{Z}$ yielding the distribution

$$\vec{p}_k | \mathcal{D}, \mathcal{B}_k^* \sim \text{Dirichlet}\left(\frac{N_{k1}}{K} + \alpha_d \frac{A_{k1}}{A}, \dots, \frac{N_{kM}}{K} + \alpha_d \frac{A_{kM}}{A}\right), \quad (10)$$

which we use in place of the posterior distribution of $\vec{p}_k$ given $\mathcal{B}_k^*$ and $\mathcal{D}$. Here $N_{km} = \sum_{i=1}^n \mathbb{1}_{(x_i \in B_{km}^*)}$ denotes the number of observations that fall into the bin $B_{km}^* \in \mathcal{B}_k^*$.

The resulting algorithm is a fast way to generate independent samples from an approximate modular posterior for $f(\mathcal{D})$. This sampler runs almost instantaneously on a personal laptop computer even for sample sizes of $n \approx 40,000$, which would be prohibitive for traditional Markov chain Monte Carlo algorithms for density estimation models. Moreover, the samples appear to appropriately reflect our uncertainty in the underlying data-generating density in our experiments.

Appendix G. Additional results from the analysis of the synthetic sky survey data

Including a diversity of sizes among the synthetic galaxy clusters led to datasets that more closely resembled the observed data, and it also made the true clusters more challenging to recover with both clustering methods. Hence, we simulated the weights of the active components from a symmetric Dirichlet distribution with a small concentration parameter. The relative weights of the “galaxy clusters” for one of the 100 synthetic datasets we analyzed are visualized in Figure 16. The specific synthetic data set associated with these weights is shown in Figure 17.

Figure 19 shows how the performance of DBSCAN is highly sensitive to the choice of tuning parameter. It is interesting to note that the optimal parameters in this application are far

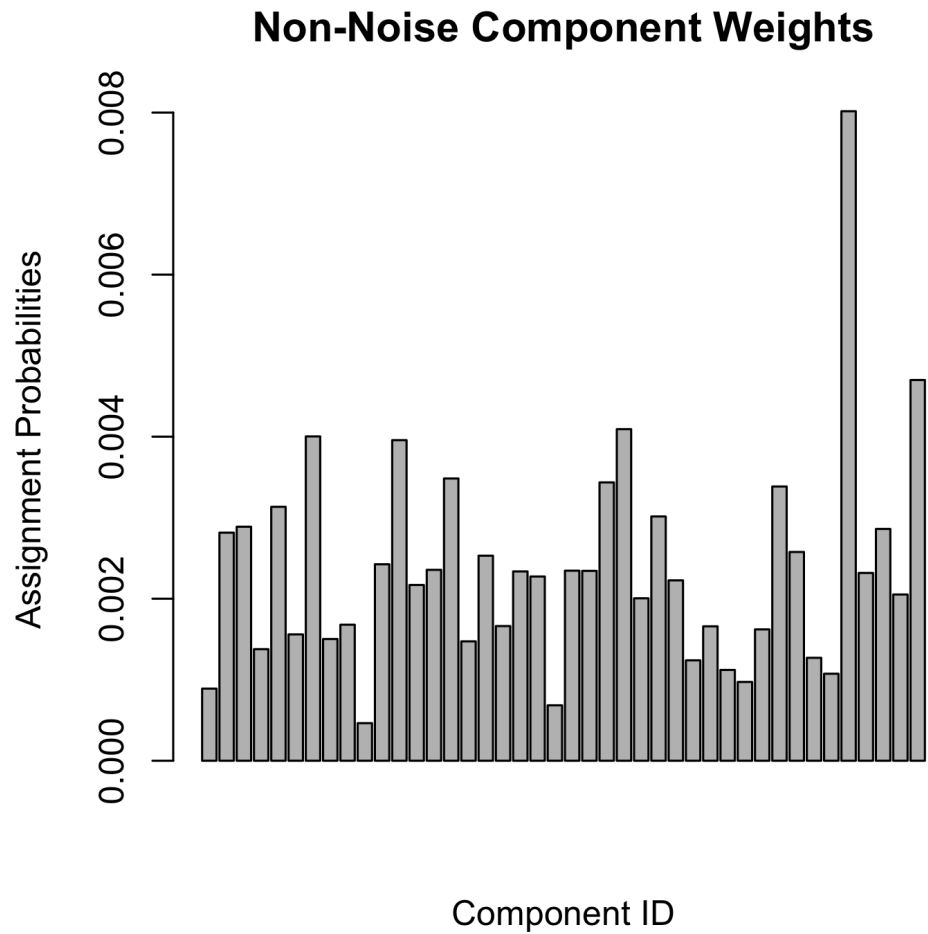


Figure 16: Relative sizes (mixture weights) of the non-noise components in one of our synthetic sky survey datasets.

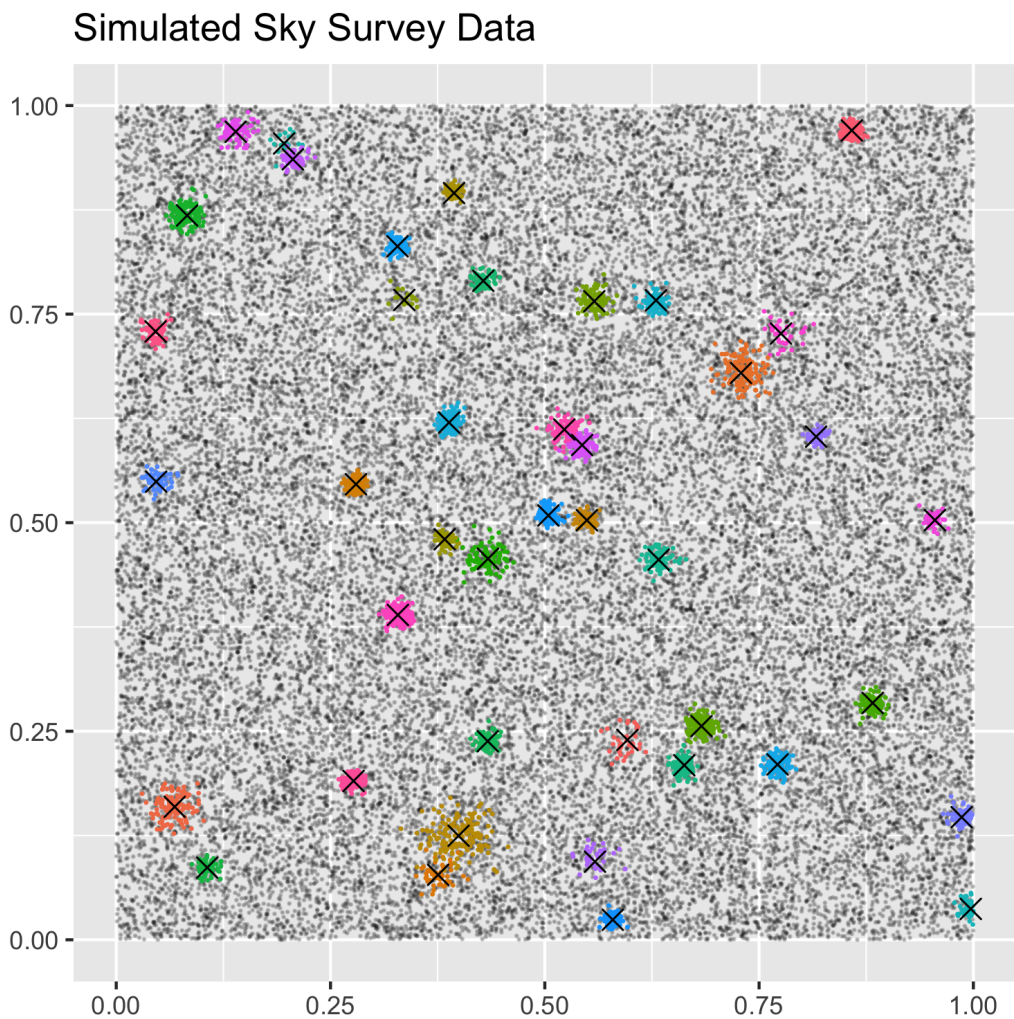


Figure 17: One of our synthetic sky survey datasets. Observations drawn from one of the high-density components are given bright colors, and each of their centers is marked with an $\times$. Observations drawn from the uniform background are colored grey and made translucent.

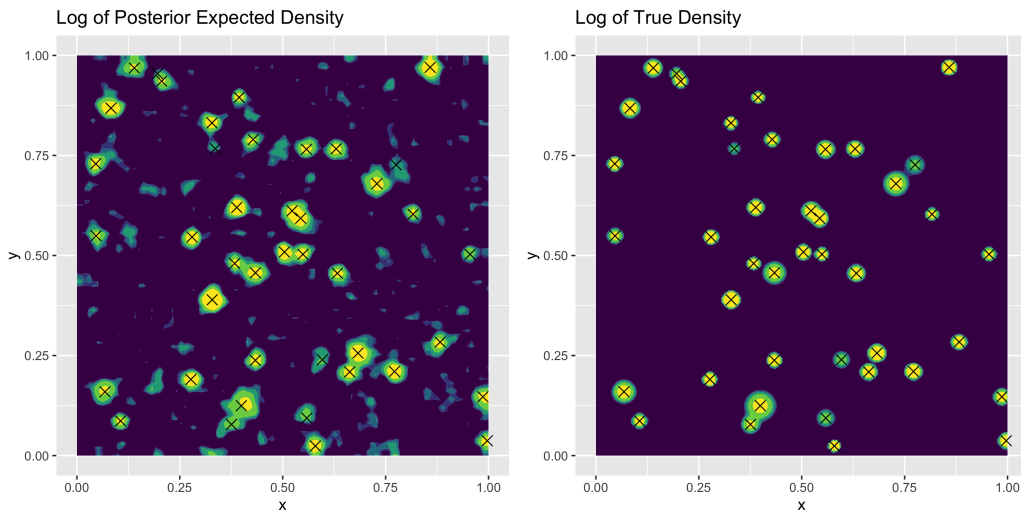


Figure 18: Comparison of $\log(\hat{f})$ and $\log(f)$, where $\hat{f}$ is the posterior expectation of f under our mixture of random histograms model fitted to the data in Figure 17.

from the values suggested by the heuristics proposed in Schubert et al. (2017), suggesting that in general they will be highly context dependent. We show the performance of optimally tuned DBSCAN in Figure 20, noting that this tuning procedure required knowledge of the ground truth. The bounds of the 95% credible ball of the BALLET point estimate for the synthetic data are shown in Figure 21. The associated BALLET point estimate is shown in Figure 4 of the main document. The complete results of the sensitivity and specificity of the various point estimates and bounds considered, averaged over the 100 synthetic datasets, are presented in Table 2.

	DBSCAN	DBSCAN ¹	BALLET Lower	BALLET Est.	BALLET Upper	BALLET Plugin
Sensitivity	0.86	0.79	0.62	0.78	0.89	0.78
Specificity	0.49	0.99	0.99	0.99	0.96	0.99
Exact Match	0.45	0.88	0.90	0.87	0.83	0.88

Table 2: Averaged results from applying BALLET and DBSCAN to 100 replicates of the synthetic sky survey data. For BALLET, we also provide the performance of upper and lower bounds for a 95% credible ball centered at the point estimate. For DBSCAN, we provide averaged sensitivity and specificity for both our default choice of its tuning parameter and for its optimized parameter choice indicated as DBSCAN¹ (see Figure 19).

Appendix H. Additional results from analysis of the sky survey data

In this section we provide additional results from the analysis of the Edinburgh-Durham Southern Galaxy Catalogue data which appeared in Section 6 of the main text. In particular, we visualize the log of the posterior expectation of the data-generating density in Figure

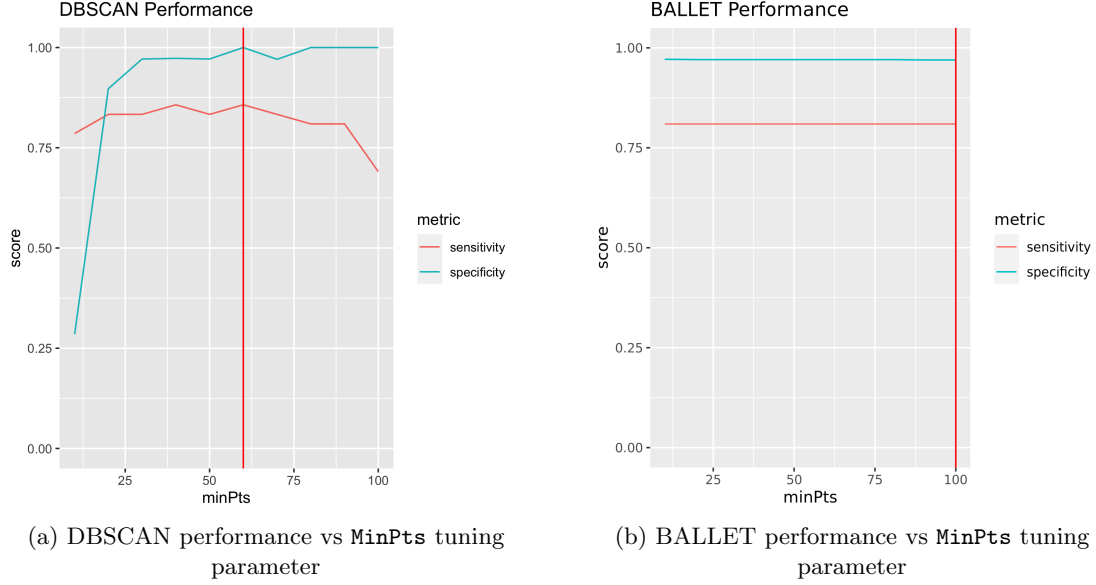


Figure 19: The performance of **BALLET** and **DBSCAN** clusters as the tuning parameter k (or equivalently **MinPts**) varies. Vertical lines call attention to the value of k that exhibits the “best” performance, as determined by the sum of the sensitivity and specificity.

22, **DBSCAN** and **BALLET** fits based on our default value of **MinPts** = $k_0 = \lceil \log_2(n) \rceil$ in Figures 23 and 24, and an alternative **DBSCAN** fit using the optimal tuning parameters from the simulation study in Figure 25. We present tabular results collecting the rate of coverage of the EDCCI and Abell catalogs, by the various point estimates and bounds we have considered, in Tables 1 and 3, respectively.

	DBSCAN	DBSCAN ¹	BALLET Lower	BALLET Est.	BALLET Upper	BALLET Plugin
Sensitivity	0.40	0.37	0.21	0.40	0.56	0.40
Specificity	0.15	0.43	0.73	0.40	0.34	0.42
Exact Match	0.13	0.35	0.67	0.26	0.29	0.28

Table 3: **DBSCAN** and **BALLET** clustering coverage of the suspected galaxy clusters listed in the Abell catalog. The column labeled **DBSCAN** reports the performance of the method with the default value of **MinPts** = 16, while **DBSCAN¹** shows the performance of the method with the optimal value of **MinPts** = 60 chosen based on our simulation study.

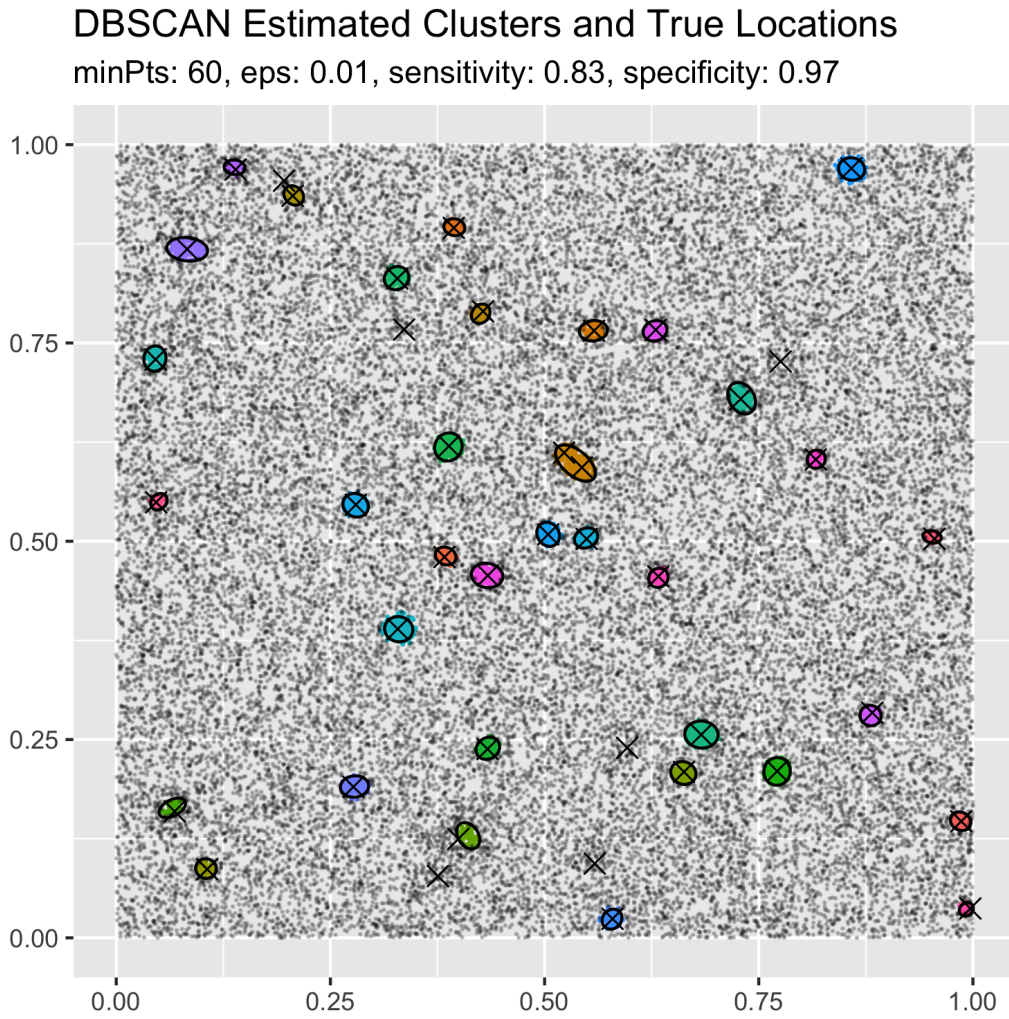


Figure 20: The result of fitting DBSCAN to the particular synthetic sky survey data using the optimal value of `MinPts` based on our simulation study.

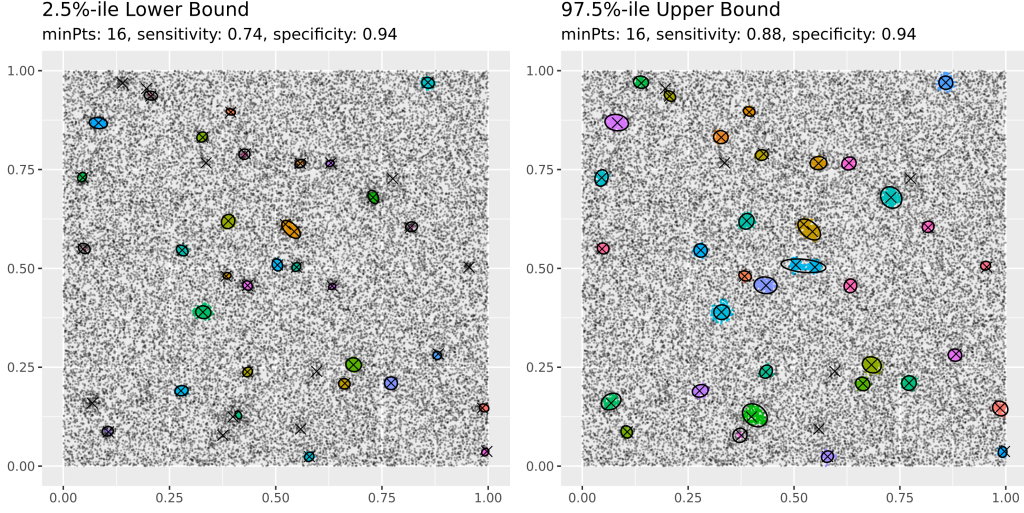


Figure 21: Upper and lower bounds for the 95% credible ball centered at our BALLET clustering estimate for the particular synthetic dataset shown in Figure 17.

Appendix I. Theory details from Section 4

I.1 Proof of Theorem 1

The proof is a simple application of the metric properties of D . In particular,

$$\begin{aligned} D\{\hat{\psi}_M(\mathcal{X}_n), \psi(f_0)\} &\leq E_{f \sim P_M(\cdot|\mathcal{X}_n)} D\{\tilde{\psi}(f), \psi(f_0)\} + E_{f \sim P_M(\cdot|\mathcal{X}_n)} D\{\tilde{\psi}(f), \hat{\psi}_M(\mathcal{X}_n)\} \\ &\leq 2E_{f \sim P_M(\cdot|\mathcal{X}_n)} D\{\tilde{\psi}(f), \psi(f_0)\}, \end{aligned}$$

where the first line follows by taking expectation with respect to the posterior distribution $P_M(\cdot|\mathcal{X}_n)$ after using the triangle inequality and symmetry for the metric D , while the second line follows by noting that the second term in the right hand side of the first line is no greater than the first term, since $\hat{\psi}_M(\mathcal{X}_n)$ is given by (8). Noting further that D is bounded above by one, we obtain

$$\begin{aligned} E_{f \sim P_M(\cdot|\mathcal{X}_n)} D\{\tilde{\psi}(f), \psi(f_0)\} &\leq P_M(f : \rho(f, f_0) > K_n \epsilon_n | \mathcal{X}_n) + \sup_{f : \rho(f, f_0) \leq K_n \epsilon_n} D\{\tilde{\psi}(f), \psi(f_0)\} \\ &= \tau_1(\mathcal{X}_n) + \tau_2(\mathcal{X}_n), \end{aligned}$$

where τ_1 is defined in Assumption 2 and τ_2 and constant K_n are as defined in Assumption 3. Since $K_n \rightarrow \infty$, these assumptions show that $\tau_1(\mathcal{X}_n), \tau_2(\mathcal{X}_n) \xrightarrow{P} 0$ as $n \rightarrow \infty$.

I.2 Proof of Theorem 2

In order to simplify the presentation of our proof we first introduce some notation. We note that any sub-partition $\mathcal{C} = \{C_1, \dots, C_k\} \in \mathcal{P}(\mathcal{X}_n)$ defines a binary “co-clustering” relation $\mathcal{C}_R : \mathcal{X}_n \times \mathcal{X}_n \rightarrow \{0, 1\}$ on pairs of data points, namely

$$\mathcal{C}_R(x, y) \doteq \mathbb{1}_{(x \notin A, y \notin A)} + \sum_{h=1}^k \mathbb{1}_{(x \in C_h, y \in C_h)}$$

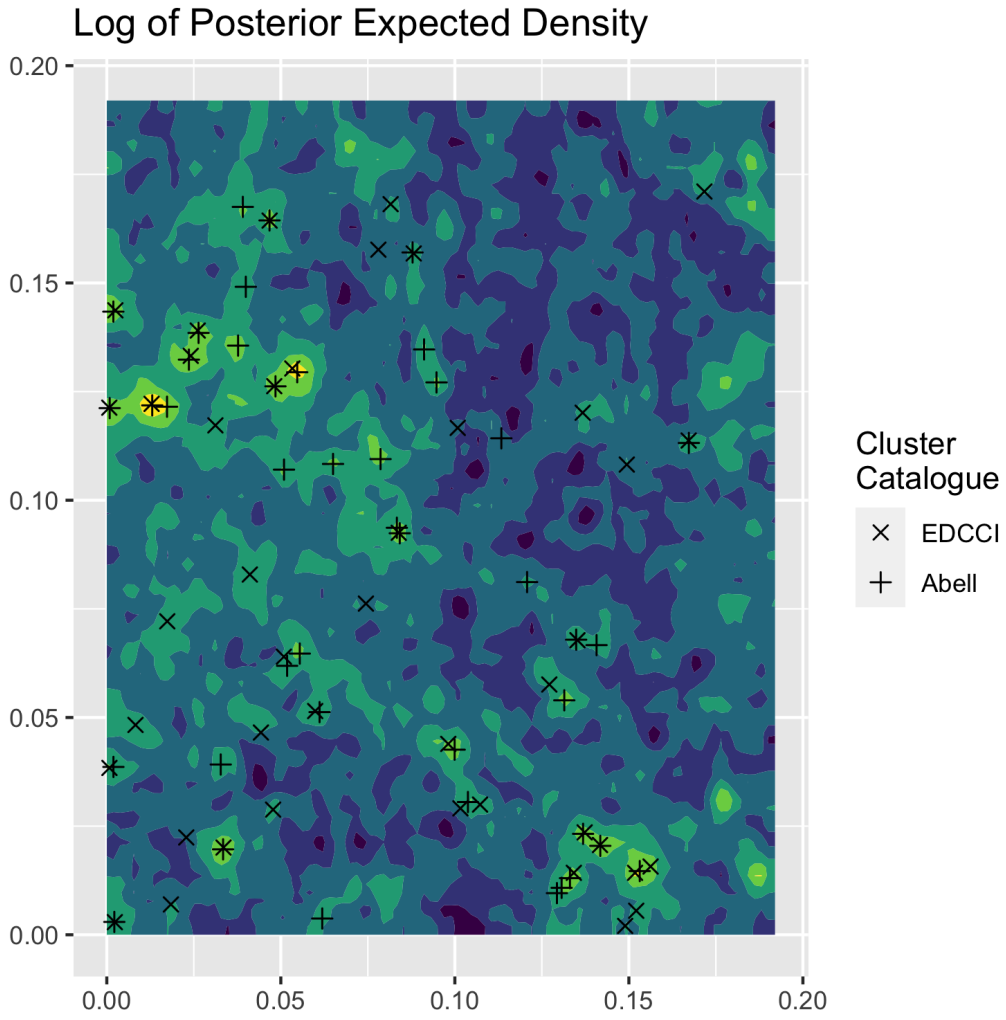


Figure 22: Log of the posterior expectation of the density for the Edinburgh-Durham Southern Galaxy Catalogue data under our mixture of random histograms model. For reference, we have superimposed galaxy clusters reported in the EDCCI and Abell cluster catalogs.

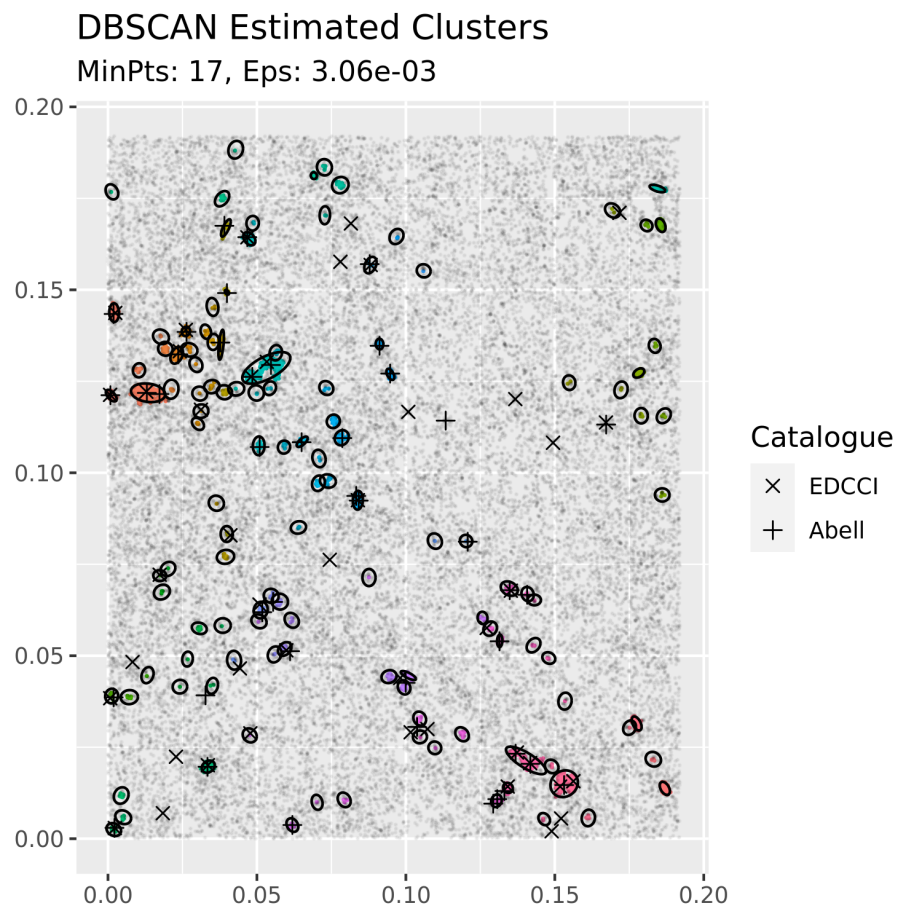


Figure 23: Result of applying DBSCAN to the Edinburgh-Durham Southern Galaxy Catalogue data using our default value of `MinPts`. Cluster centers from the two previously proposed cluster catalogs are plotted with black '+'s (Abell Catalog) and 'X's (EDCCI).

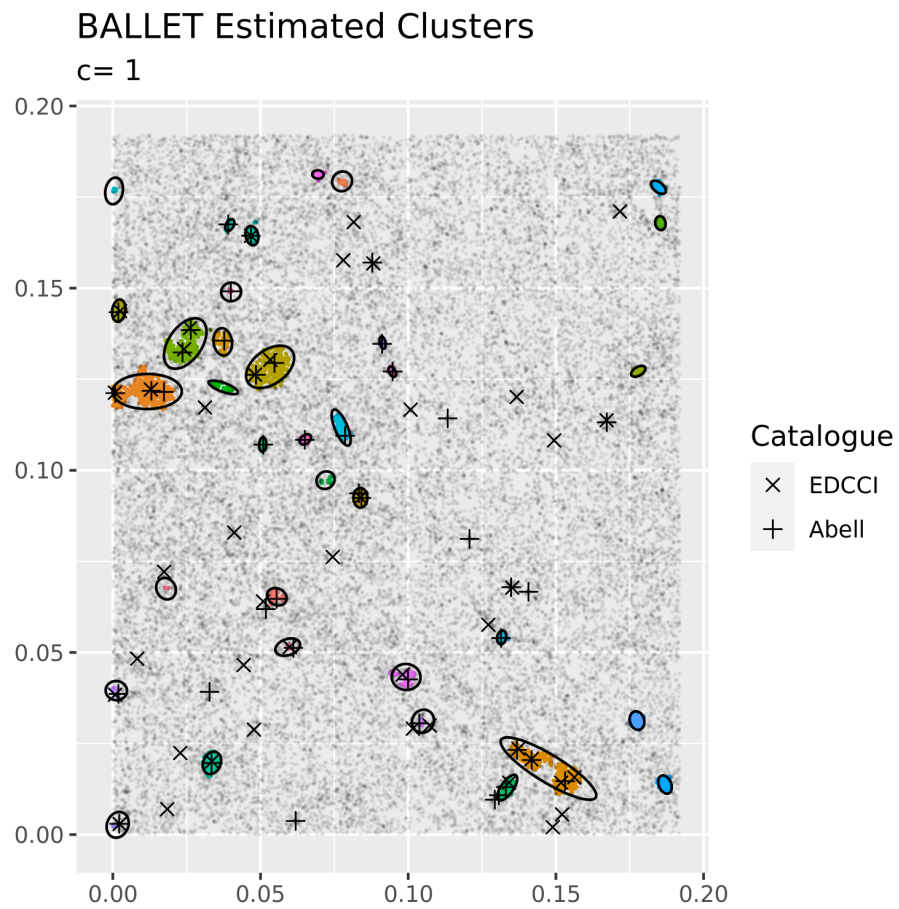


Figure 24: Results of applying BALLET Clustering to the Edinburgh-Durham Southern Galaxy Catalogue data, with 95% credible bounds presented in Figure 5.

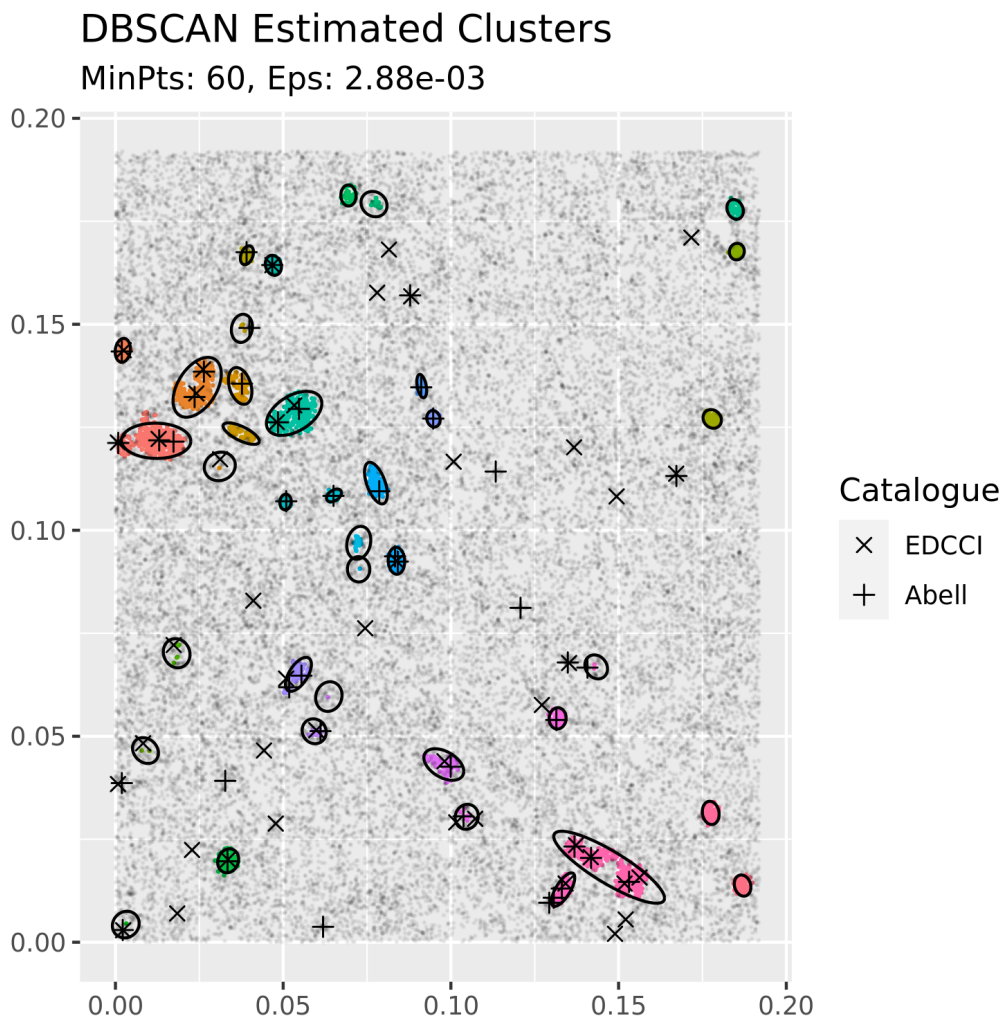


Figure 25: Result of applying DBSCAN to the Edinburgh-Durham Southern Galaxy Catalogue data using the tuning parameter that had optimal performance in our simulation study (Figure 19).

where $A = \cup_{h=1}^k C_h$ is the set of active points in $\mathcal{C}$. In other words, $\mathcal{C}_R(x, y) = 1$ if $x, y \in \mathcal{X}_n$ are both noise points or if they belong to a common cluster in $\mathcal{C}$, and $\mathcal{C}_R(x, y) = 0$ otherwise. Given $\mathcal{C}$, we can also obtain an indicator function of active points $\mathcal{C}_A : \mathcal{X}_n \rightarrow \{0, 1\}$ such that $\mathcal{C}_A(x) = 1$ if and only if $x \in A$. In fact, knowing the binary functions $\mathcal{C}_R$ and $\mathcal{C}_A$ is sufficient to uniquely recover the sub-partition $\mathcal{C} \in \mathcal{P}(\mathcal{X}_n)$. Indeed, this follows because $\mathcal{C}_R$ is an equivalence relation on $\mathcal{X}_n$, and the sub-partition $\mathcal{C}$ can be recovered by dropping the inactive subset $\mathcal{C}_A^{-1}(0)$ from the equivalence partition of $\mathcal{X}_n$ induced by $\mathcal{C}_R$.

We also introduce the following subscript-free notation for summation of a symmetric function $F : \mathcal{X}_n \times \mathcal{X}_n \rightarrow \mathbb{R}$ over pairs of distinct data points that lie in $S \subseteq \mathcal{X}$:

$$\sum_{x \neq y \in \mathcal{X}_n \cap S} F(x, y) \doteq \sum_{\substack{1 \leq i < j \leq n \\ x_i, x_j \in S}} F(x_i, x_j) = \frac{1}{2} \sum_{\substack{i, j \in [n] \\ x_i, x_j \in S}} F(x_i, x_j) \mathbb{1}_{(i \neq j)}.$$

Proof of Theorem 2 Similar to analyses of Binder's loss, the first step in our proof is to note that $L_{\text{IA-Binder}}$ can be written as a sum of pairwise losses $\phi_{x,y}$ over pairs $x, y \in \mathcal{X}_n$. In particular, fix any $\mathcal{C}, \mathcal{C}' \in \mathcal{P}(\mathcal{X}_n)$, and let $A = \mathcal{C}_A^{-1}(1)$, $A' = \mathcal{C}'_A^{-1}(1)$ and $I = \mathcal{C}_A^{-1}(0)$, $I' = \mathcal{C}'_A^{-1}(0)$ denote the active and inactive sets of $\mathcal{C}$ and $\mathcal{C}'$, respectively. Taking $a = b$ and $m = m_{ia} = m_{ai}$ in (4), we note

$$\begin{aligned} L_{\text{IA-Binder}}(\mathcal{C}, \mathcal{C}') &= m(n-1)(|A \cap I'| + |I \cap A'|) + a \sum_{\substack{1 \leq i < j \leq n \\ x_i, x_j \in A \cap A'}} \mathbb{1}_{\{\mathcal{C}_R(x_i, x_j) \neq \mathcal{C}'_R(x_i, x_j)\}} \\ &= \sum_{x \neq y \in \mathcal{X}_n} \phi_{x,y}(\mathcal{C}, \mathcal{C}') \end{aligned} \quad (11)$$

where

$$\phi_{x,y}(\mathcal{C}, \mathcal{C}') = m \mathbb{1}_{\{\mathcal{C}_A(x) \neq \mathcal{C}'_A(x)\}} + m \mathbb{1}_{\{\mathcal{C}_A(y) \neq \mathcal{C}'_A(y)\}} + a \mathbb{1}_{\{\mathcal{C}_R(x,y) \neq \mathcal{C}'_R(x,y)\}} \mathbb{1}_{\{\mathcal{C}_A(x) = \mathcal{C}'_A(x) = \mathcal{C}_A(y) = \mathcal{C}'_A(y)\}}.$$

In order to obtain (11), we have used the fact that the last term in $\phi_{x,y}(\mathcal{C}, \mathcal{C}')$ is zero when either one of x or y is outside the set $A \cap A'$, and the fact that the summation $\sum_{x \neq y \in \mathcal{X}_n}$ over the first two terms in $\phi_{x,y}(\mathcal{C}, \mathcal{C}')$ is equal to $m(n-1)(|A \cap I'| + |I \cap A'|)$.

Now we shall use (11) to show that $D = \binom{n}{2}^{-1} L_{\text{IA-Binder}}$ is a metric that is bounded above by one when $a, m \leq 1$. Note that at most one out of the three indicator variables in $\phi_{x,y}$ can be non-zero for any instance, and hence $\phi_{x,y}$ is bounded above by one (in fact by $\max(a, m) \leq 1$) for each of the $\binom{n}{2}$ summation variables $x \neq y \in \mathcal{X}_n$. This shows that D is also bounded above by one. Further, the symmetry of D in its arguments follows from the symmetry of $\phi_{x,y}$ in its arguments for every $x \neq y \in \mathcal{X}_n$.

Next suppose $D(\mathcal{C}, \mathcal{C}') = 0$. Since the functions $\phi_{x,y}$ are non-negative, this shows that $\phi_{x,y}(\mathcal{C}, \mathcal{C}') = 0$ for each $x \neq y \in \mathcal{X}_n$. Since $2m \geq a > 0$, the functions $\mathcal{C}_A$ and $\mathcal{C}'_A$ are equal (or equivalently that $A = A'$), and further that $\mathcal{C}_R(x, y) = \mathcal{C}'_R(x, y)$ either when $x, y \in A = A'$ or $x, y \in I = I'$. The latter condition is sufficient to show that the relations $\mathcal{C}_R$ and $\mathcal{C}'_R$ are equal since $\mathcal{C}_R(x, y) = 0 = \mathcal{C}'_R(x, y)$ when $x \in A, y \in I$ or $x \in I, y \in A$. Since the binary functions $\mathcal{C}_A$ and $\mathcal{C}_R$ determine the sub-partition $\mathcal{C}$, we have $\mathcal{C} = \mathcal{C}'$.

Finally, to demonstrate that D satisfies the triangle inequality, it suffices to show that for each $x \neq y \in \mathcal{X}_n$, we have the triangle inequality $\phi_{x,y}(\mathcal{C}, \mathcal{C}'') \leq \phi_{x,y}(\mathcal{C}, \mathcal{C}') + \phi_{x,y}(\mathcal{C}', \mathcal{C}'')$

for any sub-partitions $\mathcal{C}, \mathcal{C}', \mathcal{C}'' \in \mathcal{P}(\mathcal{X}_n)$. Indeed when either $\mathcal{C}_A(x) \neq \mathcal{C}_A''(x)$ or $\mathcal{C}_A(y) \neq \mathcal{C}_A''(y)$, the triangle inequality for $\phi_{x,y}$ follows from the inequality:

$$\mathbb{1}_{\{\mathcal{C}_A(z) \neq \mathcal{C}_A''(z)\}} \leq \mathbb{1}_{\{\mathcal{C}_A(z) \neq \mathcal{C}_A'(z)\}} + \mathbb{1}_{\{\mathcal{C}_A'(z) \neq \mathcal{C}_A''(z)\}} \quad z \in \{x, y\}.$$

Otherwise, let us assume that the previous condition does not hold. Let us further suppose that $\phi_{x,y}(\mathcal{C}, \mathcal{C}'') > 0$ or else there is nothing to show. This means that we are under the case $\phi_{x,y}(\mathcal{C}, \mathcal{C}'') = a$, $\mathcal{C}_A(x) = \mathcal{C}_A''(x) = \mathcal{C}_A(y) = \mathcal{C}_A''(y)$, and $\mathcal{C}_R(x, y) \neq \mathcal{C}_R''(x, y)$. If $\mathcal{C}_A'(x) \neq \mathcal{C}_A(x) = \mathcal{C}_A''(x)$ (or analogously $\mathcal{C}_A'(y) \neq \mathcal{C}_A(y) = \mathcal{C}_A''(y)$) then the triangle inequality is satisfied as $\phi_{x,y}(\mathcal{C}, \mathcal{C}') + \phi_{x,y}(\mathcal{C}', \mathcal{C}'') \geq m\mathbb{1}_{\{\mathcal{C}_A(x) \neq \mathcal{C}_A'(x)\}} + m\mathbb{1}_{\{\mathcal{C}_A'(x) \neq \mathcal{C}_A''(x)\}} = 2m \geq a = \phi_{x,y}(\mathcal{C}, \mathcal{C}'')$. Otherwise, the only remaining case is that $\mathcal{C}_A(x) = \mathcal{C}_A'(x) = \mathcal{C}_A''(x) = \mathcal{C}_A(y) = \mathcal{C}_A'(y) = \mathcal{C}_A''(y)$. Then the triangle inequality is satisfied since

$$\begin{aligned} \phi_{x,y}(\mathcal{C}, \mathcal{C}'') &= a\mathbb{1}_{\{\mathcal{C}_R(x,y) \neq \mathcal{C}_R''(x,y)\}} \leq a\mathbb{1}_{\{\mathcal{C}_R(x,y) \neq \mathcal{C}_R'(x,y)\}} + a\mathbb{1}_{\{\mathcal{C}_R'(x,y) \neq \mathcal{C}_R''(x,y)\}} \\ &= \phi_{x,y}(\mathcal{C}, \mathcal{C}') + \phi_{x,y}(\mathcal{C}', \mathcal{C}''). \end{aligned}$$

Hence, we have verified the triangle inequality for $\phi_{x,y}$, and hence also for D . Combined with the non-negativity of D , we have shown that D is a metric. $\blacksquare$

I.3 Proof of Theorem 4

Letting $\mathcal{X} = \mathbb{R}^d$, we begin with the necessary assumptions on the unknown data density $f_0 : \mathcal{X} \rightarrow \mathbb{R}$ and the threshold level $\lambda > 0$. Let $S_\lambda = \{x \in \mathbb{R}^d : f_0(x) \geq \lambda\}$ denote the level set of the unknown data density f_0 at threshold $\lambda \in (0, \infty)$. We make the following assumptions.

Assumption 4. (*Continuity with vanishing tails*) The density $f_0 : \mathbb{R}^d \rightarrow [0, \infty)$ is continuous and satisfies $\lim_{\|x\| \rightarrow \infty} f_0(x) = 0$.

Lemma 9. *If Assumption 4 holds then f_0 is uniformly continuous.*

Proof Fix any $\epsilon > 0$. Then since f_0 has vanishing tails, there is a $K > 0$ such that $\sup_{x \in \mathbb{R}^d \setminus [-K, K]^d} f_0(x) \leq \epsilon/2$, and since f_0 is continuous on the compact set $H = [-K - 1, K + 1]^d$, there is a $\delta \in (0, 1)$ such that $|f_0(x) - f_0(y)| \leq \epsilon$ whenever $\|x - y\| \leq \delta$ and $x, y \in H$. Finally if $x, y \in \mathbb{R}^d$ are such that $\|x - y\| \leq 1$ and $\{x, y\} \cap (\mathbb{R}^d \setminus H) \neq \emptyset$ then $x, y \in \mathbb{R}^d \setminus [-K, K]^d$. Thus $|f_0(x) - f_0(y)| \leq f_0(x) + f_0(y) \leq \epsilon/2 + \epsilon/2 = \epsilon$. Hence we have shown that there is a $\delta \in (0, 1)$ such that $|f_0(x) - f_0(y)| \leq \epsilon$ whenever $\|x - y\| \leq \delta$ and $x, y \in \mathbb{R}^d$. Since $\epsilon > 0$ is arbitrary, f_0 is uniformly continuous. $\blacksquare$

Assumption 5. (*Fast mass decay around level λ*) There are constants $C, \bar{\epsilon} > 0$ such that $\int_{\{x \in \mathbb{R}^d : |f_0(x) - \lambda| \leq \epsilon\}} f_0(x) dx \leq C\epsilon$ for all $\epsilon \in (0, \bar{\epsilon})$.

Assumption 5 is adapted from Rinaldo and Wasserman (2010), and intuitively prevents the density from being too flat around the level λ . In particular, if f_0 satisfies $\|\nabla f_0(x)\| > 0$ for Lebesgue-almost-every x , then Lemma 4 in Rinaldo and Wasserman (2010) shows that

Assumption 5 will hold for Lebesgue-almost-every $\lambda \in (0, \|f_0\|_\infty)$. Additionally, if f_0 is smooth and has a compact support, the authors show that the set of $\lambda \in (0, \|f_0\|_\infty)$ for which Assumption 5 does not hold is finite.

Assumption 6. (*Stable connected components at level λ*) For any $\lambda_l < \lambda_h \in [\lambda - \bar{\varepsilon}, \lambda + \bar{\varepsilon}]$, and $x, y \in S_{\lambda_h}$:

1. If x, y are disconnected in S_{λ_h} , then x, y are also disconnected in S_{λ_l} .
2. If x, y are connected in S_{λ_l} , then x, y are also connected in S_{λ_h} .

Informally, Assumption 6 states that the connected components of the level-set $S_{\lambda'}$ do not merge or split as λ' varies between $(\lambda - \bar{\varepsilon}, \lambda + \bar{\varepsilon})$. When combined with Assumption 4, this assumption ensures that the level set clusters vary continuously with respect to the level λ . Various versions of such assumptions have previously appeared in the literature like Assumption C2 in Rinaldo and Wasserman (2010) and Definition 2.1 in Sriperumbudur and Steinwart (2012).

We now prove some intermediate theory on level set estimation that will be useful in the proof of Theorem 4. Given data points $\mathcal{X}_n = \{x_1, \dots, x_n\}$ suppose we have a density estimator f that approximates f_0 . For a suitably small choice of $\delta > 0$, we estimate the level set S_λ by the δ diameter tube around the active data points, namely:

$$T_\delta(A_{f,\lambda}) = \bigcup_{x \in A_{f,\lambda}} B(x, \delta/2),$$

where $A_{f,\lambda} = \{x \in \mathcal{X}_n : f(x) \geq \lambda\}$ is the set of active data points and $B(x, \delta/2)$ is the open ball of radius $\delta/2$ around x . To emphasize that $T_\delta(A_{f,\lambda})$ is an estimator for S_λ , we denote it as $\hat{S}_{\delta,\lambda}(f) \doteq T_\delta(A_{f,\lambda})$ in the sequel.

The following lemma shows that the level set estimator $\hat{S}_{\delta,\lambda}(f)$ approximates the level sets of the original density S_λ as long as the quantities $\|f_0 - f\|_\infty$ and $\delta > 0$ are suitably small. This result extends Lemma 3.2 in Sriperumbudur and Steinwart (2012) to the case when f is an arbitrary approximation to f_0 . Our proof hinges on using Corollary 12 below rather than a specific kernel density estimator as in Sriperumbudur and Steinwart (2012).

Lemma 10. *Suppose $\mathcal{X} = \mathbb{R}^d$ and $f_0 : \mathcal{X} \rightarrow [0, \infty)$ is uniformly continuous. Then*

$$H_{f_0}(\eta) \doteq \max\{h \geq 0 : \sup_{x,y \in \mathcal{X}, \|x-y\| \leq h} |f_0(x) - f_0(y)| \leq \eta\} \quad (12)$$

is a positive number for each $\eta > 0$. Given observations $x_1, \dots, x_n$ drawn independently from f_0 with $n \geq 16$, with probability at least $1 - 1/n$ we have

$$S_{(\lambda + \|f_0 - f\|_\infty + \eta)} \subseteq \hat{S}_{\delta,\lambda}(f) \subseteq S_{(\lambda - \|f_0 - f\|_\infty - \eta)},$$

uniformly over all functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$, and constants $\eta, \lambda > 0$ such that $\delta \in [r_{n,\lambda,d}, 2H_{f_0}(\eta)]$, where $r_{n,\lambda,d} \doteq 2 \left(\frac{16d \ln n}{nv_d \lambda} \right)^{1/d}$ and v_d is the volume of the unit Euclidean ball in $\mathbb{R}^d$.

Before we prove the above lemma, we will establish Corollary 12 which provides a lower-bound on the parameter δ to ensure that the δ -ball centered around any point in the level set S_λ will contain at least one observed sample. This is a corollary of the uniform law of large numbers result from Boucheron et al. (2005). We use the following version:

Lemma 11. *(Chaudhuri and Dasgupta, 2010, Theorem 15) Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ to $\{0, 1\}$ with VC dimension $d < \infty$, and let P be a probability distribution on $\mathcal{X}$. Let E denote the expectation with respect to P . Suppose n points are drawn independently from P , and let E_n denote expectation with respect to this sample. Then for any $\delta > 0$,*

$$-\min(\beta_n^2 + \beta_n \sqrt{Eg}, \beta_n \sqrt{E_n g}) \leq Eg - E_n g \leq \min(\beta_n^2 + \beta_n \sqrt{E_n g}, \beta_n \sqrt{Eg})$$

holds for all $g \in \mathcal{G}$ with probability at least $1 - \delta$, where $\beta_n = \sqrt{(4/n)\{d \ln 2n + \ln(8/\delta)\}}$.

Corollary 12. *Suppose $\mathcal{X}_n = \{x_1, \dots, x_n\}$ are drawn independently from f_0 and $n \geq 16$. Then with probability at least $1 - 1/n$, we have $\mathcal{X}_n \cap B \neq \emptyset$ for each Euclidean ball $B \subseteq \mathbb{R}^d$ such that $\int_B f_0(x) dx \geq \frac{16d \ln n}{n}$.*

Proof Let $\mathcal{G} = \{\mathbb{1}_{B(x,r)} | x \in \mathbb{R}^d \text{ and } r > 0\}$ be the class of indicator functions of all the Euclidean balls, and note that the VC dimension of spheres in $\mathbb{R}^d$ is $d + 1$ (e.g. Wainwright (2019)). Lemma 11 then states that with probability at least $1 - 1/n$,

$$P(B) - P_n(B) \leq \beta_n \sqrt{P(B)}$$

for any Euclidean ball $B \subseteq \mathbb{R}^d$, where $P_n(B) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(x_i \in B)}$ is the empirical distribution function and $\beta_n = \sqrt{(4/n)\{(d+1) \ln(2n) + \ln(8n)\}}$. In particular, as long as this event holds and $P(B) > \beta_n^2$, one has $P_n(B) > 0$ and hence $\mathcal{X}_n \cap B \neq \emptyset$. The proof is completed by noting that $\beta_n^2 \leq \frac{16d \ln n}{n}$ whenever $n \geq 16$. $\blacksquare$

Proof of Lemma 10 With probability at least $1 - 1/n$ the event in Corollary 12 holds; we will henceforth condition on this event. Next, let $v_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ be the volume of the unit Euclidean sphere in d dimensions and note that $\lambda v_d (\delta/2)^d \geq \frac{16d \ln n}{n}$ whenever $\delta \geq r_{n,\lambda,d} \doteq 2 \left(\frac{16d \ln n}{n v_d \lambda} \right)^{1/d}$. This shows that for any $x \in \mathcal{X}$

$$\mathcal{X}_n \cap B(x, \delta/2) \neq \emptyset \quad \text{whenever} \quad \inf_{y \in B(x, \delta/2)} f_0(y) \geq \lambda, \quad (13)$$

and further since $\delta/2 \leq H_{f_0}(\eta)$ that

$$\sup_{y \in B(x, \delta/2)} |f_0(y) - f_0(x)| \leq \eta. \quad (14)$$

We are now ready to prove our main statement in Lemma 10. We first show the inclusion $\hat{S}_{\delta,\lambda}(f) \subseteq S_{(\lambda - \|f_0 - f\|_\infty - \eta)}$. Indeed, for any $x \in \hat{S}_{\delta,\lambda}(f)$ there is a $y \in \mathcal{X}_n$ such that $x \in B(y, \delta/2)$ and $f(y) \geq \lambda$. The inequalities

$$f_0(x) \geq f_0(y) - \eta \geq f(y) - |f_0(y) - f(y)| - \eta \geq \lambda - |f_0(y) - f(y)| - \eta$$

then show $x \in S_{(\lambda - \|f_0 - f\|_\infty - \eta)}$. Since $x \in \hat{S}_{\delta, \lambda}(f)$ was arbitrary our inclusion follows.

Next, we show the inclusion $S_{(\lambda + \|f_0 - f\|_\infty + \eta)} \subseteq \hat{S}_{\delta, \lambda}(f)$. Pick an $x \in S_{(\lambda + \|f_0 - f\|_\infty + \eta)}$ and note by (14) that $\inf_{y \in B(x, \delta/2)} f_0(y) \geq f_0(x) - \eta \geq \lambda + \|f_0 - f\|_\infty$. Thus (13) shows the existence of some $z \in B(x, \delta/2) \cap \mathcal{X}_n$. Further $f(z) \geq f_0(z) - |f_0(z) - f(z)| \geq f_0(x) - \eta - \|f - f_0\|_\infty \geq \lambda$ since $f_0(z) \geq f_0(x) - \eta$ and $x \in S_{(\lambda + \|f_0 - f\|_\infty + \eta)}$. Thus we have shown that $x \in \hat{S}_{\delta, \lambda}(f)$. Since $x \in S_{(\lambda + \|f_0 - f\|_\infty + \eta)}$ was arbitrary our inclusion follows. ■

We now discuss consequences of Lemma 10 for level set clustering of data $\mathcal{X}_n$. As discussed in Section 2.4, we use the surrogate clustering $\tilde{\psi}_{\delta, \lambda}(f)$ of data $\mathcal{X}_n$ defined in (2), which computes the graph-theoretic connected components (Dasgupta et al., 2008) of the δ -neighborhood graph $G_\delta(A_{f, \lambda})$ having vertices $A_{f, \lambda} = \{x \in \mathcal{X}_n \mid f(x) \geq \lambda\}$ and edges $E = \{(x, y) \in A_{f, \lambda} \times A_{f, \lambda} \mid \|x - y\| < \delta\}$. The following known result (e.g. Lemma 1 in Wang et al. (2019)) connects the surrogate clustering $\tilde{\psi}_{\delta, \lambda}(f)$ to the level-set estimator $\hat{S}_{\delta, \lambda}(f)$ defined in the last section. We provide an independent proof here for completeness.

Lemma 13. *The surrogate clustering $\tilde{\psi}_{\delta, \lambda}(f) \in \mathcal{P}(\mathcal{X}_n)$ coincides with the partition of $A_{f, \lambda} = \{x \in \mathcal{X}_n \mid f(x) \geq \lambda\}$ induced by the topological connected components of the level set estimator $\hat{S}_{\delta, \lambda}(f)$.*

Proof For any two distinct choice $x, y \in A_{f, \lambda}$ we will show that x and y lie in the same connected component of the graph $G_\delta(A_{f, \lambda})$ if and only if they are path connected in the set $\hat{S}_{\delta, \lambda}(f)$.

Indeed, suppose that x, y are in the same connected component of $G_\delta(A_{f, \lambda})$. Then for some $2 \leq m \leq n$ there are points $\{x_i\}_{i=1}^m \subseteq A_{f, \lambda}$ with $x_1 = x$, $x_m = y$ and $\|x_i - x_{i+1}\| < \delta$ for $i = 1, \dots, m-1$. These conditions ensure that the interval $[x_i, x_{i+1}] \doteq \{tx_i + (1-t)x_{i+1} : t \in [0, 1]\}$ is entirely contained within $\hat{S}_{\delta, \lambda}(f)$. Thus there is a continuous path from x to y that entirely lies within $\hat{S}_{\delta, \lambda}(f)$, which ensures that x, y are path connected in $\hat{S}_{\delta, \lambda}(f)$.

Conversely, suppose that $x, y \in A_{f, \lambda}$ are path connected in $\hat{S}_{\delta, \lambda}(f)$. Thus there is a continuous path $\varphi : [0, 1] \rightarrow \hat{S}_{\delta, \lambda}(f)$ such that $\varphi(0) = x$ and $\varphi(1) = y$. Based on φ , we can define two mappings $T : A_{f, \lambda} \rightarrow [0, 1]$ and $F : [0, 1] \rightarrow A_{f, \lambda}$ given by

$$T(z) = \sup\{t \in [0, 1] : \varphi(t) \in B(z, \delta/2)\} \quad \text{and} \quad F(t) \in \arg \min_{z \in A_{f, \lambda}} \|z - \varphi(t)\|.$$

We must have $\varphi(t) \in B(F(t), \delta/2)$ for each $t \in [0, 1]$ since the image of the path φ lies entirely in $\hat{S}_{\delta, \lambda}(f)$. Further, for each $z \in A_{f, \lambda}$ such that $T(z) \in [0, 1]$, it must be the case that $\|\varphi(T(z)) - z\| = \delta/2$ due to the continuity of φ .

Starting with $t_0 = 0$ and $x_0 = F(t_0) = x$, recursively define $t_i = T(x_{i-1}) \in [0, 1]$ and $x_i = F(t_i) \in A_{f, \lambda}$ for each $i \geq 1$. By the definition of T , we note that $t_i = T(x_{i-1}) \geq t_{i-1}$ since $\varphi(t_{i-1}) \in B(x_{i-1}, \delta/2)$ holds given that $x_{i-1} = F(t_{i-1})$ for each $i \geq 1$. In fact, $\|x_i - x_{i-1}\| < \delta$ since $\|\varphi(t_i) - x_{i-1}\| \leq \frac{\delta}{2}$ follows by using the continuity of φ and $t_i = T(x_{i-1})$, while $\|\varphi(t_i) - x_i\| < \delta/2$ follows since $x_i = F(t_i)$. Thus we can show that $x_0, x_1, \dots$, is an infinite path in $G_\delta(A_{f, \lambda})$ starting from $x_0 = x \in A_{f, \lambda}$.

Next, we claim that the path $x_0, x_1, \dots, x_m$ in $G_\delta(A_{f, \lambda})$ will terminate at $x_m = F(t_m) = y$, where m is smallest integer such that $t_m = 1$. Thus the proof will be complete once we

show that such an $m \in \mathbb{N}$ will exist. Whenever $t_{i-1} < 1$, we can observe that $t_{i-1} \neq t_i$ since $\varphi(t_i) \notin B(x_{i-1}, \delta/2)$ but $\varphi(t_{i-1}) \in B(x_{i-1}, \delta/2)$. Further as long as $t_{i-1} < 1$, we must also have $x_i \notin \{x_0, \dots, x_{i-1}\}$ because $\varphi(t_i) \in B(x_i, \delta/2)$ but $\varphi(t_i) \notin \cup_{j=0}^{i-1} B(x_j, \delta/2)$ since $t_i > \max(t_0, \dots, t_{i-1})$. Hence we have shown that for each $i \geq 1$, the points $x_0, \dots, x_i \in A_{f,\lambda}$ will be distinct as long as $t_{i-1} < 1$. Since $A_{f,\lambda}$ is a finite set, there must be $m \in \mathbb{N}$ such that $t_m = 1$ and $x_m = F(t_m) = \varphi(t_m) = y$. Thus x, y are connected by a path in $G_\delta(A_{f,\lambda})$. $\blacksquare$

When Lemma 10 holds and Assumption 6 is satisfied, the topological connected components of $\hat{S}_{\delta,\lambda}(f)$ will be close to those of the level set S_λ if $\|f - f_0\|_\infty$ and δ are suitably small. To formally define this relationship we start with the following definition.

Definition 14. Consider the binary co-clustering relations $T, \hat{T}_{\delta,f} : \mathcal{X} \times \mathcal{X} \rightarrow \{0, 1\}$ defined as follows. For any $x, y \in \mathcal{X}$, we define $T(x, y) = 1$ if x and y either both fall outside the level set S_λ or if they lie in the same topological connected component of S_λ , otherwise we let $T(x, y) = 0$. The estimated quantity $\hat{T}_{\delta,f}(x, y)$ is defined similarly as above, but with S_λ replaced by $\hat{S}_{\delta,\lambda}(f)$.

Lemma 15. Suppose that Assumption 6 is satisfied and the conclusion of Lemma 10 holds with $\epsilon \doteq \|f - f_0\|_\infty + \eta \leq \bar{\epsilon}$. Then whenever $T(x, y) \neq \hat{T}_{\delta,f}(x, y)$ for some $x, y \in \mathcal{X}$, it must follow that $\{x, y\} \cap S_{(\lambda-\epsilon)} \setminus S_{(\lambda+\epsilon)} \neq \emptyset$.

Proof Fix any pair $x, y \in \mathcal{X}$. It suffices to show that $T(x, y) = \hat{T}_{\delta,f}(x, y)$ whenever $\{x, y\} \cap S_{(\lambda-\epsilon)} \setminus S_{(\lambda+\epsilon)} = \emptyset$. We will consider the following cases:

1. Case $x, y \in S_{(\lambda+\epsilon)}$. Assumption 6 states that the topological connectivity between x, y as points in $S_{(\lambda')}$ remains unchanged as long as $\lambda' \in [\lambda - \bar{\epsilon}, \lambda + \bar{\epsilon}]$. Further Lemma 10 shows that

$$S_{(\lambda+\epsilon)} \subseteq \hat{S}_{\delta,\lambda}(f) \subseteq S_{(\lambda-\epsilon)}. \quad (15)$$

Thus if $T(x, y) = 1$, points x, y will be connected in $S_{(\lambda+\epsilon)}$ and hence also in $\hat{S}_{\delta,\lambda}(f)$, and thus we must have $\hat{T}_{\delta,f}(x, y) = 1$. Conversely, if $T(x, y) = 0$, then x, y are disconnected in $S_{(\lambda-\epsilon)}$ and hence also in $\hat{S}_{\delta,\lambda}(f)$, giving $\hat{T}_{\delta,f}(x, y) = 0$.

2. Case $x, y \notin S_{(\lambda-\epsilon)}$. Then $T(x, y) = 1$ since $x, y \notin S_\lambda$. But by eq. (15), $x, y \notin \hat{S}_{\delta,\lambda}(f)$ and thus $\hat{T}_{\delta,f}(x, y) = 1$.
3. Case $x \in S_{(\lambda+\epsilon)}$ and $y \notin S_{(\lambda-\epsilon)}$ (or vice-versa). Then $T(x, y) = 0$ since $x \in S_\lambda$ but $y \notin S_\lambda$. Equation (15) shows that $x \in \hat{S}_{\delta,\lambda}(f)$ and $y \notin \hat{S}_{\delta,\lambda}(f)$, and thus $\hat{T}_{\delta,f}(x, y) = 0$.

In any case, we have shown that $T(x, y) = \hat{T}_{\delta,f}(x, y)$ if the condition $\{x, y\} \cap S_{(\lambda-\epsilon)} \setminus S_{(\lambda+\epsilon)} \neq \emptyset$ does not hold. $\blacksquare$

If Assumption 5 holds in addition to the result in Lemma 15, then one immediately notes that for samples X, Y drawn independently at random from f_0 we have

$$\begin{aligned} P_{f_0}\{T(X, Y) \neq \hat{T}_{\delta, f}(X, Y)\} &\leq P_{f_0}[\{X, Y\} \cap S_{(\lambda-\epsilon)} \setminus S_{(\lambda+\epsilon)} \neq \emptyset] \\ &\leq 2P_{f_0}\{X \in S_{(\lambda-\epsilon)} \setminus S_{(\lambda+\epsilon)}\} = 2 \int_{\{x: |f_0(x)-\lambda| \leq \epsilon\}} f_0(x) dx \leq 2C\epsilon. \end{aligned}$$

where P_{f_0} denotes the probability under independent draws X, Y from f_0 . This suggests that if $\|f - f_0\|_\infty$ and $\delta > 0$ are suitably small, so that ϵ can be chosen to be small, then for any fixed pairs of indices $1 \leq i < j \leq n$, the data points x_i, x_j will, with probability at least $1 - C\epsilon$, be identically co-clustered by the surrogate function $\tilde{\psi}_{\delta, \lambda}$ and the level-set function ψ_λ , that is, points x_i, x_j will either be in the same cluster in both $\tilde{\psi}_{\delta, \lambda}$ and ψ_λ , or they will be in different clusters of both $\tilde{\psi}_{\delta, \lambda}$ and ψ_λ . The following theorem builds on this intuition to bound $D\{\tilde{\psi}_{\delta, \lambda}(f), \psi_\lambda(f_0)\}$ where $D = \binom{n}{2}^{-1} L_{\text{IA-Binder}}$ is the loss from Theorem 2.

Theorem 16. *Let f_0 and $\lambda > 0$ satisfy Assumptions 4 to 6, and let $\mathcal{X}_n = \{x_1, \dots, x_n\}$ be independent draws from f_0 . Then, whenever $n \geq 16$, with probability at least $1 - \frac{n+1}{n^2}$*

$$\sup_{\delta \in [r_{n, \lambda, d}, 2H_{f_0}(\epsilon)]} \sup_{f: \|f - f_0\|_\infty \leq \epsilon} D\{\tilde{\psi}_{\delta, \lambda}(f), \psi_\lambda(f_0)\} \leq 8 \left(C\epsilon + \sqrt{\frac{\ln n}{n}} \right) \quad \text{for every } \epsilon \in (0, \bar{\epsilon}/2), \quad (16)$$

where $\tilde{\psi}_{\delta, \lambda}$ is the surrogate clustering defined in eq. (2), ψ_λ is the true level set clustering defined in Section 2.3, $D = \binom{n}{2}^{-1} L_{\text{IA-Binder}}$ is the loss from Theorem 2, $\eta \mapsto H_{f_0}(\eta)$ is defined in (12), $r_{n, \lambda, d} \doteq 2 \left(\frac{16d \ln n}{nv_d \lambda} \right)^{1/d}$, and v_d is the volume of the unit Euclidean ball in d dimensions.

Proof By Lemma 9, the assumptions of Lemma 10 are satisfied. Thus, if we take $\eta = \epsilon \in (0, \bar{\epsilon}/2)$ in Lemma 10, we see that the condition

$$S_{(\lambda+2\epsilon)} \subseteq \hat{S}_{\delta, \lambda}(f) \subseteq S_{(\lambda-2\epsilon)} \quad (17)$$

holds uniformly over all $f : \mathcal{X} \rightarrow \mathbb{R}$ with $\|f - f_0\|_\infty \leq \epsilon$ and $\delta \in [r_{n, \lambda, d}, 2H_{f_0}(\epsilon)]$ with probability at least $1 - 1/n$. Henceforth, let us suppose that this event holds. Recall the true and estimated co-clustering relations T and $\hat{T}_{\delta, f}$ from Definition 14. By Lemma 15, for any f, δ such that $\|f - f_0\|_\infty \leq \epsilon$ and $\delta \in [r_{n, \lambda, d}, 2H_{f_0}(\epsilon)]$, we see that if $T(x, y) \neq \hat{T}_{\delta, f}(x, y)$ for some $x, y \in \mathcal{X}$, then one of x or y must lie in the region $\Delta(\epsilon) \doteq S_{(\lambda-2\epsilon)} \setminus S_{(\lambda+2\epsilon)} \subseteq \mathcal{X}$.

Next we note that only a small fraction of observed data points $\mathcal{X}_n$ lie in the region $\Delta(\epsilon) \subseteq \mathcal{X}$. We use Hoeffding's inequality to establish this, noting that the event

$$\hat{P}\{\Delta(\epsilon)\} - P_{f_0}\{\Delta(\epsilon)\} \leq \sqrt{\frac{\ln n}{n}}$$

holds with probability at least $1 - 1/n^2$, where $\hat{P}(A) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(x_i \in A)}$ denotes the empirical measure of any $A \subseteq \mathcal{X}$, and $P_{f_0}\{\Delta(\epsilon)\} = \int_{\Delta(\epsilon)} f_0(x) dx$ denotes its population measure under

the density f_0 . Under Assumption 5 we have $P_{f_0}\{\Delta(\epsilon)\} = \int_{\{x: |f_0(x) - \lambda| \leq 2\epsilon\}} f_0(x) dx \leq 2C\epsilon$ and thus:

$$\hat{P}\{\Delta(\epsilon)\} \leq 2C\epsilon + \sqrt{\frac{\ln n}{n}}. \quad (18)$$

By the union bound, the events (17) and (18) will simultaneously hold with probability at least $1 - \frac{n+1}{n^2}$. We henceforth assume that these events hold. We are now ready to establish (16). Fix any $\epsilon \in (0, \bar{\epsilon}/2)$, $\delta \in [r_{n,\lambda,d}, 2H_{f_0}(\epsilon)]$, and f with $\|f - f_0\|_\infty \leq \epsilon$, and, for brevity, let $\hat{\mathcal{C}}_f, \mathcal{C}_0 \in \mathcal{P}(\mathcal{X}_n)$ denote $\tilde{\psi}_{\delta,\lambda}(f)$ and $\psi_\lambda(f_0)$ respectively. Starting from the representation (11) in the proof of Theorem 2, we note that:

$$\begin{aligned} D(\hat{\mathcal{C}}_f, \mathcal{C}_0) &= \frac{1}{n(n-1)} \sum_{i \in [n]} \sum_{j \in [n] \setminus \{i\}} \phi_{x_i, x_j}(\hat{\mathcal{C}}_f, \mathcal{C}_0) \\ &= \frac{1}{n(n-1)} \sum_{i \in [n]} \sum_{j \in [n] \setminus \{i\}} \left[m \mathbb{1}_{\{\hat{\mathcal{C}}_{f,A}(x_i) \neq \mathcal{C}_{0,A}(x_i)\}} + m \mathbb{1}_{\{\hat{\mathcal{C}}_{f,A}(x_j) \neq \mathcal{C}_{0,A}(x_j)\}} \right. \\ &\quad \left. + a \mathbb{1}_{\{\hat{\mathcal{C}}_{f,R}(x_i, x_j) \neq \mathcal{C}_{0,R}(x_i, x_j)\}} \mathbb{1}_{\{\hat{\mathcal{C}}_{f,A}(x_i) = \mathcal{C}_{0,A}(x_i) = \hat{\mathcal{C}}_{f,A}(x_j) = \mathcal{C}_{0,A}(x_j)\}} \right] \\ &= \frac{2m}{n} \sum_{i \in [n]} \mathbb{1}_{\{\hat{\mathcal{C}}_{f,A}(x_i) \neq \mathcal{C}_{0,A}(x_i)\}} + \frac{a}{n(n-1)} \sum_{i \in [n]} \sum_{j \in [n] \setminus \{i\}} \mathbb{1}_{\{\hat{\mathcal{C}}_{f,R}(x_i, x_j) \neq \mathcal{C}_{0,R}(x_i, x_j)\}} \mathbb{1}_{(x_i, x_j \in A_{f,\lambda} \cap A_{f_0,\lambda})} \\ &= \frac{2m}{n} \sum_{i \in [n]} \mathbb{1}_{(x_i \in A_{f,\lambda} \Delta A_{f_0,\lambda})} + \frac{a}{n(n-1)} \sum_{i \in [n]} \sum_{j \in [n] \setminus \{i\}} \mathbb{1}_{\{\hat{T}_{\delta,f}(x_i, x_j) \neq T(x_i, x_j)\}} \mathbb{1}_{(x_i, x_j \in A_{f,\lambda} \cap A_{f_0,\lambda})}. \end{aligned}$$

Indeed, for the third equality, we have used that the last summand in the second equation (i.e. the term in the third line) is non-zero only when $x_i, x_j \in A_{f,\lambda} \cap A_{f_0,\lambda}$, where $A_{f,\lambda} = \{x \in \mathcal{X}_n : f(x) \geq \lambda\}$ and $A_{f_0,\lambda} = S_\lambda \cap \mathcal{X}_n$ are the active sets of $\hat{\mathcal{C}}_f$ and $\mathcal{C}_0$, respectively. For the subsequent equality, Δ symbolizes the symmetric difference between sets. Here we note by definition that the co-clustering relation $\mathcal{C}_{0,R}$ is the relation T restricted to $\mathcal{X}_n$. Further, restricting to the points in $A_{f,\lambda}$, Lemma 13 shows that the co-clustering relation $\hat{\mathcal{C}}_{f,R}$ defined via $\tilde{\psi}_{\delta,\lambda}(f)$ is equal to the co-clustering relation $\hat{T}_{\delta,f}$ defined via the connected components of $\hat{S}_{\delta,\lambda}(f)$, i.e. $\hat{\mathcal{C}}_{f,R}(x, y) = \hat{T}_{\delta,f}(x, y)$ for any $x, y \in A_{f,\lambda}$.

In order to complete the proof, we note the inequality $\mathbb{1}_{\{T(x,y) \neq \hat{T}_{\delta,f}(x,y)\}} \leq \mathbb{1}_{\{x \in \Delta(\epsilon)\}} + \mathbb{1}_{\{y \in \Delta(\epsilon)\}}$ and inclusion $A_{f,\lambda} \Delta A_{f_0,\lambda} \subseteq \Delta(\epsilon) \cap \mathcal{X}_n$. While the inequality follows from the argument noted at the beginning of this proof, the inclusion follows since $\mathbb{1}_{\{f_0(x) \geq \lambda\}} = \mathbb{1}_{\{f(x) \geq \lambda\}}$ whenever $x \in \mathcal{X} \setminus \Delta(\epsilon)$ and $\|f - f_0\|_\infty \leq 2\epsilon$. We thus obtain the bound:

$$D(\hat{\mathcal{C}}_f, \mathcal{C}_0) \leq 2(m+a)\hat{P}\{\Delta(\epsilon)\} \leq 8\left(C\epsilon + \sqrt{\frac{\ln n}{n}}\right).$$

Since $\epsilon \in (0, \bar{\epsilon}/2)$, $\delta \in [r_{n,\lambda,d}, H_{f_0}(\epsilon)]$, and f with $\|f - f_0\|_\infty \leq \epsilon$ were arbitrary, we have shown that (16) holds. $\blacksquare$

The proof of Theorem 4 now follows as a special case of the above theorem. Indeed, suppose f_0 is an α -Hölder continuous function so that $|f_0(x) - f_0(y)| \leq C_\alpha |x - y|^\alpha$ for

some constant $C_\alpha > 0$. Then from (12) we find that $H_{f_0}(\eta) \geq (\eta/C_\alpha)^{1/\alpha}$ for any $\eta > 0$. Thus we can take $\epsilon = \max(\gamma, C_\alpha(2\delta)^\alpha)$ in Theorem 16 to obtain Theorem 4 with $\bar{\gamma} = \bar{\epsilon}/2$, $\bar{\delta} = \frac{1}{2}(\bar{\epsilon}/2C_\alpha)^{1/\alpha}$ and $C_0 = 8(1+C)(1+C_\alpha)2^\alpha$.

I.4 Proof of Lemma 5

Here we show that our data-adaptive choice of $\delta = \hat{\delta}$ from (3) based on the k -nearest neighbor distance $\delta_k(x) \doteq \inf\{r > 0 : |B(x, r) \cap \mathcal{X}_n| \geq k\}$ will satisfy conditions of Theorem 4.

The argument of our proof starts with the following corollary of Lemma 11 used in Chaudhuri and Dasgupta (2010) and later works like Dasgupta and Kpotufe (2014); Jiang (2017) to study properties of k -nearest neighbor density estimates.

Lemma 17 (Lemma 2 in Dasgupta and Kpotufe (2014)). *Suppose P is a probability measure on $\mathbb{R}^d$ and $\hat{P}(A) = n^{-1} \sum_{i=1}^n \mathbf{I}\{X_i \in A\}$ is the empirical distribution based on n i.i.d. samples $X_1, \dots, X_n$ from P . Pick $0 < t < 1$ and let $C_{t,n} \doteq 16 \log(2/t) \sqrt{d \log n}$. If $k \geq d \log n$ then with probability at least $1 - t$, for every ball $B \subseteq \mathbb{R}^d$ we have:*

$$\begin{aligned} P(B) \geq C_{t,n} \frac{\sqrt{d \log n}}{n} &\implies \hat{P}(B) > 0 \\ P(B) \geq k/n + C_{t,n} \frac{\sqrt{k}}{n} &\implies \hat{P}(B) \geq \frac{k}{n} \text{ and} \\ P(B) \leq k/n - C_{t,n} \frac{\sqrt{k}}{n} &\implies \hat{P}(B) < \frac{k}{n}. \end{aligned}$$

This leads to the following corollary for the behavior of our k nearest neighbor distance based on data $\mathcal{X}_n = \{x_1, \dots, x_n\}$ drawn independently from the assumed distribution $P_0(A) \doteq \int_A f_0(x) dx$.

Corollary 18. *Suppose $k \geq (32)^2 d \log n$. Then with probability at least $1 - 2e^{-\frac{1}{32} \sqrt{\frac{k}{d \log n}}}$ uniformly over $x \in \mathbb{R}^d$ and $r > 0$ we have:*

$$\begin{aligned} \delta_k(x) \leq r &\quad \text{if } P_0(B(x, r)) \geq \frac{3k}{2n} \quad \text{and} \\ \delta_k(x) \geq r &\quad \text{if } P_0(B(x, r)) \leq \frac{k}{2n}. \end{aligned}$$

Proof We will take $t = 2e^{-\frac{1}{32} \sqrt{\frac{k}{d \log n}}}$ in Lemma 17 noting that $C_{t,n} = \frac{\sqrt{k}}{2}$. Thus Lemma 17 shows that with probability $1 - 2e^{-\frac{1}{32} \sqrt{\frac{k}{d \log n}}}$:

$$\begin{aligned} P_0(B) \geq \frac{3k}{2n} &\implies \hat{P}(B) \geq \frac{k}{n} \quad \text{and} \\ P_0(B) \leq \frac{k}{2n} &\implies \hat{P}(B) < \frac{k}{n}. \end{aligned}$$

for each $x \in \mathcal{X}$ and $r > 0$ with $B = B(x, r)$ and $\hat{P}(B) = \frac{|B \cap \mathcal{X}_n|}{n}$. The proof is completed by noting that $\delta_k(x) = \inf\{r | \hat{P}(B(x, r)) \geq k/n\}$. Hence when $\hat{P}(B(x, r)) \geq \frac{k}{n}$ we must have $\delta_k(x) \leq r$ and when $\hat{P}(B(x, r)) < \frac{k}{n}$ we must have $\delta_k(x) \geq r$. $\blacksquare$

Now we are ready to prove Lemma 5.

Proof of Lemma 5

By Assumption 4 and Lemma 9, f_0 is uniformly continuous and bounded. Thus there are constants $\bar{r} > 0$ and $M > 0$ such that

$$\sup_{x \in \mathcal{X}} f_0(x) \leq M < \infty$$

and

$$\sup_{\substack{x, y \in \mathcal{X} \\ \|x - y\| \leq \bar{r}}} |f_0(x) - f_0(y)| \leq \lambda/4.$$

We will assume that $k \in [L \log n, n/L]$ for a suitably large constant $L > 0$ that is independent of n , which can be determined by examining the details of this proof. For example, we will assume that L is large enough so that the event in Corollary 18 holds with high probability.

First let us show that $\hat{\delta}$ from (3) will be less than $\bar{\delta}$. This will follow if for any $x_i \in A_{\lambda, \hat{f}}$ we can show that $\delta_k(x_i) \leq r_0 = \min(\bar{r}, \bar{\delta}/2)$. Indeed, since $\|\hat{f} - f_0\|_\infty \leq \lambda/2$, we must have $f_0(x_i) \geq \hat{f}(x_i) - \|f_0 - \hat{f}\|_\infty \geq \lambda - \lambda/2 = \lambda/2$. Further, since $r_0 \leq \bar{r}$, we must have $\inf_{y \in B(x_i, r_0)} f_0(y) \geq \lambda/4$. This shows that

$$P_0(B(x_i, r_0)) \geq \frac{\lambda}{4} v_d(r_0)^d \geq \frac{3}{2L} \geq \frac{3k}{2n}$$

as long as $L \geq \frac{6}{\lambda v_d(r_0)^d}$. By Corollary 18, we must have $\delta_k(x_i) \leq r_0$ as required.

Next, let us show that $\hat{\delta} \geq r_{n, \lambda, d}$. Since $\hat{\delta} \geq \inf_{x_i \in A_{\lambda, \hat{f}}} \delta_k(x_i)$ we will in-fact show that $\delta_k(x) \geq r_{n, \lambda, d}$ for any $x \in \mathcal{X}$. Indeed, this will follow from Corollary 18 once we can show that $P_0(B(x, r_{n, \lambda, d})) \leq \frac{k}{2n}$. From the definition of $r_{n, \lambda, d} = 2 \left(\frac{16d \ln n}{nv_d \lambda} \right)^{1/d}$ and the maximum value M for f_0 , we can note that

$$P_0(B(x, r_{n, \lambda, d})) \leq M v_d(r_{n, \lambda, d})^d = 2^{d+4} \frac{Md \ln n}{\lambda n} \leq \frac{L \log n}{2n} \leq \frac{k}{2n}$$

as long as $L \geq \frac{2^{d+5} Md}{\lambda}$. $\blacksquare$

Appendix J. Selecting the level λ

The level set threshold $\lambda > 0$ is an important parameter for our analysis, and its choice needs to align well with the nature of clustering that we seek. In order to improve interpretation and comparison of level set clusters across different density models and clustering methods,

following Cuevas et al. (2001); Scrucca (2016), we choose the fraction of noise points $\nu \in (0, 1)$ rather than the actual density level $\lambda > 0$. Indeed, there is a one-to-one association between the two parameters when our true data-generating distribution has a continuous density. Our experiments here demonstrate at least the following three possible ways to practically choose the level λ , depending on the goals of our clustering analysis.

1. *A known value of the level λ .* In our sky-survey analysis (Section 6), the clustering of interest corresponded to an approximately known value of λ motivated by scientific considerations. While our analysis in Section 6 directly used this threshold λ , we note in Section K that exploring the persistence of clusters across nearby choices of λ may improve clustering accuracy. Indeed, even if the target level λ is known exactly, the need for checking persistence of clusters across nearby levels has also appeared in theoretical studies of level set clustering (Steinwart, 2011; Sriperumbudur and Steinwart, 2012; Jiang, 2017).
2. *Finding the level λ to separate a noisy background.* Often, our clusters of interest will be connected components of regions with significantly large data density values, separated by noisy regions of comparatively much lower density values. For example, this is the case for our toy data example from Figure 1 and our illustrative data examples in Section 5 if we are interested in the connected components of the obvious regions of non-negligible data density. (Note: depending on the density model used for the RNA-seq example, there is perhaps still some ambiguity about whether some observations bordering the major regions should be called noisy or not.) For these datasets, motivated by DBSCAN (Ester et al., 1996; Schubert et al., 2017), we have found the following elbow heuristic useful: we sort the values of the logarithm of the density estimates $\{\log \hat{f}(x_i)\}_{i=1}^n$ at the observations, and use the ‘kneedle’ algorithm (Satopaa et al., 2011) to find a so-called elbow (or knee) in the plot of the logarithm of the density estimates versus their ranks (see Figures 27 and 28). The intuition here is that a noisy observation x_i will have a much smaller value of $\log \hat{f}(x_i)$ compared to a non-noisy observation x_i , and since the fraction of noisy observations is assumed to be small, this will reflect as an elbow in our plot. Figure 26 shows the **BALLET** clusters for the illustrative challenge datasets, based on the level selected using this elbow heuristic.
3. *Finding nuanced clusters by varying the density λ .* A careful choice of the level λ can reveal more nuanced clustering structure in the data, whereby what seemed like a single cluster at a lower value of λ can split into more than one cluster when a higher value of λ is used. Indeed, this has motivated estimation of an entire hierarchical clustering tree as λ varies (see Wang et al., 2019; Campello et al., 2019; Steinwart et al., 2023, and references therein), but additional strategies are then needed to obtain a flat clustering from the hierarchical clustering tree (Campello et al., 2013; Scrucca, 2016). Here, particularly for the RNA-seq dataset, we visualize the **BALLET** clusters for a range of values of $\nu \in \{5\%, 10\%, 15\%\}$ (see Figures 3, 11 and 12). Some of the clusters when $\nu = 5\%$ are seen to split further when we choose $\nu = 10\%$. In Figure 15, we show the persistent clusters (Section K) for this dataset obtained by post-processing the results corresponding to the noise levels $\nu \in \{5\%, 10\%, 15\%\}$.

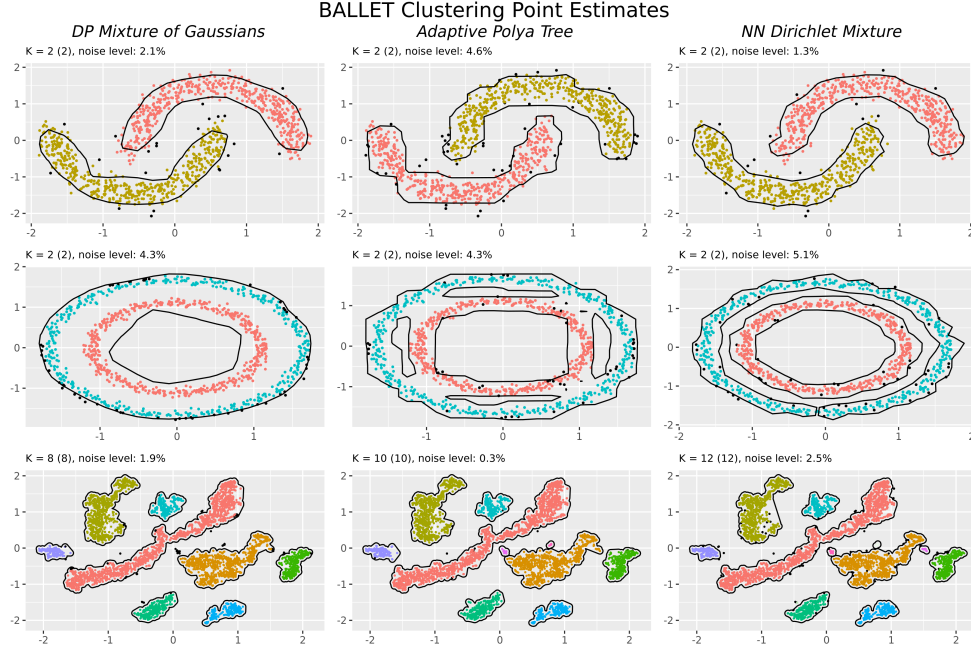


Figure 26: BALLET clustering point estimates obtained under the three different density models shown in Figure 10 with the level chosen using the elbow heuristic (see Figure 27).

We note that a related notion of post-processing of the output of level set clustering methods across different levels has been explored in Steinwart (2011); Sriperumbudur and Steinwart (2012); Steinwart (2015) to consistently estimate the smallest level where the true density has more than one connected component, but their aim is different from what we need here.

Appendix K. Persistent Clustering

K.1 Motivation: robustness to the choice of level λ

A key problem with level set clustering is that we may not exactly know the level (Campello et al., 2019) or, worse yet, that our results can be sensitive to the exact level that we choose for our analysis. Here we describe how to summarize clustering results across multiple values of the level by visualizing a cluster tree (Zappia and Oshlack, 2018), and reduce our sensitivity to any single choice of the level by identifying clusters that remain active or “persistent” across all the levels in the tree.

As described in Section 7, we expect the level set clusters of our Edinburgh-Durham Southern Galaxy Catalogue data to be sensitive to the exact value of the level $\lambda = (1+c)\bar{f}$, determined by the scientific constant c . Since c is believed to be around one (Jang, 2006), our preliminary analysis of this data in Section 6 proceeded with the assumption that $\lambda = 2\bar{f}$, or equivalently that $c = 1$. Here we summarize our results from computing the BALLET clusters at various density levels corresponding to the values $c \in \{0.8, 0.9, \dots, 1.2\}$.

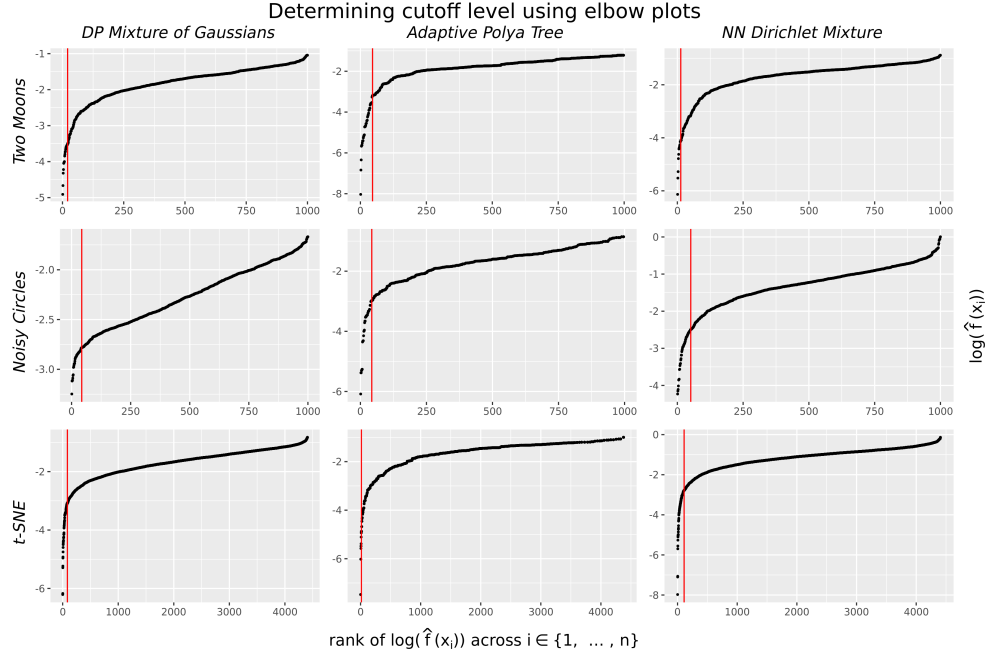


Figure 27: Elbow heuristic to choose the level for the illustrative challenge datasets across density models (Figure 9) based on sorting the log of posterior median density $\hat{f}$ at observations $\{x_i\}_{i=1}^n$ for each dataset and model pair. The elbow value (red vertical line) was automatically determined using the ‘kneedle’ algorithm of Satopaa et al. (2011). Figure 26 shows the corresponding **BALLET** clusters.

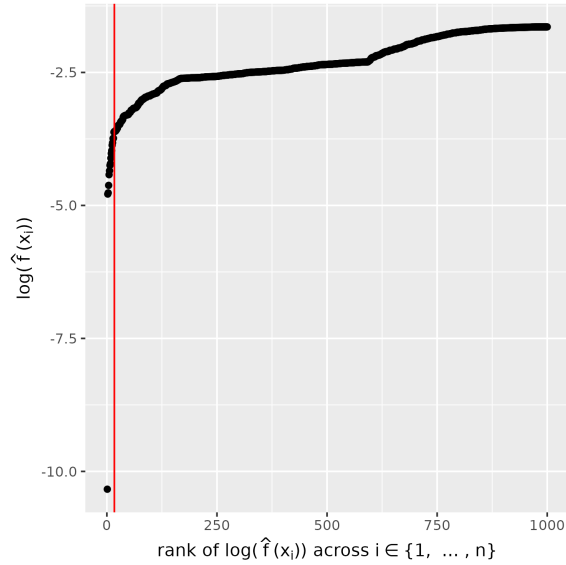


Figure 28: Elbow plot illustrating our selection of the level in Figure 1 based on sorting the log of posterior median density $\hat{f}$ evaluated at the observations $\{x_i\}_{i=1}^n$. The elbow value (red vertical line) was automatically determined using the ‘kneedle’ algorithm of Satopaa et al. (2011).

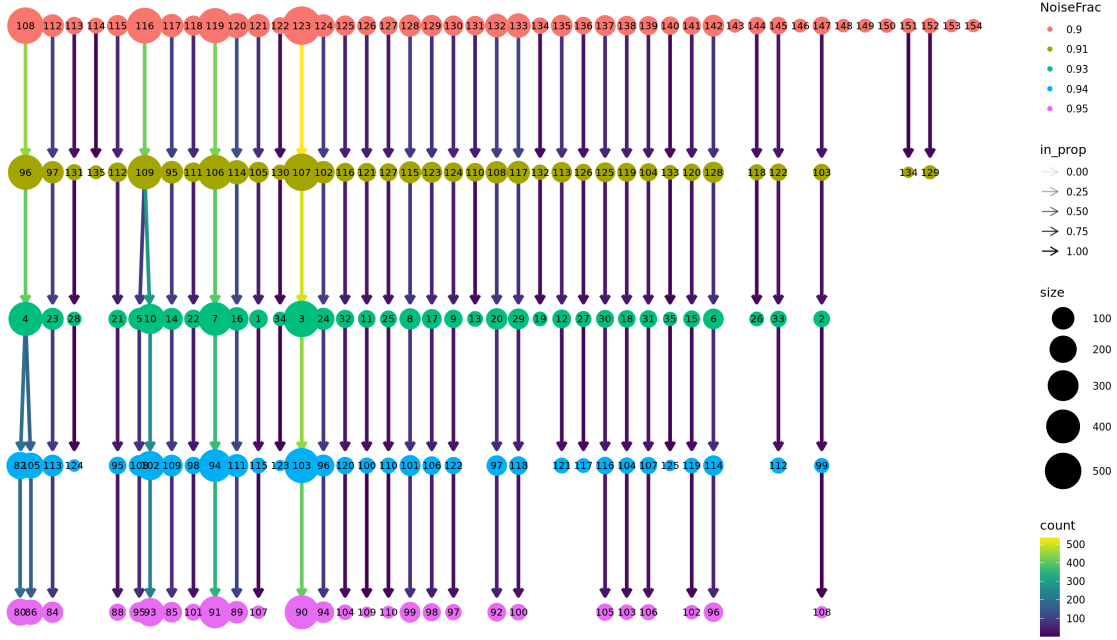


Figure 29: The BALLET cluster tree (Zappia and Oshlack, 2018) for the Edinburgh-Durham Southern Galaxy Catalogue data across multiple density levels corresponding to $c \in \{0.8, 0.9, \dots, 1.2\}$. The nodes in each row are the BALLET clusters for the fixed level $\lambda = (1 + c)\bar{f}$, where c increases as we go down the tree. An edge between nodes in two successive levels indicates an overlap between the two corresponding clusters. While most clusters at the top level ($c = 0.8$) have a unique child in the tree at each lower level (as c increases), some clusters at the top level split into multiple children or did not have any descendant in the bottom levels. For each cluster at the bottom level, the *persistent clustering* algorithm finds its topmost ascendant in the tree below any (potential) split.

K.2 Visualizing the cluster tree

It is well known (Hartigan, 1975; Campello et al., 2019; Menardi, 2015) that the level set clusters across different levels of the same density are nested in a way that can be organized into a tree. In particular, given two clusters from two different levels of the same density, it is the case that either both the clusters are disjoint, or one of the clusters is contained inside the other.

We empirically found that our BALLET estimates across various levels could similarly be organized into a tree. We visualized this tree in Figure 29 by modifying code for the `clustree` package in R (Zappia and Oshlack, 2018). We see that while BALLET found 44 clusters at the lower level ($c = .8$), it only found 27 clusters at the higher level ($c = 1.2$), indicating that more than a third of the lower level clusters disappear as the choice of the level is slightly increased. Further, in this process, two of the lower level clusters are also seen to split into two clusters each.

	BALLET (persistent)	BALLET ($c = 1$)
Sensitivity (EDCCI)	0.69	0.67
Specificity (EDCCI)	0.74	0.69
Exact Match (EDCCI)	0.48	0.51
Sensitivity (Abell)	0.40	0.40
Specificity (Abell)	0.44	0.40
Exact Match (Abell)	0.26	0.26

Table 4: Comparing results from **BALLET** persistent clusters across $c \in \{.8, \dots, 1.2\}$ to the **BALLET** point estimate at $c = 1$. Persistent clustering improves the specificity for both the catalogs without losing sensitivity.

K.3 Persistent Clustering

Given the sensitivity of level set clusters to the choice of level, we now describe a simple algorithm that processes the cluster tree to extract clusters that are active (persistent) across all the levels in the tree. Some clusters can split into multiple sub-clusters as we increase our level in the cluster tree (i.e. go down the tree). In such cases we will only focus on the cluster’s descendants at the time of the last split.

Suppose a cluster tree like Figure 29 is given. Starting from each cluster at the bottom row of the tree, the *Persistent Clustering* algorithm involves walking up the tree until we (i) either hit the top row of the tree, or (ii) hit a node whose parent has more than one child. The collection of clusters corresponding to the final nodes obtained from these runs will be called *persistent clusters*.

BALLET persistent clusters for the Edinburgh-Durham Southern Galaxy Catalogue data are shown in Figure 30. Table 4 compares the performance of **BALLET** persistent clusters to those at the fixed level ($c = 1$). We find that persistent clustering improves specificity on both the Abell and EDCCI catalogs without loss in sensitivity.

While we have motivated the idea of persistent clustering by the practical concern of robustness, the idea of obtaining a single clustering by cutting the cluster tree at locally adaptive levels has been explored before in the algorithmic level set clustering literature (Campello et al., 2019, 2015). Such methods are useful when we want to recover density-based clusters that can only be separated by considering differing values of the levels (Figure 31).

Appendix L. Other clustering methods

There are a wide variety of clustering algorithms (e.g. Wani (2024); Xu and Tian (2015)) because no single notion of clustering is useful across all applications (von Luxburg et al., 2011; Hennig, 2015). Here our focus has been on Bayesian statistical approaches to clustering (Wade, 2023) that account for sampling variability within the data and have the ability to use application-dependent prior information. In principle, our density-based clustering framework allows for the combination of statistical inference with any flexible clustering notion required by the application (provided the clustering $\psi(f)$ can be computed given the population density f).

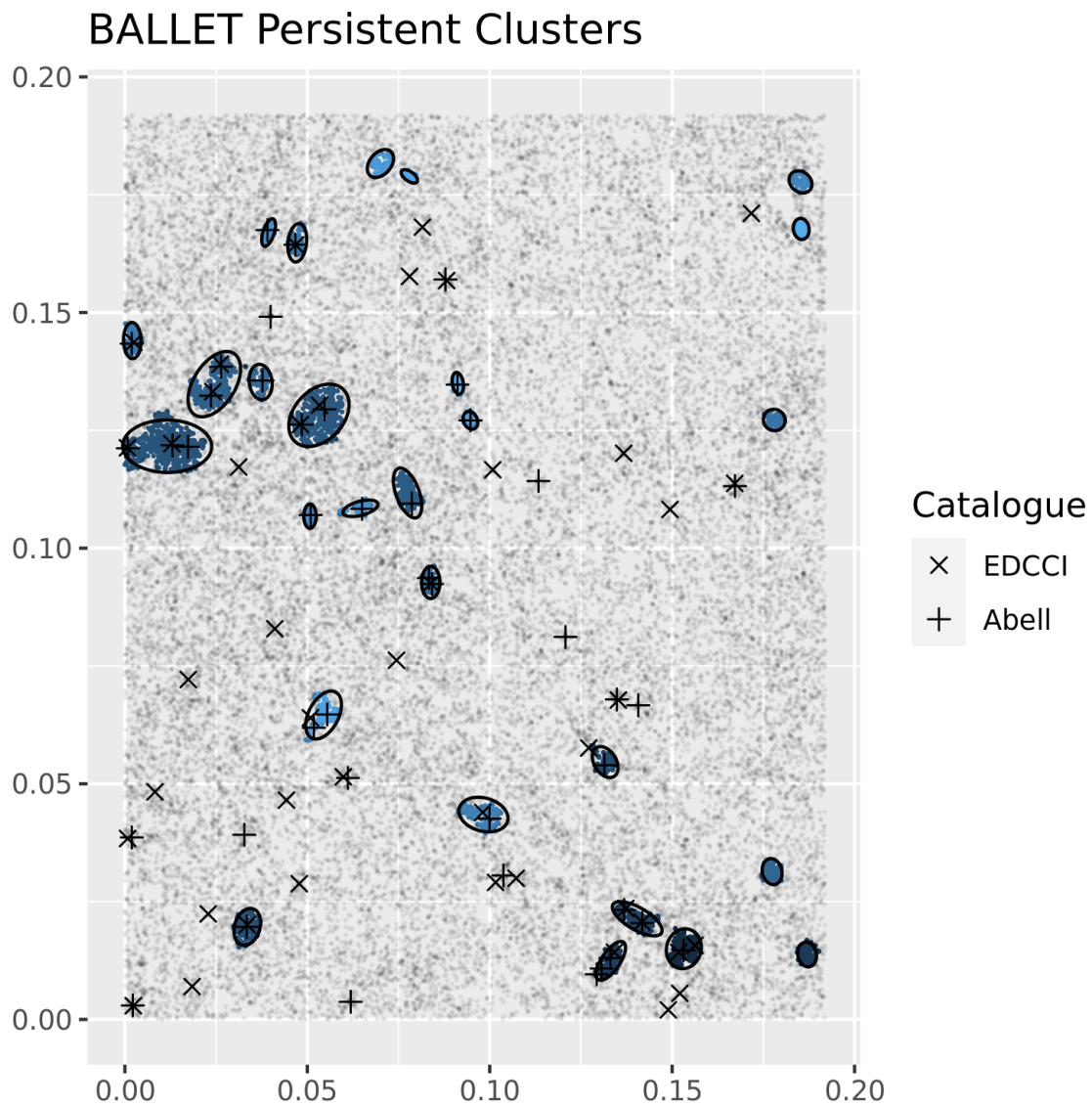


Figure 30: The BALLET persistent clustering estimate for the Edinburgh-Durham Southern Galaxy Catalogue data across levels $c \in \{.8, \dots, 1.2\}$.

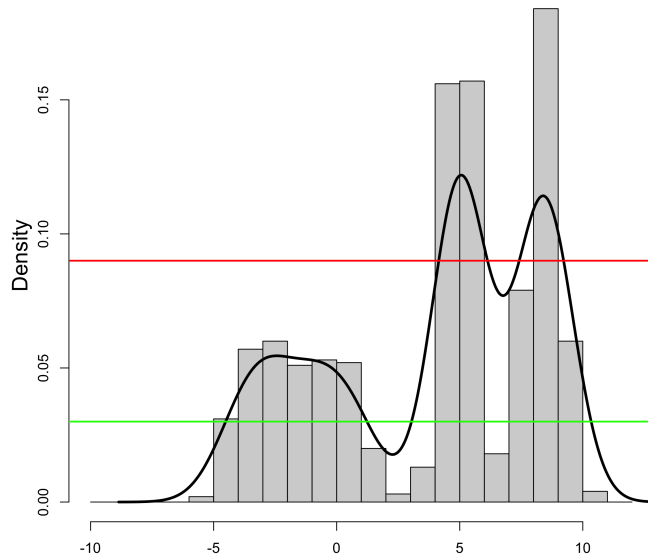


Figure 31: An example of a situation in which we might want to cluster data according to locally adaptive levels.

As an example of our framework, our **BALLET** methodology shows the ability to find arbitrarily shaped clusters in comparison to Gaussian mixture models, which have been predominantly used for Bayesian clustering (Wade, 2023). While additional algorithmic approaches like spectral and hierarchical clustering (Wani, 2024) also have the ability to find arbitrarily shaped clusters, their clustering can be sensitive to the presence of even a few noisy observations. This is seen in Figure 32 with the addition of six new observations to a sample of $n = 600$ observations from one of the datasets considered in Section E.

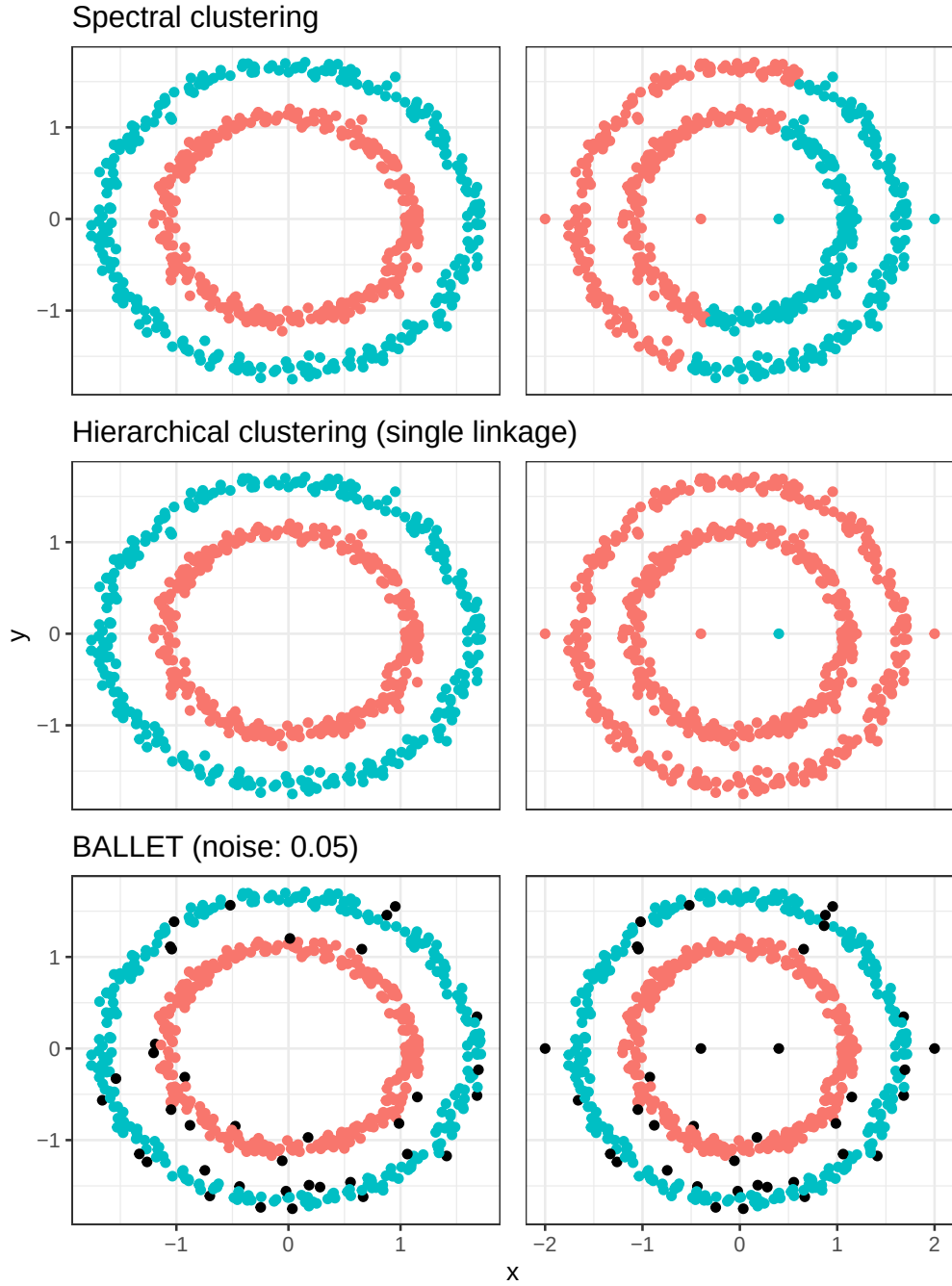


Figure 32: Spectral and hierarchical clustering results change when we add six equally spaced observations on the $y = 0$ line to $n = 600$ observations sampled from one of the datasets in Section E (left: original clustering, right: clustering with six observations added). BALLET clustering based on $\nu = 5\%$ noise points is largely unaffected here as most of these additional points are declared to be noise.

References

- George O Abell, Harold G Corwin Jr, and Ronald P Olowin. A catalog of rich clusters of galaxies. *Astrophysical Journal Supplement Series*, 70:1–138, 1989.
- Homayun Afrabandpey, Tomi Peltola, Juho Piironen, Aki Vehtari, and Samuel Kaski. A decision-theoretic approach for model interpretability in Bayesian framework. *Machine Learning*, 109:1855–1876, 2020.
- Bryon Aragam, Chen Dan, Eric P. Xing, and Pradeep Ravikumar. Identifiability of non-parametric mixture models and Bayes optimal clustering. *The Annals of Statistics*, 48(4):2277–2302, 2020. doi: 10.1214/19-aos1887.
- Ery Arias-Castro and Wanli Qiao. A unifying view of modal clustering. *Information and Inference: A Journal of the IMA*, 12(2):897–920, 2023.
- Amir Bashir, Carlos M. Carvalho, P. Richard Hahn, and M. Beatrix Jones. Post-processing posteriors over precision matrices to produce sparse graph estimates. *Bayesian Analysis*, 14(4):1075–1090, 2019. doi: 10.1214/18-ba1139.
- Mario Beraha, Raffaele Argiento, Jesper Møller, and Alessandra Guglielmi. MCMC computations for Bayesian mixture models using repulsive point processes. *Journal of Computational and Graphical Statistics*, 31(2):422–435, 2022.
- James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.
- Panthadeep Bhattacharjee and Pinaki Mitra. A survey of density based clustering algorithms. *Frontiers of Computer Science*, 15(1):151308, September 2020. doi: 10.1007/s11704-019-9059-3.
- Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*, volume 246. Springer, 2015.
- David A Binder. Bayesian cluster analysis. *Biometrika*, 65(1):31–38, 1978.
- Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.
- Henrique Bolfarine, Carlos M Carvalho, Hedibert F Lopes, and Jared S Murray. Decoupling shrinkage and selection in Gaussian linear factor analysis. *Bayesian Analysis*, 19(1):181–203, 2024.
- Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.
- Diana Cai, Trevor Campbell, and Tamara Broderick. Finite mixture models do not reliably learn the number of components. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 1158–1169, 2021.

- Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. *Data Mining and Knowledge Discovery*, 27:344–371, 2013.
- Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):1–51, 2015.
- Ricardo J. G. B. Campello, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *WIREs Data Mining and Knowledge Discovery*, 10(2):e1343, 2019. doi: 10.1002/widm.1343.
- Ismaël Castillo. On Bayesian supremum norm contraction rates. *The Annals of Statistics*, 42(5):2058–2091, 2014.
- Ismaël Castillo. Pólya tree posterior distributions on densities. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 53(4):2074–2102, 2017.
- José E. Chacón. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, November 2015. doi: 10.1214/15-STS526.
- Noirrit Kiran Chandra, Antonio Canale, and David B Dunson. Escaping the curse of dimensionality in Bayesian model-based clustering. *Journal of Machine Learning Research*, 24(144):1–42, 2023.
- Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, volume 23, 2010.
- Yannis Chaumeny, Johan van der Molen Moris, Anthony C. Davison, and Paul D. W. Kirk. Bayesian nonparametric mixture inconsistency for the number of components: How worried should we be in practice? *arXiv preprint arXiv:2207.14717*, 2022.
- Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112(520):1684–1696, 2017.
- Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Estimating the number of clusters. *The Canadian Journal of Statistics*, 28(2):367–382, 2000.
- Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Cluster analysis: A further approach based on density estimation. *Computational Statistics & Data Analysis*, 36(4):441–459, June 2001. doi: 10.1016/S0167-9473(00)00052-9.
- David B Dahl, Devin J Johnson, and Peter Müller. Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics*, 31(4):1189–1201, 2022.
- Sanjoy Dasgupta and Samory Kpotufe. Optimal rates for k-NN density and mode estimation. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

- Sanjoy Dasgupta, Christos Papadimitriou, and Umesh Vazirani. *Algorithms*. McGraw Hill, 2008.
- Luc Devroye and Gary L Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.
- Alexander Dombowsky and David B. Dunson. Bayesian clustering via fusing of localized densities. *Journal of the American Statistical Association*, 120(551):1775–1786, January 2025. doi: 10.1080/01621459.2024.2427935.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.
- Edward B. Fowlkes and Colin L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.
- Beatrice Franzolini and Giovanni Rebaudo. Entropy regularization in probabilistic clustering. *Statistical Methods & Applications*, 33(1):37–60, 2024.
- Arno Fritsch and Katja Ickstadt. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2):367–391, 2009.
- Sylvia Frühwirth-Schnatter and Saumyadipta Pyne. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11(2):317–336, 2010.
- Ghislaine Gayraud and Judith Rousseau. Rates of convergence for a Bayesian level set estimation. *Scandinavian Journal of Statistics*, 32(4):639–660, 2005.
- Ghislaine Gayraud and Judith Rousseau. Consistency results on nonparametric Bayesian estimation of level sets using spatial priors. *Test*, 16:90–108, 2007.
- Subhashis Ghosal and Aad W van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, 2017.
- Evarist Giné and Richard Nickl. Rates of contraction for posterior distributions in L_r -metrics, $1 \leq r \leq \infty$. *The Annals of Statistics*, 39(6):2883–2911, 2011.
- Aritra Guha, Nhat Ho, and XuanLong Nguyen. On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4):2159–2188, 2021. doi: 10.3150/20-BEJ1275.
- Eduardo Gutiérrez-Peña and Stephen G Walker. Statistical decision problems and Bayesian nonparametric methods. *International Statistical Review*, 73(3):309–330, 2005.
- P Richard Hahn and Carlos M Carvalho. Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015.
- John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

- Christian Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53–62, 2015.
- Nhat Ho and XuanLong Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44(6):2726–2755, 2016.
- Nhat Ho and XuanLong Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *SIAM Journal on Mathematics of Data Science*, 1(4):730–758, 2019.
- Jennifer Jang and Heinrich Jiang. DBSCAN++: Towards fast and scalable density clustering. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3019–3029, 2019.
- Woncheol Jang. Nonparametric density estimation and clustering in astronomical sky surveys. *Computational Statistics & Data Analysis*, 50(3):760–774, 2006.
- Heinrich Jiang. Density level set estimation on manifolds with DBSCAN. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1684–1693, July 2017.
- Daniel R Kowal. Bayesian subset selection and variable importance for interpretable prediction and classification. *Journal of Machine Learning Research*, 23(108):1–38, 2022a.
- Daniel R Kowal. Fast, optimal, and targeted predictions using parameterized decision analysis. *Journal of the American Statistical Association*, 117(540):1875–1886, 2022b.
- Daniel R Kowal and Daniel C Bourgeois. Bayesian function-on-scalars regression for high-dimensional data. *Journal of Computational and Graphical Statistics*, 29(3):629–638, 2020.
- John W Lau and Peter J Green. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558, 2007.
- Michael Lavine. Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics*, 20(3):1222–1235, 1992.
- Peter J Lenk. Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543, 1991.
- Wei Li and Subhashis Ghosal. Posterior contraction and credible regions for level sets. *Electronic Journal of Statistics*, 15(1):2647–2689, January 2021. doi: 10.1214/21-EJS1846.
- Fei Liu, MJ Bayarri, JO Berger, et al. Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150, 2009.
- SL Lumsden, RC Nichol, CA Collins, and L Guzzo. The Edinburgh-Durham southern galaxy catalogue. IV – The cluster catalogue. *Monthly Notices of the Royal Astronomical Society*, 258:1–22, 1992.
- Li Ma. Adaptive shrinkage in Pólya tree type models. *Bayesian Analysis*, 12(3):779–805, 2017.

- Gertraud Malsiner-Walli, Sylvia Frühwirth-Schnatter, and Bettina Grün. Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2):285–295, 2017.
- Mario Medvedovic and Siva Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206, 2002.
- Giovanna Menardi. A review on modal clustering. *International Statistical Review*, 84(3):413–433, 2015. doi: 10.1111/insr.12109.
- Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.
- Zacharie Naulet. Adaptive Bayesian density estimation in sup-norm. *Bernoulli*, 28(2):1284–1308, 2022.
- RC Nichol, CA Collins, L Guzzo, and S Lumsden. The Edinburgh/Durham southern galaxy catalogue. In *Digitised Optical Sky Surveys: Proceedings of the Conference on ‘Digitised Optical Sky Surveys’*, pages 335–344, 1992.
- Francesca Petralia, Vinayak Rao, and David Dunson. Repulsive mixtures. In *Advances in Neural Information Processing Systems*, volume 25, 2012.
- Riccardo Rastelli and Nial Friel. Optimal Bayesian estimators for latent variable cluster models. *Statistics and Computing*, 28:1169–1186, 2018.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: high-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 1527 – 1535, 2018.
- Tommaso Rigon, Amy H Herring, and David B Dunson. A generalized Bayes framework for probabilistic clustering. *Biometrika*, 110(3):559–578, 2023.
- Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, October 2010. doi: 10.1214/10-AOS797.
- Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a “kneedle” in a haystack: Detecting knee points in system behavior. In *31st International Conference on Distributed Computing Systems Workshops*, pages 166–171, 2011. doi: 10.1109/ICDCSW.2011.20.
- Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, 42(3):1–21, 2017.
- Luca Scrucca. Identifying connected components in Gaussian finite mixture models for clustering. *Computational Statistics and Data Analysis*, 93:5–17, 2016. doi: 10.1016/j.csda.2015.01.006.

- Weining Shen and Subhashis Ghosal. Posterior contraction rates of density derivative estimation. *Sankhya A*, 79:336–354, 2017.
- Bharath Sriperumbudur and Ingo Steinwart. Consistency and rates for clustering with DBSCAN. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 1090–1098, March 2012.
- Ingo Steinwart. Adaptive density level set clustering. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 703–738, December 2011.
- Ingo Steinwart. Fully adaptive density-based clustering. *The Annals of Statistics*, 43(5): 2132–2167, October 2015. doi: 10.1214/15-AOS1331.
- Ingo Steinwart, Bharath K Sriperumbudur, and Philipp Thomann. Adaptive clustering using kernel density estimators. *Journal of Machine Learning Research*, 24(275):1–56, 2023.
- Briana JK Stephenson, Amy H Herring, and Andrew Olshan. Robust clustering with subpopulation-specific deviations. *Journal of the American Statistical Association*, 115 (530):521–537, 2020.
- Werner Stuetzle and Rebecca Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2): 397–418, 2010.
- Surya T Tokdar. Towards a faster implementation of density estimation with logistic Gaussian process priors. *Journal of Computational and Graphical Statistics*, 16(3):633–655, 2007.
- Ulrike von Luxburg, Robert C. Williamson, and Isabelle Guyon. Clustering: science or art? In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27, pages 65–79, 2011.
- Sara Wade. Bayesian cluster analysis. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2247):20220149, 2023.
- Sara Wade and Zoubin Ghahramani. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626, 2018. doi: 10.1214/17-ba1073.
- Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.
- Daren Wang, Xinyang Lu, and Alessandro Rinaldo. DBSCAN: Optimal rates for density-based cluster estimation. *Journal of Machine Learning Research*, 20(170):1–50, 2019. URL <https://jmlr.org/papers/v20/18-470.html>.
- Aasim Ayaz Wani. Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. *PeerJ Computer Science*, 10:e2286, 2024.

- Spencer Woody, Carlos M Carvalho, and Jared S Murray. Model interpretation through lower-dimensional posterior summarization. *Journal of Computational and Graphical Statistics*, 30(1):144–161, 2021.
- Fangzheng Xie and Yanxun Xu. Bayesian repulsive Gaussian mixture model. *Journal of the American Statistical Association*, 115(529):187–203, 2020.
- Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2:165–193, 2015.
- Luke Zappia and Alicia Oshlack. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience*, 7(7), 2018.