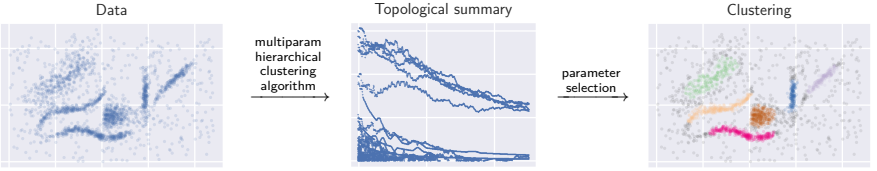
Density-based clustering and multiparameter persistence

Luis Scoccola
Michigan State University
jww Alexander Rolle

Workshop on Metrics in Multiparameter Persistence,
Lorentz Center, July 21st, 2021



Data from (McInnes, Healy, Astels, 2017).

Clustering

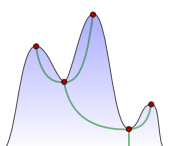
Clustering problem: Group points of dataset X into "natural" classes.

Not well-defined.

Def: $\mathsf{C}(X) = \{\text{families of non-empty, disjoint subsets of } X\}$ (poset).

Density-based clustering:

- ▷ Assume $X \subseteq \mathbb{R}^d$ sampled from p.d.f. $f : \mathbb{R}^d \to \mathbb{R}$. Let $S = \operatorname{supp}(f)$.
- ▷ Define the hierarchical clustering $H(f) : \mathbb{R} \to \mathsf{C}(S)$
$$r \mapsto \pi_0\{f \geq r\}.$$



- ▷ *Problem 1*: estimate $H(f)$ from X.
- ▷ *Problem 2*: extract a flat clustering from the estimate.

Problems: Estimate $H(f)$ and extract flat clustering from estimate.

Outline:

1. Hierarchical clusterings and correspondence-interleaving distance.
2. Stable algorithms to approximate $H(f)$ from sample.
3. A stable flattening construction.
4. Parameter selection.

See (Rolle, S., 2021) for details.

1. Hierarchical clustering

Definition: A (multiparameter) hierachical clustering is any poset map

$$\mathbb{R}^n \to \mathsf{C}(X).$$

Example: Single-linkage $\mathrm{SL}(X) : \mathbb{R} \to \mathsf{C}(X)$

$$r \mapsto \pi_0(\mathrm{VR}(X)(r)).$$

- ▷ Being (multi)persistent objects, get notion of interleaving.
- ▷ Let $R \subseteq X \times Y$ correspondence. Hierarchical clusterings H, E of X, Y are **interleaved wrt R** if $\pi_X^*(H)$ and $\pi_Y^*(E)$ are interleaved.

Definition: The **correspondence-interleaving distance**

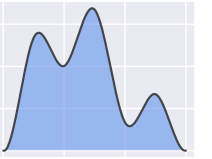
$$d_{\mathrm{CI}}(H, E) = \inf\{\varepsilon : \ H \text{ and } E \text{ are } (\varepsilon, \ldots, \varepsilon)\text{-int. wrt some correspondence}\}.$$

2. Stable hierarchical clustering

Fix X metric probability space and $K : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ non-increasing kernel.

The **kernel filtration** of X at $s, k > 0$ is:

$$X_{[s,k]} = \left\{ x \in X \,:\, \int_{y \in X} K\left(\frac{d_X(x,y)}{s}\right) \mathrm{d}\mu_X \geq k \right\}$$



Kernel linkage $\mathbb{L}^K(X) : \mathbb{R}^3 \to \mathrm{C}(X)$

$$(s,t,k) \mapsto \mathrm{SL}\left(X_{[s,k]}\right)(t)$$

Given $0 < x, y \leq \infty$, restrict $\mathbb{L}^K(X)$ to $\lambda^{x,y}$.

Get $\lambda^{x,y}\texttt{-link}(X) : \mathbb{R} \to \mathrm{C}(X)$

- $\lambda^{\infty,y}$-`link` = robust single-linkage of (Chaudhuri, Dasgupta).
- $\lambda^{x,\infty}$-`link` = Rips-graph filtered by density estimate.
- Also recover π_0 of degree-Rips (Lesnick, Wright).

Thm (Rolle, S.): Let $x, y < \infty$.

- Kernel linkage and $\lambda^{x,y}$-`link` are stable wrt Gromov–Hausdorff–Prokhorov distance and correspondence-interleaving distance.
- $\lambda^{x,y}$-`link` is continuous in x and y.
- $\lambda^{x,y}$-`link` is consistent.

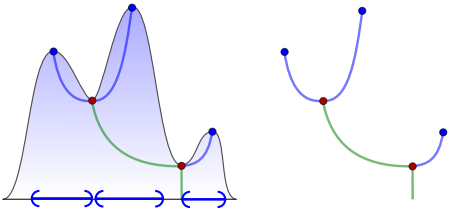
Stability does not hold for vertical or horizontal lines.

Solves problem 1.

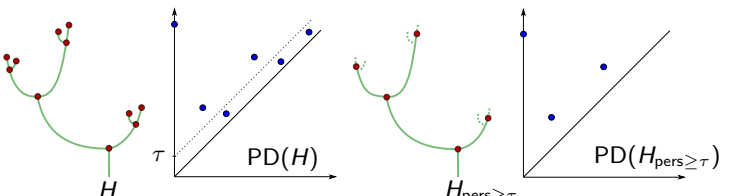
3. Stable flattening

Flatten H by taking its **leaves**:

Stable if leaves have large persistence.



Before flattening remove branches with persistence less than τ: $H_{\text{pers}\geq\tau}$.



(Chazal, Guibas, Oudot, Skraba): Choose τ using PD(H).

Solves problem 2.

Takeaway: Want PD(H) to have large gap in total persistences.

4. Parameter selection

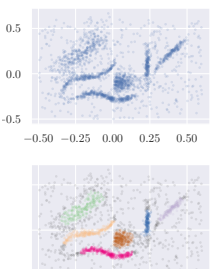
Given dataset X, have clustering algorithm $\lambda, \tau \mapsto \mathrm{PF}(\lambda\text{-}\texttt{link}(X), \tau)$.

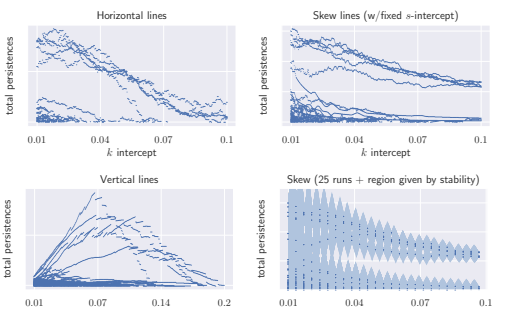
Goal: Choose λ st $\mathrm{PD}(\lambda\text{-}\texttt{link}(X))$ has good separation in total pers.

Do this by looking at total pers. as λ varies in some family of lines.

Def: the **total persistence diagram** of a persistence diagram $\{(x_i, y_i)\}$ is the multiset of persistences $\{y_i - x_i\}$.

Example: Total persistences of 3 families of curves.





Stability gives "100%-confidence region".

Remarks

- Visualize total persistence of all slices at once?
- Robustness (to outliers)?
  - Consider other filtrations.
  - Consider other distances between hierarchical clusterings.
- A subquadratic implementation (current work).
- Can use our implementation to compute persistent homology of slices of "Kernel-Rips" and degree-Rips.

- Campello, Moulavi, Sander. Density-based clustering based on hierarchical density estimates.
- Carlsson, Mémoli. Multiparameter hierarchical clustering methods.
- Chaudhuri, Dasgupta. Rates of convergence for the cluster tree.
- Chazal, Guibas, Oudot, Skraba. Persistence-based clustering in Riemannian manifolds.
- Cohen-Steiner, Edelsbrunner, Morozov. Vines and vineyards by updating persistence in linear time.
- McInnes, Healy, Astels. hdbscan: Hierarchical density based clustering.
- Rolle, S.. Stable and consistent density-based clustering.
  Code: `https://github.com/LuisScoccola/gamma-linkage`

Thank you for your attention!