Bayesian Level Set Clustering

Miheer Dewaskar David Buch David Dunson

Email: mdewaskar@unm.edu
Department of Mathematics and Statistics
University of New Mexico, USA.

19th International Joint Conference on CFE and CMStatistics

Outline

1. Bayesian versus broader clustering literature
2. Bayesian density based clustering
3. Application: Finding clusters of galaxies in the night sky.

Outline

1. Bayesian versus broader clustering literature
2. Bayesian density based clustering
3. Application: Finding clusters of galaxies in the night sky.

Broader clustering literature

Clustering is basically to divide observations into groups. Many approaches:

1. Similarity-based clustering (K-means, PAM, SLINK, Spectral Clustering).
2. Density-based clustering (DBSCAN, Mean-Shift).
3. Model-based clustering (Mixture models)
4. ... Projective clustering, Neural Network based clustering, etc.

Application determines the right clustering approach (Hennig 2015, von Luxburg, Williamson, Guyon, 2011)

Broader clustering literature

Clustering is basically to divide observations into groups. Many approaches:

1. Similarity-based clustering (K-means, PAM, SLINK, Spectral Clustering).
2. Density-based clustering (DBSCAN, Mean-Shift).
3. Model-based clustering (Mixture models)
4. ... Projective clustering, Neural Network based clustering, etc.

Application determines the right clustering approach (Hennig 2015, von Luxburg, Williamson, Guyon, 2011)

What kind of clusters do we wish to find?

- Cluster "nearby" observations $\Longrightarrow$ Similarity-based
- Arbitrary-shaped but well-separated clusters $\Longrightarrow$ Density-based
- Simple model for observations in each group $\Longrightarrow$ Model-based

Broader clustering literature

Clustering is basically to divide observations into groups. Many approaches:

1. Similarity-based clustering (K-means, PAM, SLINK, Spectral Clustering).
2. Density-based clustering (DBSCAN, Mean-Shift).
3. Model-based clustering (Mixture models)
4. . . . Projective clustering, Neural Network based clustering, etc.

Application determines the right clustering approach (Hennig 2015, von Luxburg, Williamson, Guyon, 2011)

What kind of clusters do we wish to find?

- Cluster "nearby" observations $\implies$ Similarity-based
- Arbitrary-shaped but well-separated clusters $\implies$ Density-based
- Simple model for observations in each group $\implies$ Model-based

Typical “Bayesian clustering” is model-based

Starting from a simple (e.g. Gaussian) component kernel $g(y|\theta)$:

$$x_1, \dots, x_n \sim \sum_{k=1}^{K} \pi_k g(\cdot|\theta_k) \quad \text{or} \quad \begin{cases} z_i|\boldsymbol{\pi} \sim \text{Categorical}(\pi_1, \dots, \pi_K) \\ x_i|z_i \sim g(\cdot|\theta_{z_i}), \ \text{for } i = 1, \dots, n \end{cases}$$

where $z_i \in \{1, \dots, K\}$ is the cluster membership of x_i, and $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ are the component weights, and $\{\theta_k\}_{k=1}^K$ are component parameters.

Typical "Bayesian clustering" is model-based

Starting from a simple (e.g. Gaussian) component kernel $g(y|\theta)$:

$$x_1, \ldots, x_n \sim \sum_{k=1}^{K} \pi_k g(\cdot|\theta_k) \quad \text{or} \quad \begin{cases} z_i|\boldsymbol{\pi} \sim \text{Categorical}(\pi_1, \ldots, \pi_K) \\ x_i|z_i \sim g(\cdot|\theta_{z_i}), \text{ for } i = 1, \ldots, n \end{cases}$$

where $z_i \in \{1, \ldots, K\}$ is the cluster membership of x_i, and $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ are the component weights, and $\{\theta_k\}_{k=1}^K$ are component parameters.

One of the following two priors are commonly used for tractability:

Mixture of Finite Mixtures (Miller & Harrison 2018) with $K < \infty$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$
$$\theta_k \sim G_0(\cdot) \text{ for } k = 1, \ldots, K$$
$$K \sim p_K(\cdot)$$

Dirichlet Process Mixture (e.g. Lo 1984; Neal 2000) with $K = \infty$

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots) \sim \text{StickBreaking}(\alpha)$$
$$\theta_k \sim G_0(\cdot) \text{ for } k = 1, 2, \ldots$$
$$\alpha \sim \text{Gamma}(a, b)$$

A decision is required to obtain final clustering

Conditional on data $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, we get a joint posterior distribution on $\mathbf{z} = (z_1, \ldots, z_n)$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, and $\{\theta_k\}_{k=1}^K$ (and possibly K).

- $\mathscr{C} = \{C_{h_1}, \ldots, C_{h_H}\}$ is the partition of $\mathcal{X}_n$ induced by $\mathbf{z}$, i.e. $C_h \doteq \{x_i : z_i = h\}$.
- This induces a posterior on $\mathscr{P}(\mathcal{X}_n)$, the set of all partitions of $\mathcal{X}_n$.
- In practice, we have a sample of partitions $\{\mathscr{C}^{(s)}\}_{s=1}^S$ from MCMC.

A decision is required to obtain final clustering

Conditional on data $\mathcal{X}_n = \{x_1, \ldots, x_n\}$, we get a joint posterior distribution on $\mathbf{z} = (z_1, \ldots, z_n)$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, and $\{\theta_k\}_{k=1}^K$ (and possibly K).

- $\mathscr{C} = \{C_{h_1}, \ldots, C_{h_H}\}$ is the partition of $\mathcal{X}_n$ induced by $\mathbf{z}$, i.e. $C_h \doteq \{x_i : z_i = h\}$.
- This induces a posterior on $\mathscr{P}(\mathcal{X}_n)$, the set of all partitions of $\mathcal{X}_n$.
- In practice, we have a sample of partitions $\{\mathscr{C}^{(s)}\}_{s=1}^S$ from MCMC.

Wade and Ghahramani (2018): For a loss $L(\cdot, \cdot)$ on $\mathscr{P}(\mathcal{X}_n)$ (e.g. VI, Binder's), choose the clustering $\hat{\mathscr{C}}$ that minimizes the posterior expected loss:

$$\hat{\mathscr{C}} \approx \arg \min_{\mathscr{C}' \in \mathscr{P}(\mathcal{X}_n)} \frac{1}{S} \sum_{s=1}^{S} L(\mathscr{C}^{(s)}, \mathscr{C}').$$

Solve this optimization using salso package (Dahl, Johnson, Müller, 2022).

Thus when L is a metric $\hat{\mathscr{C}}$ is a posterior Fréchet mean, "averaging" $\{\mathscr{C}^{(s)}\}_{s=1}^n$

Why we like Bayesian clustering

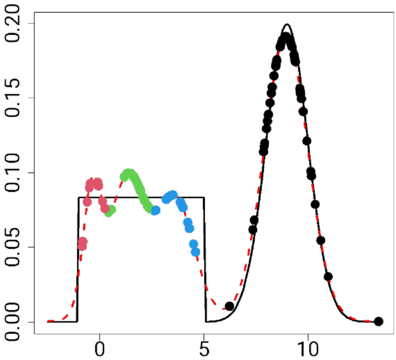
- A statistical/density-based approach to clustering: Data $x_1, \ldots, x_n$ are assumed to be samples from a larger population f, and the clustering is actually driven by inference of f.
- Quantify uncertainty of clustering: Bayesian methods naturally provide a posterior distribution on the space of partitions $\mathscr{P}(\mathcal{X}_n)$ rather than a point estimate.
- Focus on careful modeling of the data using domain-specific prior information rather than experiment with a zillion clustering methods.

The last advantage seems distinctly Bayesian.

See Wade (2023) for a survey on Bayesian clustering.

Limitations of model-based clustering

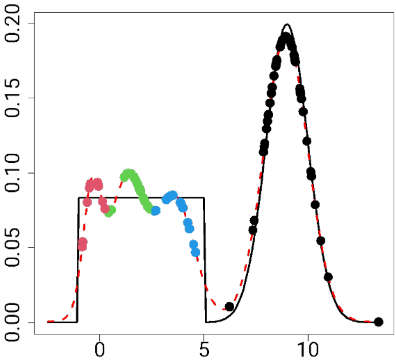
Issue: True clusters are split when the kernel is even slightly misspecified.



solid/black line = true density; red/dashed line = posterior mean density. *shows only a random subsample of the observations.

Limitations of model-based clustering

Issue: True clusters are split when the kernel is even slightly misspecified.



solid/black line = true density; red/dashed line = posterior mean density. *shows only a random subsample of the observations.

Fixes in the Bayesian setting:

- Loss functions (Wade & Ghahramani 2018; Dahl et al. 2022)
- Mode-merging (Dombowski & Dunson 2024)
- Increasing kernel flexibility (Frühwirth-Schnatter & Pyne 2010)
- Mixtures of mixtures (Malsiner-Walli et al. 2017; Stephenson et al. 2019)
- Coarsening (Miller & Dunson, 2018)
- Gibbs posteriors (Rigon et al. 2023)
- Other types of Bayesian clustering?

Outline

Density Based Clustering

What clustering do we want in the limit of infinite data from a density f?

The answer determines a population-level clustering functional:

$$\psi : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X})$$

where

1. $\mathcal{D}(\mathcal{X}) =$ a collection of densities on $\mathcal{X}$
2. $\mathscr{P}(\mathcal{X}) =$ the set of all partitions of $\mathcal{X}$.

Examples:

- If f is an identifiable mixture model then $\psi(f)$ can be its Bayes optimal partition (e.g. Aragam et al. 2020).
- If f is multimodal then $\psi(f)$ could be partition of $\mathcal{X}$ based on the basins of attraction of its modes (e.g. Chacón 2015).
- If f is any density then $\psi_\lambda(f)$ can denote the connected components of the level set $\{f \geq \lambda\}$ (Hartigan 1975; Rinaldo et al. 2012).

Density Based Clustering

What clustering do we want in the limit of infinite data from a density f?

The answer determines a population-level clustering functional:

$$\psi : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X})$$

where

1. $\mathcal{D}(\mathcal{X}) =$ a collection of densities on $\mathcal{X}$
2. $\mathscr{P}(\mathcal{X}) =$ the set of all partitions of $\mathcal{X}$.

Examples:

- If f is an identifiable mixture model then $\psi(f)$ can be its Bayes optimal partition (e.g. Aragam et al. 2020).
- If f is multimodal then $\psi(f)$ could be partition of $\mathcal{X}$ based on the basins of attraction of its modes (e.g. Chacón 2015).
- If f is any density then $\psi_\lambda(f)$ can denote the connected components of the level set $\{f \geq \lambda\}$ (Hartigan 1975; Rinaldo et al. 2012).

Bayesian Density Based Clustering

Given data $\mathcal{X}_n \subseteq \mathcal{X}$, the clustering functional $\psi : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X})$ provides

$$\psi_n : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X}_n) \qquad \psi_n(f) \doteq \psi(f)\big|_{\mathcal{X}_n}$$

a clustering of the data points $\psi_n(f) \in \mathscr{P}(\mathcal{X}_n)$ when the true density f is known.

Bayesian Density Based Clustering

Given data $\mathcal{X}_n \subseteq \mathcal{X}$, the clustering functional $\psi : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X})$ provides

$$\psi_n : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X}_n) \qquad \psi_n(f) \doteq \psi(f)\big|_{\mathcal{X}_n}$$

a clustering of the data points $\psi_n(f) \in \mathscr{P}(\mathcal{X}_n)$ when the true density f is known.

Starting from any prior model $P_M(\cdot)$ for the data generating density f, draw posterior samples of f and compute resulting clustering:

$$f^{(1)}, \ldots, f^{(S)} \sim P_M(\cdot|\mathcal{X}_n) \implies \psi_n(f^{(1)}), \ldots, \psi_n(f^{(S)}) \in \mathscr{P}(\mathcal{X}_n).$$

Bayesian Density Based Clustering

Given data $\mathcal{X}_n \subseteq \mathcal{X}$, the clustering functional $\psi : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X})$ provides

$$\psi_n : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X}_n) \qquad \psi_n(f) \doteq \psi(f)\big|_{\mathcal{X}_n}$$

a clustering of the data points $\psi_n(f) \in \mathscr{P}(\mathcal{X}_n)$ when the true density f is known.

Starting from any prior model $P_M(\cdot)$ for the data generating density f, draw posterior samples of f and compute resulting clustering:

$$f^{(1)}, \ldots, f^{(S)} \sim P_M(\cdot|\mathcal{X}_n) \implies \psi_n(f^{(1)}), \ldots, \psi_n(f^{(S)}) \in \mathscr{P}(\mathcal{X}_n).$$

Final averaging step: Given loss $L(\cdot, \cdot)$ between clustering, we consider the clustering point-estimate as:

$$\hat{\mathscr{C}} \approx \arg \min_{\mathscr{C}' \in \mathscr{P}(\mathcal{X}_n)} \frac{1}{S} \sum_{s=1}^{S} L(\psi_n(f^{(s)}), \mathscr{C}')$$

Bayesian Density Based Clustering

Given data $\mathcal{X}_n \subseteq \mathcal{X}$, the clustering functional $\psi : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X})$ provides

$$\psi_n : \mathcal{D}(\mathcal{X}) \to \mathscr{P}(\mathcal{X}_n) \qquad \psi_n(f) \doteq \psi(f)\big|_{\mathcal{X}_n}$$

a clustering of the data points $\psi_n(f) \in \mathscr{P}(\mathcal{X}_n)$ when the true density f is known.

Starting from any prior model $P_M(\cdot)$ for the data generating density f, draw posterior samples of f and compute resulting clustering:

$$f^{(1)}, \ldots, f^{(S)} \sim P_M(\cdot|\mathcal{X}_n) \implies \psi_n(f^{(1)}), \ldots, \psi_n(f^{(S)}) \in \mathscr{P}(\mathcal{X}_n).$$

Final averaging step: Given loss $L(\cdot,\cdot)$ between clustering, we consider the clustering point-estimate as:

$$\hat{\mathscr{C}} \approx \arg \min_{\mathscr{C}' \in \mathscr{P}(\mathcal{X}_n)} \frac{1}{S} \sum_{s=1}^{S} L(\psi_n(f^{(s)}), \mathscr{C}')$$

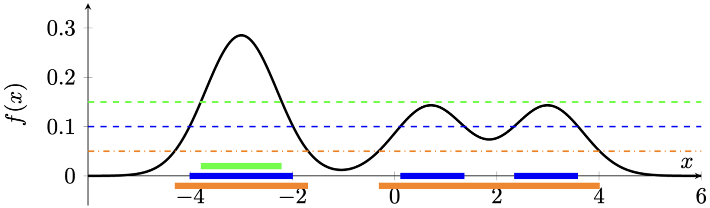
1. Expands the kinds of clustering that can be considered in the Bayesian framework.
2. Separates density estimation from clustering so that any model can be used.

Level set clustering

Active area since Wishart (1969) and Hartigan (1975).

Given f and level $\lambda > 0$, the level set clustering is a sub-partition of $\mathcal{X}$ defined as

$$\psi_\lambda(f) \doteq \text{Connected components of } \{x \in \mathcal{X} : f(x) \geq \lambda\}$$



- Heuristics to choose λ using elbow plots or a fixed fraction of noise points when clusters are well separated (Ester et al. 1996, Cuevas et al. 2001)
- In generally examine the cluster tree over all $\lambda > 0$ (Campello et al. 2015, Steinwart et al. 2023)

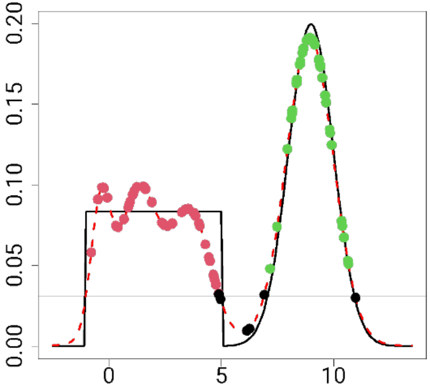
Bayesian Level Set (BALLET) Clustering

We implement the previous methodology by

- using a computable surrogate $\hat{\psi}_{\delta,\lambda}(f)$ from level set clustering literature, and
- modifying Binder's loss to give a metric on the space of sub-partitions of $\mathcal{X}_n$.

Important notes:

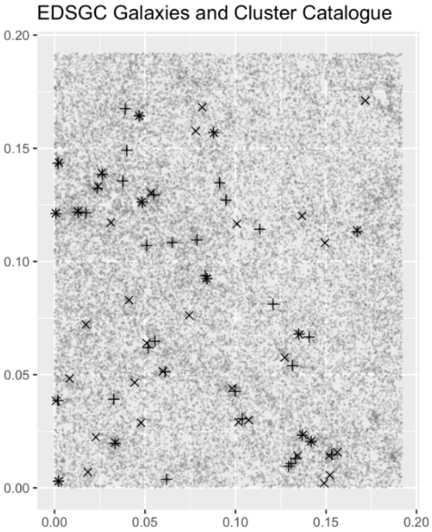
- Points with $f(x_i) < \lambda$ are declared as noise. (Black points in the figure)
- Level $\lambda > 0$ is a loss parameter and not part of the model (thus not learned from data). We use previous strategies.
- Compared to DBSCAN (Ester et al. 1996), we allow use of carefully chosen priors and quantifying clustering uncertainty.



Outline

Edinburgh-Durham Southern Galaxy Catalog

- Around $41K$ galaxies (grey points) observed in a $10^\circ \times 10^\circ$ section of the sky (Nichol et al., 1992).
- Level set clustering corresponding to scientifically motivated λ can help understand cosmological models (Jang, 2006).
- Available catalogs of *suspected* galaxy clusters for validation
  - '$+$' Abell catalog (Abell et al., 1989) – handpicked.
  - '$\times$' EDCCI (Lumsden et al., 1992) – software generated.



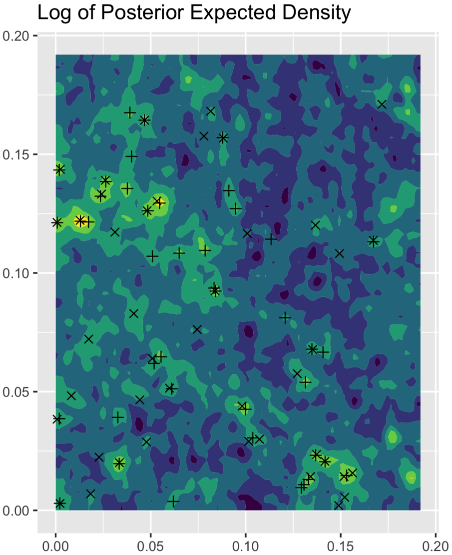
Fast density sampling using mixture of histograms

For fast sampling of density f from its posterior ($n \approx 40K$ data points), we model f as a mixture of $K = 50$ histograms

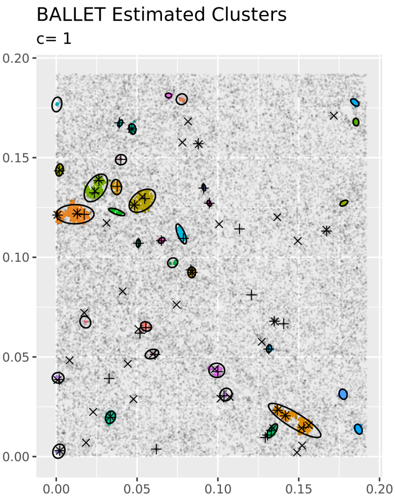
$$f(x) = \frac{1}{K} \sum_{k=1}^{K} H(x; B_k, \vec{\rho_k}),$$

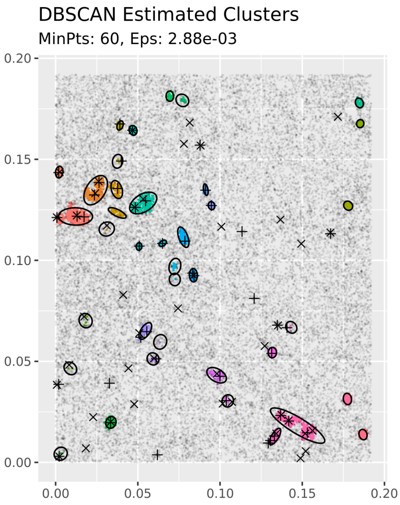
where $H(x; B_k, \vec{\rho_k})$ is a histogram density estimator with bins B_k (fixed) and weight vector $\vec{\rho}_k$.

Next we use a mean-field type variational approximation for the joint posterior of $\{\vec{\rho}_k\}_{k=1}^{K}$ by independently sampling each $\vec{\rho}_k$ based on all the data (conjugate).



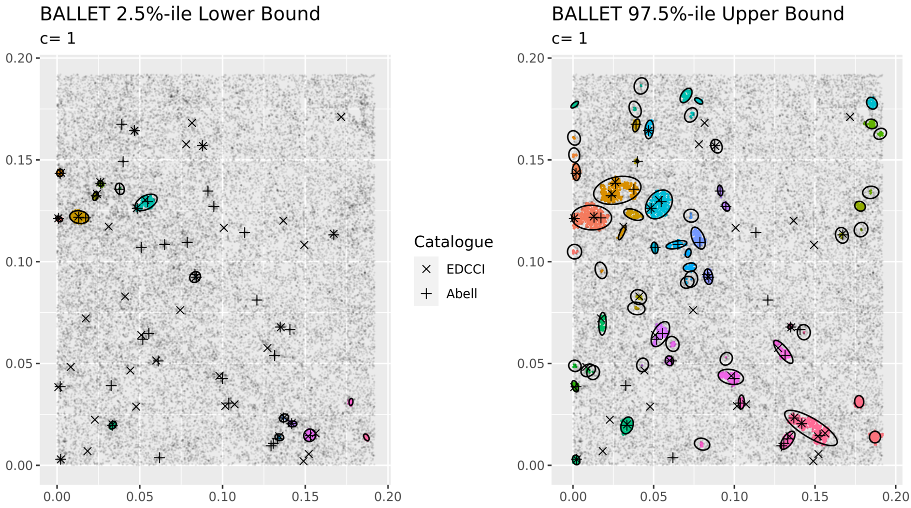
BALLET vs DBSCAN clustering





DBSCAN parameter was hand-tuned to avoid many false positives. In contrast, BALLET results were stable to the choice of its parameter δ (but not the level λ).

BALLET clustering uncertainty: 95% credible bounds



Like Wade & Ghahramani (2018) we summarize the 95% credible ball using upper and lower bounds using an associated Hasse diagram on the space of sub-partitions.

Validation of clusters against known catalogs

EDCCI catalog

Method	Sensitivity	Specificity	Exact Match
DBSCAN	0.71	0.25	0.23
DBSCAN[1]	0.69	0.63	0.45
BALLET Lower	0.29	**0.87**	**0.67**
BALLET Est.	0.67	0.69	0.51
BALLET Upper	**0.86**	0.42	0.32

Abell catalog

Method	Sensitivity	Specificity	Exact Match
DBSCAN	0.40	0.18	0.16
DBSCAN[1]	0.37	0.42	0.34
BALLET Lower	0.21	**0.73**	**0.67**
BALLET Est.	0.40	0.40	0.26
BALLET Upper	**0.56**	0.34	0.27

Conclusion

- We propose a framework for Bayesian density based clustering that separates density estimation from clustering.
- This clustering is consistent as long as the map $f \mapsto \psi(f)$ is "continuous" and the density estimation is consistent. We carefully check these conditions for `BALLET`.
- Application to the galaxy clustering problem. Compared to `DBSCAN`, `BALLET` provides clustering uncertainty and allows careful prior modeling.

Future Directions:

- Handle overlapping clusters: Approach modal clustering by the connection to level set cluster tree tree (Arias-Castro & Qiao, 2023).
- High-dimensional setting: Cluster latent factors (Chandra et al. 2023).
- Regression setting: See Chacón (2020).

Thank You!

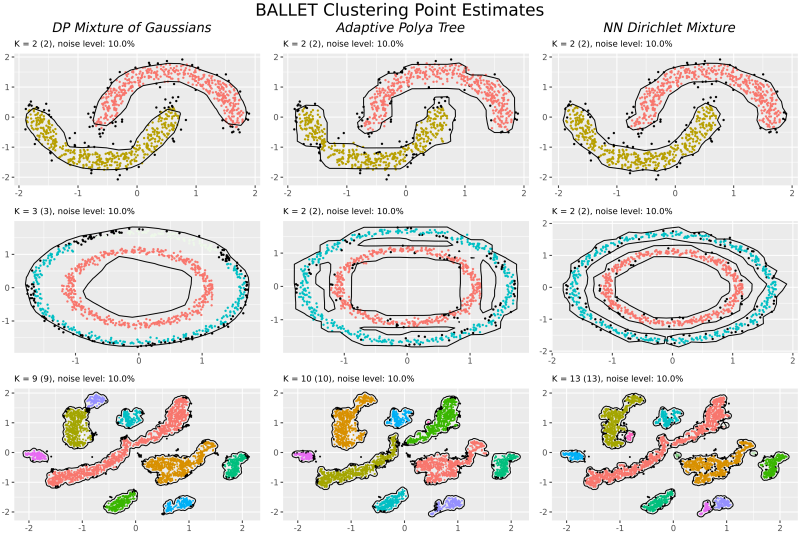
Any questions or suggestions?

Pre-print: https://arxiv.org/abs/2403.04912

Email: mdewaskar@unm.edu

Acknowledgements: This work was partially funded by grants R01-ES028804 and R01-ES035625 from the NIH and N00014-21-1-2510 from ONR.

Toy clustering across different models



BALLET implementation details

Problem: How to compute $\psi_\lambda(f)$?

Following the level set clustering literature (Rinaldo and Wasserman, 2010; Sriperumbudur and Steinwart, 2012), we use a surrogate based on the Devroye and Wise (1980) estimator for $\{f \geq \lambda\}$:

$$\tilde{\psi}_{\delta,\lambda}(f) = \mathtt{CC}(G_\delta\{x_i \in \mathcal{X}_n : f(x_i) \geq \lambda\})$$

that can be computed by single linkage clustering.

Problem How to choose δ?

- Given $\lambda > 0$, we recommend the data-adaptive choice

$$\hat{\delta} = q_{.99}\{d_k(x_i) : f(x_i) \geq \lambda\}$$

  where q is the quantile function and $d_k(x)$ is the k-NN distance of x to $\mathcal{X}_n$.
- As long as $k \gg \log n$, we show that BALLET estimator is consistent with this choice of $\hat{\delta}$.

More details: Sub-partitions (forms a lattice), choice of loss (modified Binder's applicable to sub-partitions), and solving the optimization using SALSO.

Consistency of Bayesian Density-based clustering

Suppose $x_1, \ldots, x_n \stackrel{iid}{\sim} f_0$. Assume further that:

1. The loss $L : \mathscr{P}(\mathcal{X}_n) \times \mathscr{P}(\mathcal{X}_n) \to [0,1]$ is a metric
2. There is a metric ρ on $\mathcal{D}(\mathcal{X})$ such that the posterior $P_M(\cdot|\mathcal{X}_n)$ contracts at rate $\{\epsilon_n\}$ to f_0 in the sense that for any sequence $\{K_n\} \to \infty$,

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

3. ψ_n is suitably continuous at f_0 with respect to ρ and L, i.e.

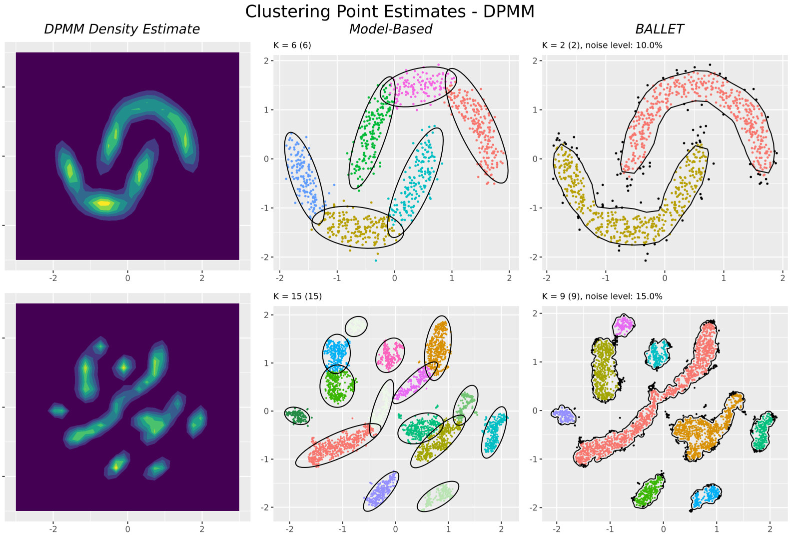
$$\tau_2(\mathcal{X}_n) = \sup_{f : \rho(f, f_0) \le K_n \epsilon_n} L(\psi_n(f), \psi_n(f_0)) \xrightarrow{P} 0 \text{ as } n \to \infty$$

Then triangle inequality shows that our Bayesian density-based clustering point $\hat{\mathscr{C}}$ is consistent for $\mathscr{C}_0 = \psi_n(f_0)$, namely

$$L(\hat{\mathscr{C}}, \mathscr{C}_0) \le 2\tau_1(\mathcal{X}_n) + 2\tau_2(\mathcal{X}_n) \xrightarrow{P} 0 \text{ as } n \to \infty.$$

In our manuscript, we verify conditions 1 & 3 for level set clustering $\psi = \psi_\lambda$ assuming that condition 2 holds for some $\epsilon_n \to 0$ with $\rho(f, g) = \|f - g\|_\infty$.

Finding arbitrary shaped clusters using DPMM



Top panel: simulated two-moons data. Bottom panel: tSNE plot of 4406 cells and 2000 genes from `https://www.reneshbedre.com/blog/tsne.html`