Topological mixture estimation

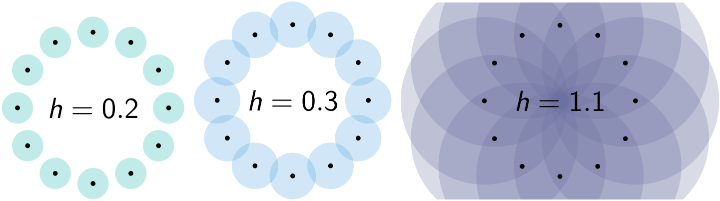
Steve Huntsman

FAST Labs (goo.gl/zWw5HQ)
Cyber Technology

17 May 2018

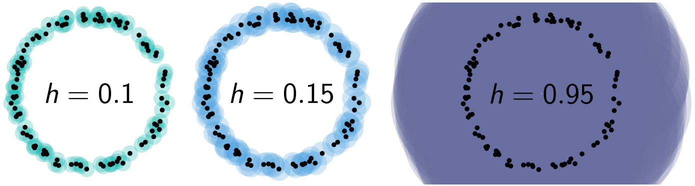
The rudiments of topological data analysis

- Approximate $X = (X_j)_{j=1}^n \subset (\mathbb{R}^d)^n$ at various scales $h > 0$
- Compute a topological invariant of the approximations
- Highlight invariants that persist as h varies
- Toy example: 12 equispaced points on S^1 ($\Delta \approx 0.2588$)
  - $\beta_0 = 12 \cdot 1_{[0,\Delta)} + 1_{[\Delta,1)} + 1_{[1,\infty)}$
  - $\beta_1 = 1_{[\Delta,1)}$



Belaboring some details

- There exists $\varepsilon > 0$ s.t. $\beta_0([0, \varepsilon)) = n$ (undersmoothing)
- $\beta_0([\text{diam } X, \infty)) = 1$ (oversmoothing)
- Persistent structure is stable under perturbations
- Example: $n = 100$, $\delta = 0.05$, $|X| \sim U([1-\delta, 1+\delta])$, $\varepsilon' \approx \delta$
  - $\beta_0 = n \cdot 1_{[0,\varepsilon)} + \text{noise}_0 \cdot 1_{[\varepsilon,\Delta)} + 1_{[\Delta,1-\varepsilon')} + 1_{[1-\varepsilon',\infty)}$
  - $\beta_1 = \text{noise}_1 \cdot 1_{[\varepsilon,\Delta)} + 1_{[\Delta,1-\varepsilon')}$

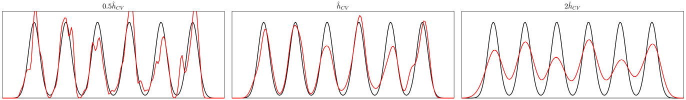


Is topological data analysis interesting or useful for $d = 1$?

Yes and YES

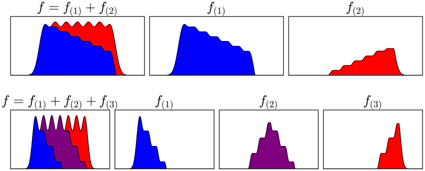
Approximate: kernel density estimate

- $X \sim f \approx \hat{f}_{h;X} := \frac{1}{n}\sum_{j=1}^{n} K_{X_j,h}$ with $K_{\mu,h}(x) := \frac{1}{h}K\left(\frac{x-\mu}{h}\right)$
- To estimate f, just select a good *bandwidth* h
  - Choice of K much less important than h as long as K is nice
- Least-squares *cross-validation* (CV) optimizes L^2 error estimate without *a priori* information about f
  - Epanechnikov is L^2 optimal choice for K, but see above
  - Gaussian is also good and very convenient
- If f is multimodal but we have no *a priori* information, the literature doesn't offer any significant improvement on CV



Compute: topologically optimal mixture

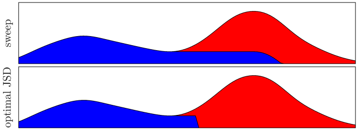
- A PDF $f \in C(\mathbb{R}^d)$ is *unimodal* iff $f^{-1}([y, \infty))$ is contractible (i.e., homotopic to a point) for all $0 < y \leq \max f$
  - For $d = 1$, contractible = interval (captures intuitive notion)
- ucat(f) is the smallest number M s.t. $f = \sum_{m=1}^{M} \pi_m p_m$ for some $\pi > 0$, $\sum_m \pi_m = 1$, and p_m unimodal: write $(\pi, p) \models f$
- Lemma: ucat is invariant under homeomorphism
- Theorem [Baryshnikov and Ghrist]: For $d = 1$, a "sweep" algorithm efficiently produces a $(\pi, p) \models f$

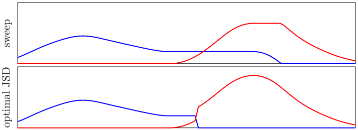


BAE SYSTEMS INSPIRED WORK

Compute: information theoretically optimal mixture

- Mixture $(\pi, p) \models f$ produced by sweep algorithm is ugly
- Want to preserve predicate $(\pi, p) \models f$ while maximizing mutual info $J(\pi, p)$ between $\Xi \sim \pi$ and $X \sim \sum_m \pi_m p_m$
- Two lemmas regarding local perturbations of piecewise affine (π, p) that preserve $(\pi, p) \models f$:
  - J is convex
  - Explicit tight bounds on perturbations
- Theorem: Greedy unimodality-preserving local perturbations obtain $\arg\max_{(\pi,p)\models f} J(\pi, p)$ in $O(\text{ucat} \cdot |\text{mesh}|)$ iterations

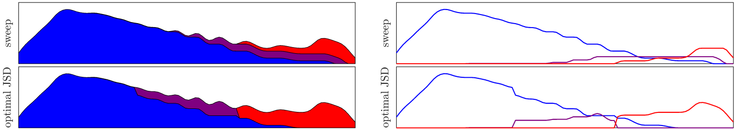




BAE SYSTEMS
INSPIRED WORK

Compute: information theoretically optimal mixture

- Mixture $(\pi, p) \models f$ produced by sweep algorithm is ugly
- Want to preserve predicate $(\pi, p) \models f$ while maximizing mutual info $J(\pi, p)$ between $\Xi \sim \pi$ and $X \sim \sum_m \pi_m p_m$
- Two lemmas regarding local perturbations of piecewise affine (π, p) that preserve $(\pi, p) \models f$:
    - J is convex
    - Explicit tight bounds on perturbations
- Theorem: Greedy unimodality-preserving local perturbations obtain $\arg\max_{(\pi,p)\models f} J(\pi, p)$ in $O(\mathrm{ucat} \cdot |\mathrm{mesh}|)$ iterations



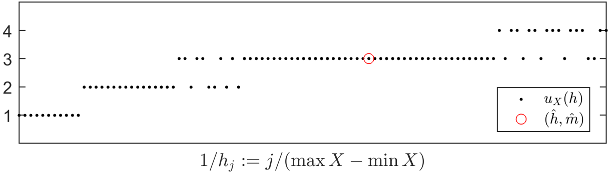
Topological density estimation (TDE)

- Approximate X at various scales $h > 0$ by $\hat{f}_{h;X}$
- Compute the topological invariant $u_X(h) := \text{ucat}(\hat{f}_{h;X})$
- Highlight the most persistent value:

$$\hat{m} := \arg\max_m \mu(u_X^{-1}(m))$$

- Topological bandwidth estimate

$$\hat{h} := \text{median}_\mu(u_X^{-1}(\hat{m}))$$

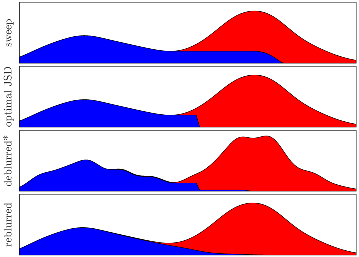


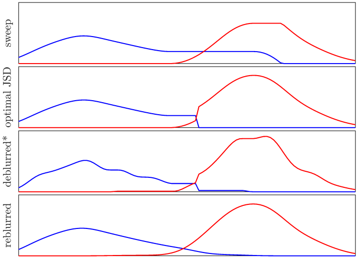
BAE SYSTEMS INSPIRED WORK

(De/re)blurring

- TDE and TME are better together than apart
- For convenience take K = Gaussian
  - Gives semigroup $K_{\mu,h} * K_{\mu',h'} = K_{\mu+\mu',(h^2+h'^2)^{1/2}}$
  - Any other choice of K yielding a semigroup would also work
- Deblur by replacing median (or similar) with inf (or similar)
- Reblur by convolving with $K_{0,\Delta h}$ to recover original density
  - Preserves topological optimality
  - Sensibly trades information-theoretical, smoothness objectives
  - Could do something similar directly on a density via Fourier deconvolution, but this is delicate and of more limited utility

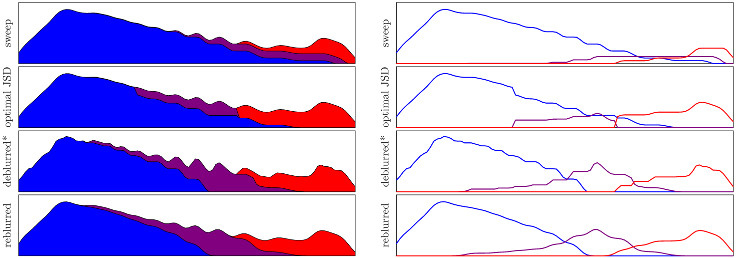
Putting it all together: $n = 272$ Old Faithful data





BAE SYSTEMS
INSPIRED WORK

Putting it all together: $n = 2107$ color index data



BAE SYSTEMS
INSPIRED WORK

$d > 1$ is hard to get but would be nice to have

- For $d = 2$ computing ucat is hard
  - ucat(f) determined by the combinatorial type of the Reeb graph of f labeled by critical values (not true for $d > 2$)
  - No explicit efficient algorithm known
  - Partial bound established in unpublished work of Hickok, Villatoro, and Wang
- For $d \gg 1$ computing ucat is undecidable
  - Will have to restrict notion of unimodality and/or employ heuristics, probably tomographic in spirit and using sheaves/Morse theory
- Determining the number of clusters/mixture components in data is an important but delicate art, with subjective results
  - Obvious that this should be informed by topology
  - Not obvious how to account for variable density and overlap
  - Random projections can help

Current and future directions

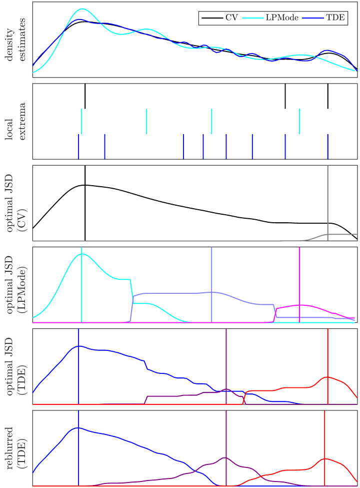
- $d = 1$:
  - Recursive TME/variable bandwidth estimation
  - Enhanced LODA algorithm
  - Component-aware fusion
  - ...
- $d > 1$: topological mixture tomography
  - Glue together 1D mixtures using some combination of random projections, sheaves, Morse theory, etc.
  - Need a decidable restriction of unimodality for nice theory; need an efficient restriction for nice practice

Thanks

steve.huntsman@baesystems.com
FAST Labs (goo.gl/zWw5HQ)

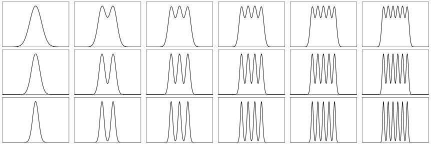
MATLAB code available at goo.gl/KoBtHb

Comparative analysis: $n = 2107$ color index data

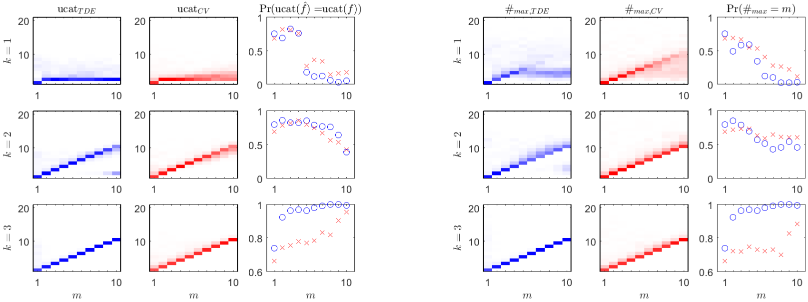


BAE SYSTEMS
INSPIRED WORK

Performance of TDE vs CV on a multimodal family



Densities f_{km} for $1 \leq k \leq 3$ and $1 \leq m \leq 6$ over $[-0.5, 1.5]$.
Rows and column indices are k and m, respectively; upper left and lower right panels respectively show f_{11} and f_{36}.



TDE: broad observations

- TDE requires no nontrivial information or parameters
- TDE's estimated lower bound on modes is itself very useful
- TDE can sense if its output is appropriate to use
  - f has ≥ 2 clearly resolvable modes $\Rightarrow$ TDE is a good choice
  - Otherwise TDE is the wrong choice
- TDE works well on multimodal distributions because unimodal decompositions are stable w.r.t. perturbations and persist over bandwidths that correctly resolve extrema
  - Maybe there is a nice theorem (exercise for the audience)
- TDE is usually faster than CV for large n
  - Both spend most of their time evaluating and summing kernels
  - CV: $2n_h n^2$ evals; TDE: $n_h n_x n$ evals; typically $n_x \ll n$ if n large
  - Unimodal decompositions contribute a marginal $O(n_h n_x \langle m \rangle_h)$