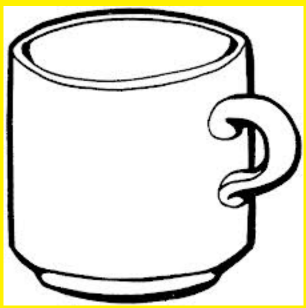
Statistical Inference For Functional Summaries of Persistent Homology

Larry Wasserman
CMU
www.stat.cmu.edu/topstat

Co-authors:
Brittany Fasy, Fabrizio Lecci
Fred Chazal, Alessandro Rinaldo

TOPSTAT

www.stat.cmu.edu/topstat

 $\longrightarrow$ $\longrightarrow$ $\longrightarrow$ 

PEOPLE

Sivaraman Balakrishnan	Yen-Chi Chen
Jessi Cisewski	Brittany Fasy
Christopher Genovese	Brian Kent
Jisu Kim	Fabrizio Lecci
Alessandro Rinaldo	Aarti Singh
Isa Verdinelli	Larry Wasserman

Honorary members: Fred Chazal, Don Sheehy.

OUTLINE

Main Topic: Random sets $K_1, \ldots, K_n \sim P$. Infer P or the "average homology" using functional summaries. Includes "meta-persistent homology." (Use persistent homology to study persistent homology.)

Time permitting:

The tyranny of tuning parmeters. Find data-driven methods for choosing tuning parameters.

Other TopStat Stuff (later this week): Inference for Persistence diagrams, Metric graphs, Lyman α reconstruction, Density trees.

Problem: Random sets $K_1, \ldots, K_n \sim P$. Infer P or the "average homology" using functional summaries.

One approach: compute persistence diagram D_i for eack K_i. Then take the Fréchet average i.e. find D to minimize

$$\sum_i d_\infty(D, D_i).$$

This turns out to involve some subtle complications. See Turner et al (2012) and Munch et al (2013).

We take a different approach. Convert each D_i into a function F_i (functional summary) and work with the functions $F_1, \ldots, F_n$. These are random fuctions:

$$\boldsymbol{F_1, \ldots, F_n \sim P}.$$

The mean is $\mu(t) = \mathbb{E}[F_i(t)]$.

FUNCTIONAL SUMMARIES

Landscapes (Bubenik 2012), Silhouettes, Barcode intensity, Persistence Intensity (Edelsbrunner, Pranav), Salience (Doraiswarmy et al).

The advantage of function-valued summaries of persistent homology is that we can analyze them using existing techniques from probability and nonparametric statistics. In particular we look at:

- means
- weak convergence
- bootstrap
- functional clustering
- meta-persistent homology

TWO SCENARIOS

Scenario 1:

$K_1, \ldots, K_n \sim P$.

$K_i \longrightarrow D_i \longrightarrow F_i$.

Goal is to infer $\mu = \mathbb{E}(F_i)$ (and other things).

There are many ways of going from K_i to D_i. In fact, we may have

$$K_i \longrightarrow \mathrm{Data} \longrightarrow D_i$$

but we ignore this (until Wed morning.)

TWO SCENARIOS

Scenario 2: We have a very large dataset

$$\mathcal{D}_N = \{Y_1, \ldots, Y_N\}$$

with N points. The data define a diagram D and functional summary F. But it may be hard to compute D when N is large.

Draw n subsamples, $S_1, \ldots, S_n$ from $\mathcal{D}_N$ where $|S_i| = m < N$. We have:
$S_i \longrightarrow D_i \longrightarrow F_i$.

Let $\mu_m = \mathbb{E}(F_i)$. Then

$$||\widehat{\mu}_m - F||_\infty \leq ||\widehat{\mu}_m - \mu_m||_\infty + ||\mu_m - F||_\infty = I + II.$$

Today we only deal with I.

BUBENIK'S LANDSCAPES

Start with a persistence diagram D or barcodes B. We regard this as a set of intervals (birth and death times):

$$B = D = \{(b_j, d_j) : \ j = 1, \ldots, m\}.$$

For simplicity we assume that $m < \infty$.

Also, we assume that

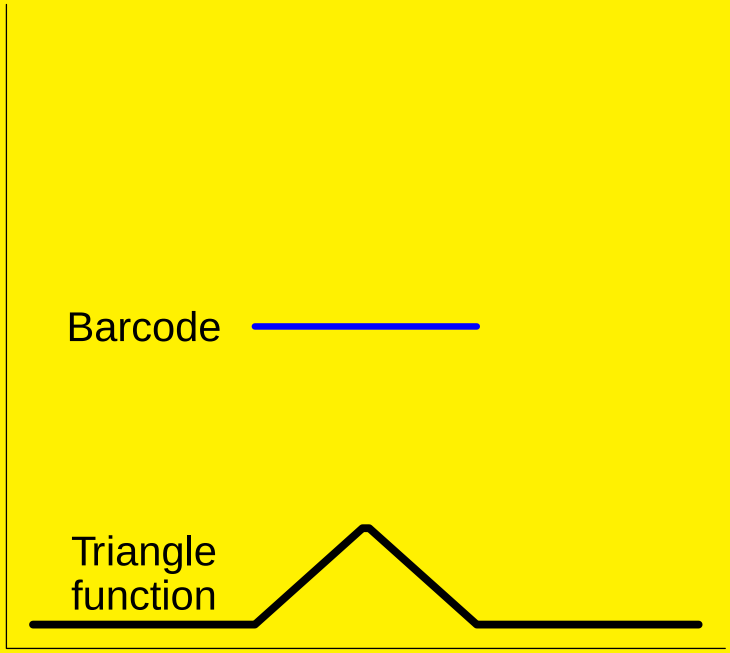
$$0 \le b_j \le d_j \le T$$

for some fixed $T < \infty$.

BUBENIK'S LANDSCAPES

Step 1: Convert each (b_j, d_j) into a triangle function:

$$T_j(t) = \left[(t - b_j) \wedge (d_j - t)\right]_+.$$



Step 2: convert the bag of triangle functions $\{T_j\}$ into a summary function such as

$$\Lambda(t) = \max_j T_j(t).$$

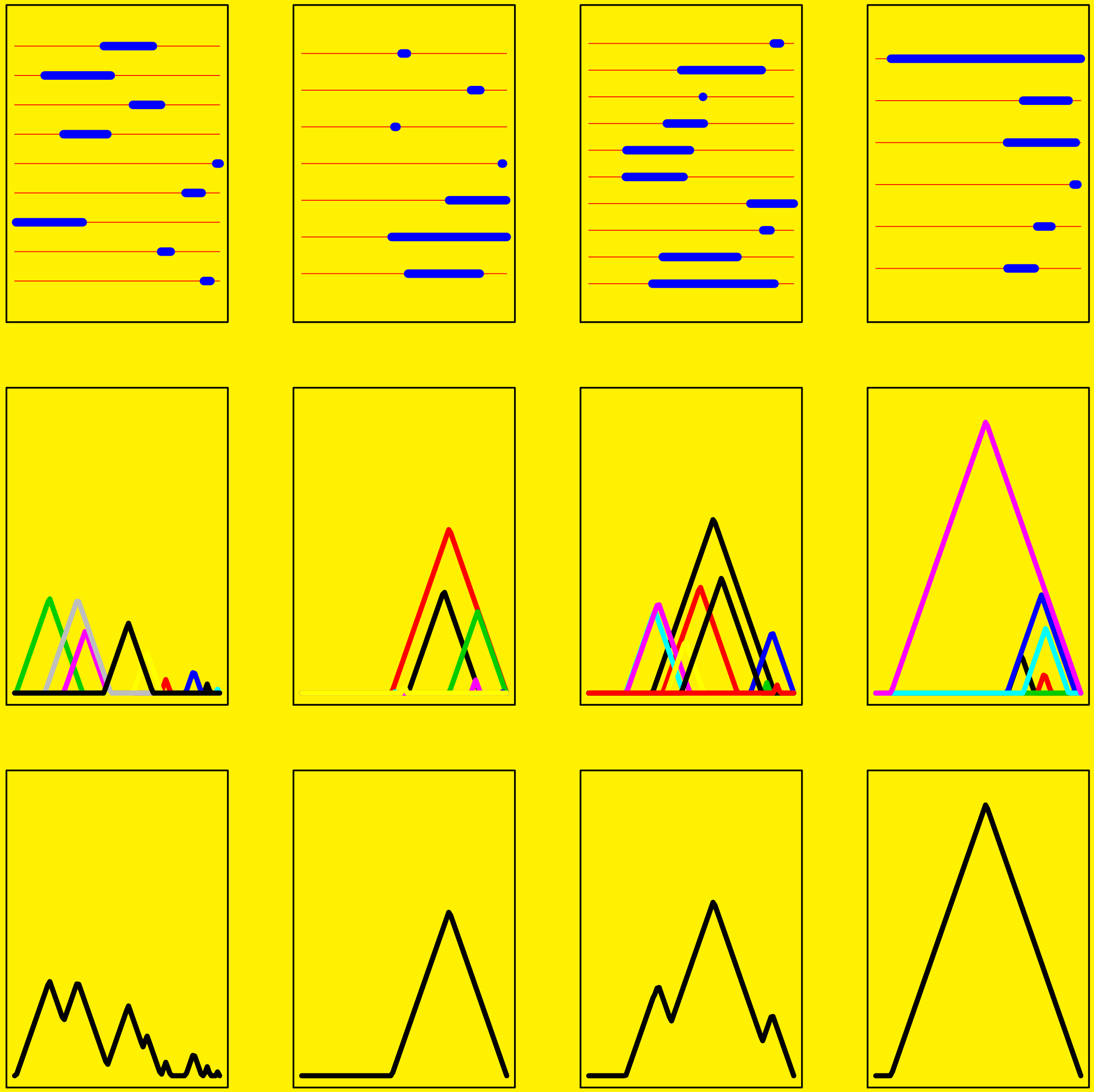
Bubenik also considers second biggest, third biggest etc. We will focus on the max for simplicity.

Note that Λ is 1-Lipschitz.

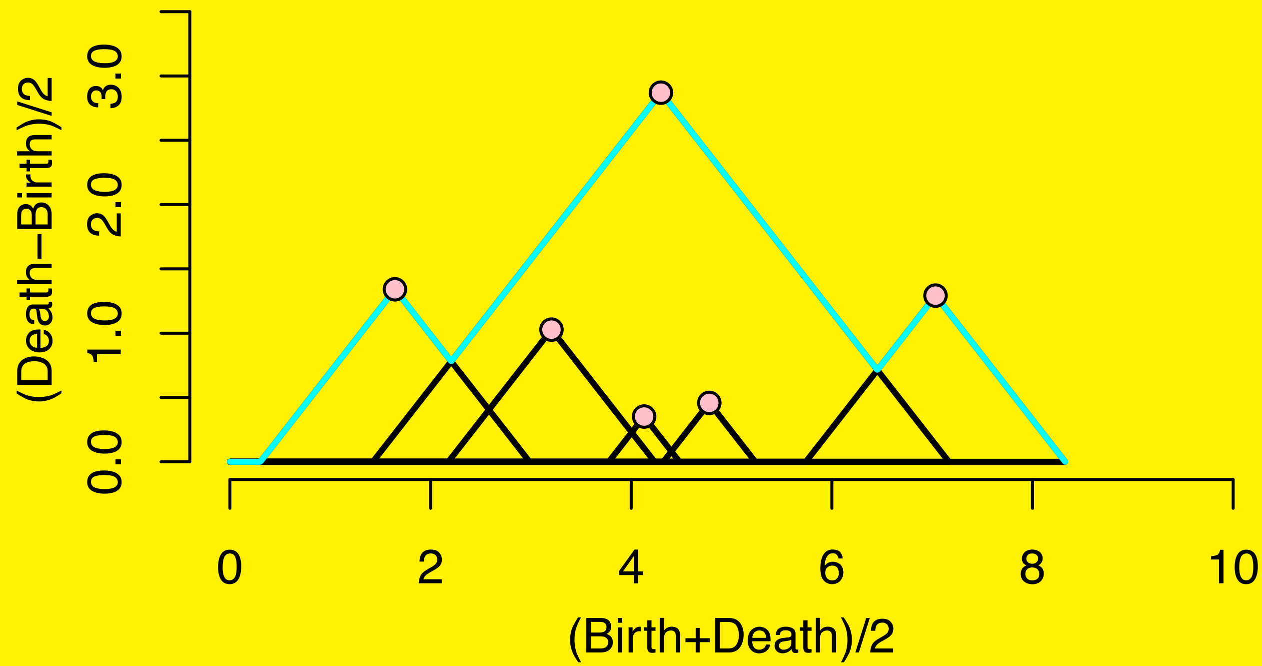
So now we have:

$$D_i \longrightarrow \{T_j\}_{j=1}^m \longrightarrow \Lambda_i$$

for $i = 1, \ldots, n$.



Rotated Persistence Diagram $\longrightarrow$ Landscapes



Let $\mathcal{L}_T$ denote the space of persistence landscapes corresponding to the set of diagrams $\mathcal{D}_T$.

Let P be a probability distribution on $\mathcal{L}_T$, and let

$$\Lambda_1, \ldots, \Lambda_n \sim P.$$

We define the mean landscape as

$$\mu(t) = \mathbb{E}[\Lambda_i(t)], \quad t \in [0, T].$$

Point estimate:

$$\widehat{\mu}(t) \equiv \overline{\Lambda}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \Lambda_i(t).$$

Ultimately we want to find ℓ, u such that

$$\mathbb{P}\Big(\ell(t) \leq \mu(t) \leq u(t) \ \text{ for all } t\Big) \geq 1 - \alpha.$$

Recall that

$$\overline{\Lambda}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\Lambda_i(t), \quad t \in [0,T].$$

Note that $\mathbb{E}(\overline{\Lambda}_n(t)) = \mu(t)$.

Bubenik (2012) showed that $\overline{\Lambda}_n$ converges pointwise to μ and that the pointwise Central Limit Theorem holds. We will show that

$$\left\{\sqrt{n}\left(\overline{\Lambda}_n(t) - \mu(t)\right)\right\}_{t\in[0,T]}$$

converges weakly to a Gaussian process on $[0,T]$ and we establish the rate of convergence.

Let $\mathcal{F} = \{f_t : \ 0 \le t \le T\}$ where $f_t : \mathcal{L}_T \to \mathbb{R}$ is defined by

$$f_t(\Lambda) = \Lambda(t).$$

Write $P(f) = \int f dP$ and let

P_n be the empirical measure: mass $1/n$ at each Λ_i.

We regard $\sqrt{n}\left(\overline{\Lambda}_n(t) - \mu(t)\right)$ as an empirical process indexed by

$f \in \mathcal{F}$. Thus, for $t \in [0, T]$, we will write

$$\sqrt{n}\left(\overline{\Lambda}_n(t) - \mu(t)\right) = \sqrt{n}(P_n - P)(f_t) = \mathbb{G}_n(t) \ = \ \mathbb{G}_n(f_t)$$

CONVERGENCE

Theorem. Let $\mathbb{G}$ be a Brownian bridge with covariance function

$$\kappa(t,s) = \int f_t(\lambda) f_s(\lambda) dP(\lambda) - \int f_t(\lambda) dP(\lambda) \int f_s(\lambda) dP(\lambda)$$

for $t, s \in [0, T]$. Then $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ (converges in distribution).

Theorem. Let $W \stackrel{d}{=} \sup_t |\mathbb{G}(f_t)|$. Then

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P}\left(\sup_t |\mathbb{G}_n(t)| \le z \right) - \mathbb{P}\left(W \le z\right) \right| = O\left(\frac{(\log n)^{7/8}}{n^{1/8}} \right) \to 0.$$

INFERENCE

Want ℓ_n, u_n such that

$$\mathbb{P}\Big(\ell_n(t) \le \mu(t) \le u_n(t) \text{ for all } t\Big) \ge 1 - \alpha - O(r_n),$$

where $r_n = o(1)$. We use the multiplier bootstrap.

Let $\xi_1^n = (\xi_1, \ldots, \xi_n)$ where $\xi_i \sim N(0,1)$. Define

$$\widetilde{\mathbb{G}}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \left(\Lambda_i(t) - \overline{\Lambda}_n(t)\right)\ ,\ t \in [0,T].$$

Everything is fixed except $\xi_1^n = (\xi_1, \ldots, \xi_n)$ which we generate. Hence, we know (can compute) $\widetilde{Z}(\alpha)$ where

$$\mathbb{P}(\sqrt{n}||\widetilde{\mathbb{G}}_n(t)||_\infty > \widetilde{Z}(\alpha)) = \alpha.$$

INFERENCE

The multiplier bootstrap confidence band is

$$\ell_n(t) = \overline{\Lambda}_n(t) - \frac{Z(\alpha)}{\sqrt{n}}, \quad u_n(t) = \overline{\Lambda}_n(t) + \frac{Z(\alpha)}{\sqrt{n}}.$$

THEOREM. We have

$$\mathbb{P}\Big(\ell_n(t) \leq \mu(t) \leq u_n(t) \text{ for all } t\Big) \geq 1-\alpha-O\left(\frac{(\log n)^{7/8}}{n^{1/8}}\right).$$

Also, $\sup_t (u_n(t) - \ell_n(t)) = O_P\left(\sqrt{\frac{1}{n}}\right)$.

IMPROVEMENT: Variable Width

Let

$$\widehat{\boldsymbol{\sigma}}_n(\boldsymbol{t}) := \sqrt{\frac{1}{n}\sum_{i=1}^{n}[f_t(\boldsymbol{\Lambda}_i)]^2 - [\overline{\boldsymbol{\Lambda}}_n(t))]^2}$$

and

$$\mathbb{H}_n(f_t) := \mathbb{H}_n(\boldsymbol{\Lambda}_1^n)(f_t) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{f_t(\boldsymbol{\Lambda}_i) - \boldsymbol{\mu}(t)}{\boldsymbol{\sigma}(t)}.$$

Multiplier bootstrap version

$$\widehat{\mathbb{H}}_n(f_t) := \widehat{\mathbb{H}}_n(\boldsymbol{\Lambda}_1^n, \boldsymbol{\xi}_1^n)(f_t) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\xi}_i\frac{f_t(\boldsymbol{\Lambda}_i) - \overline{\boldsymbol{\Lambda}}_n(t)}{\widehat{\boldsymbol{\sigma}}_n(t)}.$$

BOOTSTRAP

Let $\widehat{Q}(\alpha)$ be such that

$$\mathbb{P}\left(\sup_t \left|\widehat{\mathbb{H}}_n(\lambda_1^n, \xi_1^n)(f_t)\right| > \widehat{Q}(\alpha) \;\middle|\; \lambda_1, \ldots, \lambda_n\right) = \alpha.$$

The variable width confidence band is

$$\ell_{\sigma_n}(t) = \overline{\Lambda}_n(t) - \frac{\widehat{Q}(\alpha)\widehat{\sigma}_n(t)}{\sqrt{n}}, \quad u_{\sigma_n}(t) = \overline{\Lambda}_n(t) + \frac{\widehat{Q}(\alpha)\widehat{\sigma}_n(t)}{\sqrt{n}}.$$

THEOREM. We have

$$\mathbb{P}\Big(\ell_{\sigma_n}(t) \le \mu(t) \le u_{\sigma_n}(t) \text{ for all } t\Big) \ge 1 - \alpha - O\left(\frac{(\log n)^{1/2}}{n^{1/8}}\right).$$

BEYOND LANDSCAPES

The landscape is just one of many functions that could be used to summarize persistence.

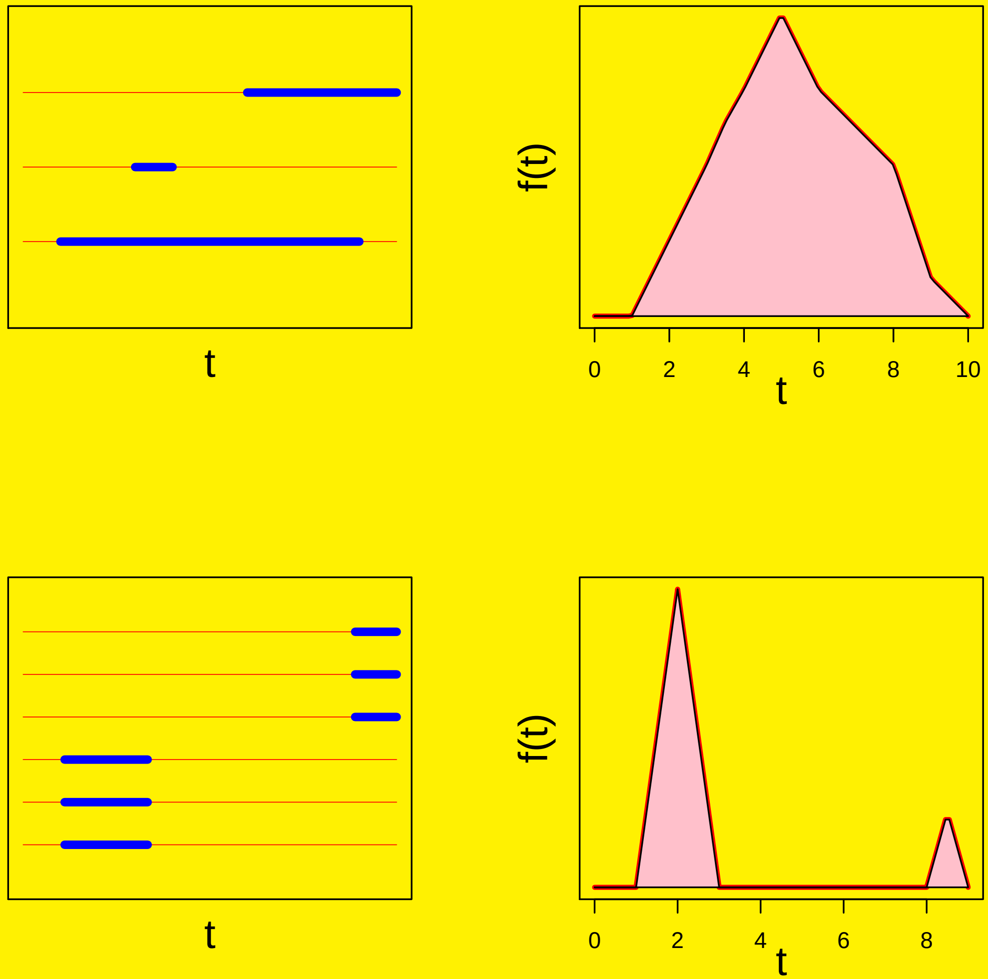
For $0 < p \leq \infty$, we define the **Power-Weighted Silhouette**

$$\phi_p(t) = \frac{\sum_{j=1}^n |b_j - a_j|^p \, T_{(a_j,b_j)}(t)}{\sum_{j=1}^n |b_j - a_j|^p}.$$

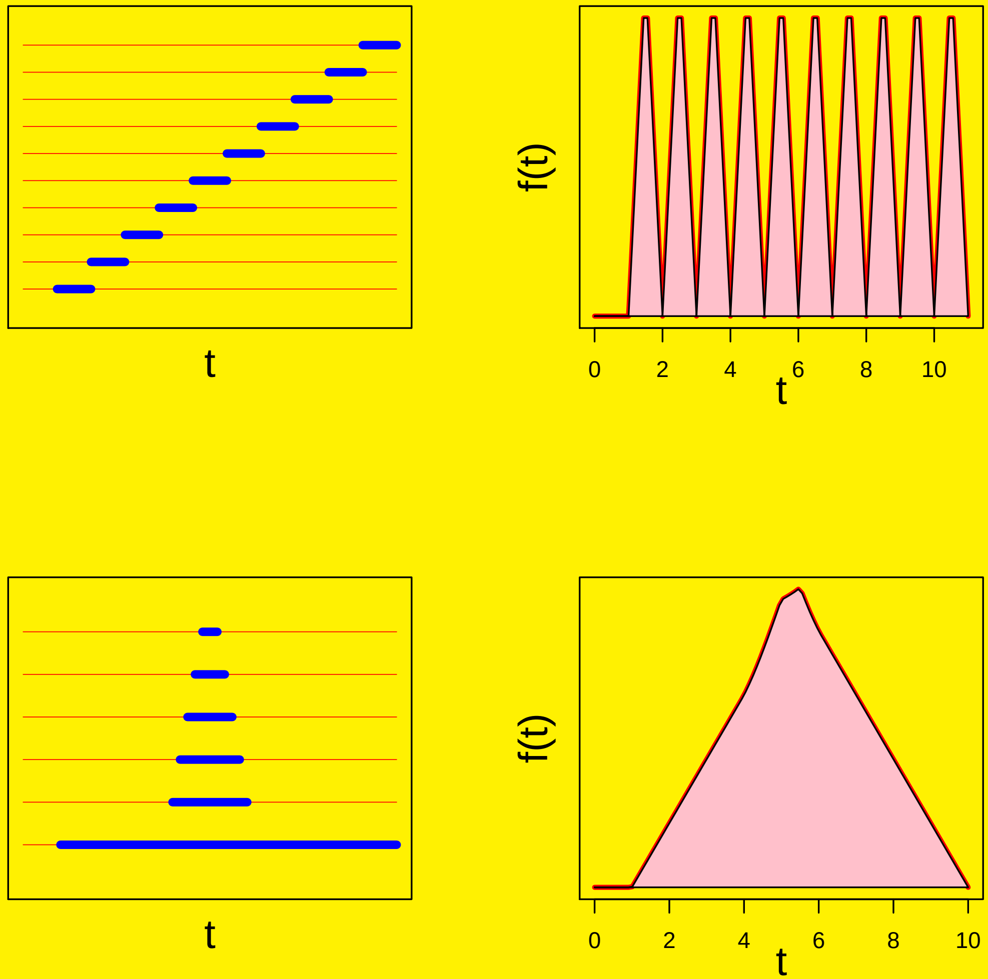
p small: $\phi_p(t)$ is dominated by small barcodes.

p large: $\phi_p(t)$ is dominated by large barcodes.

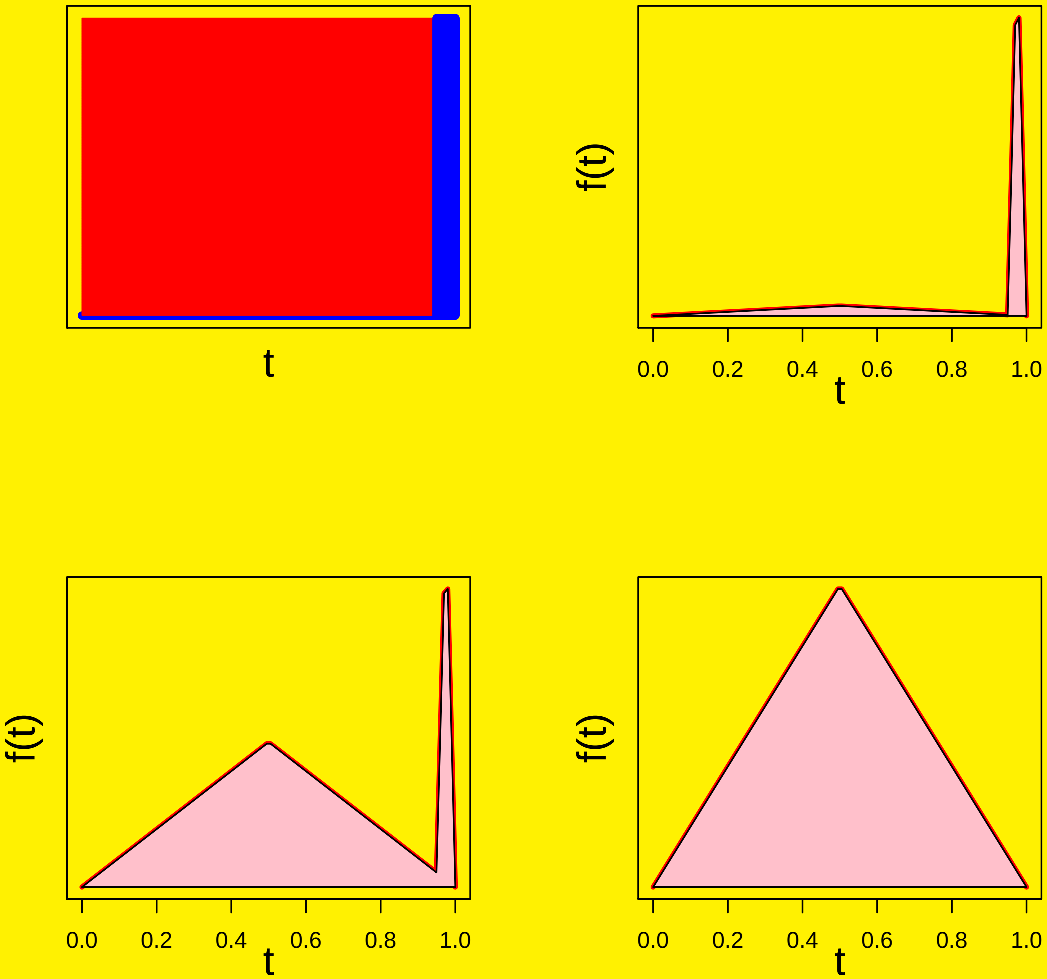
$$\boldsymbol{p = 1}$$

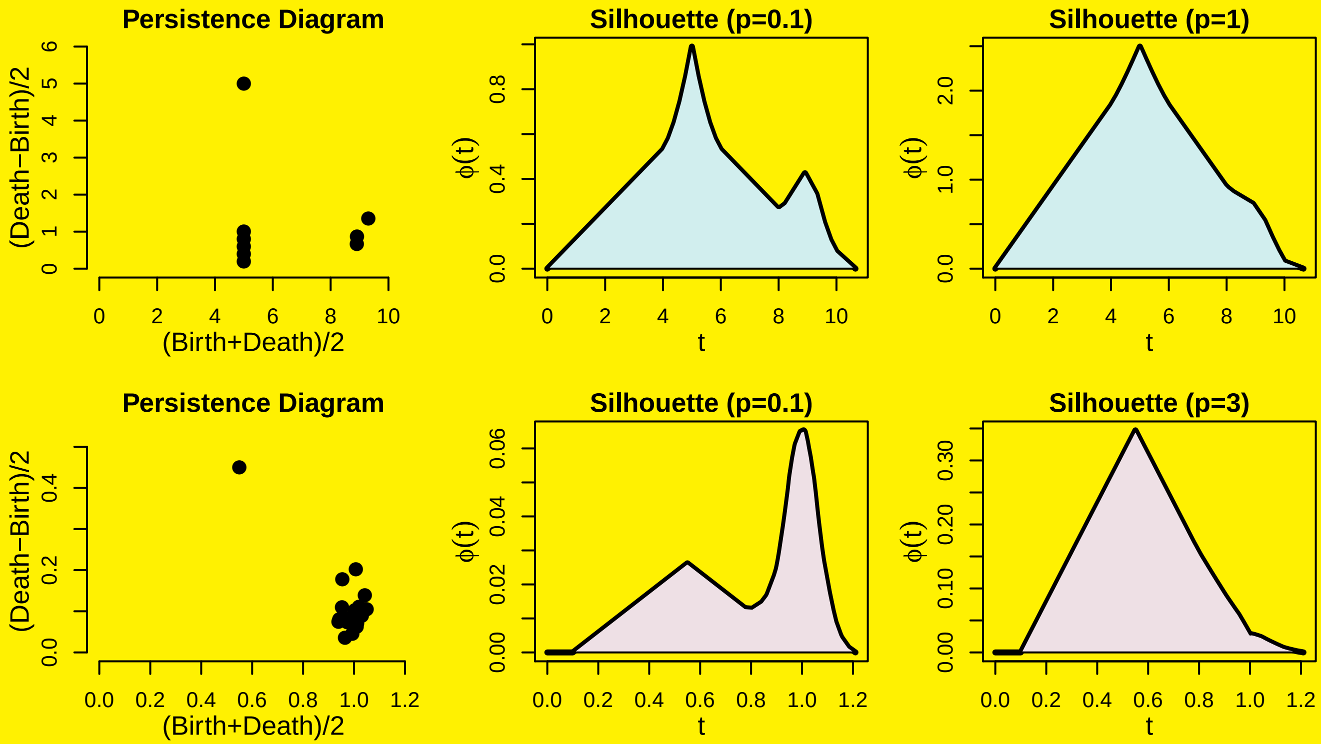


$\boldsymbol{p = 1}$



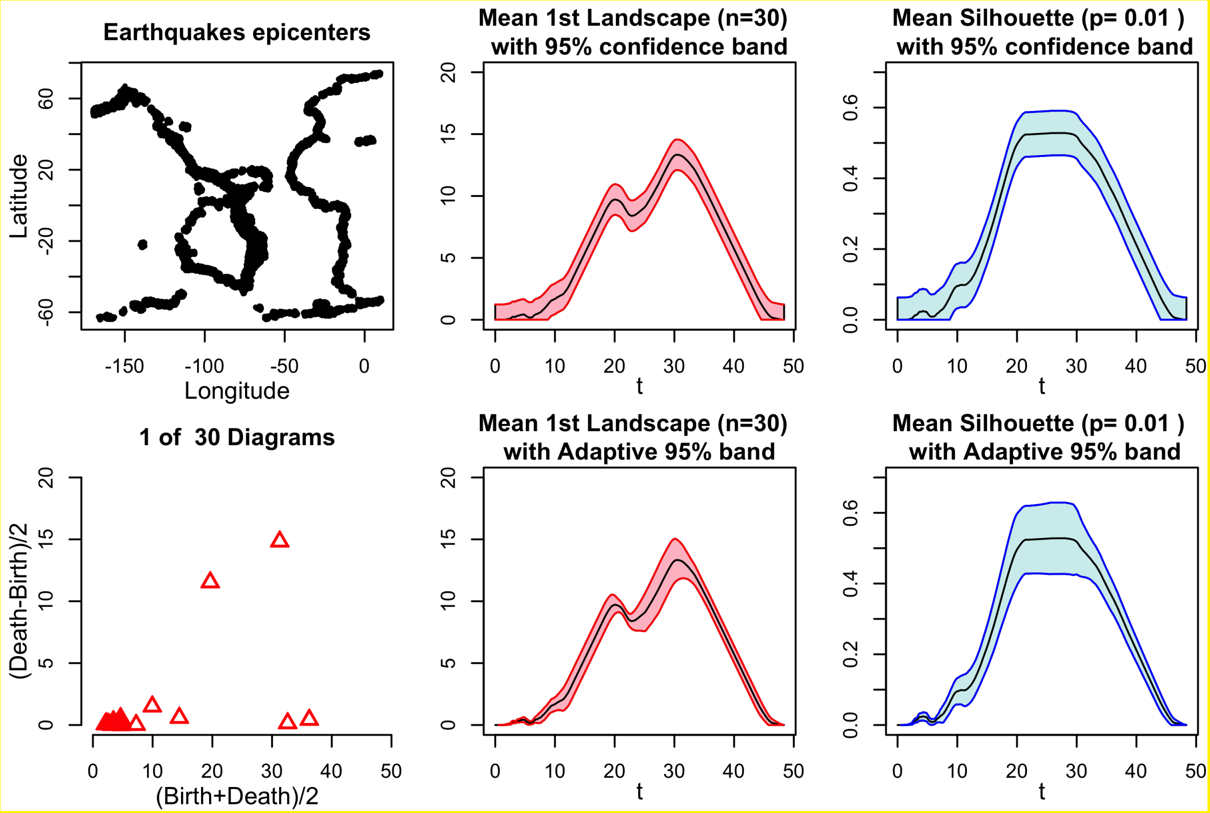
$$p = 0.1,\ 1.0, 10.0$$





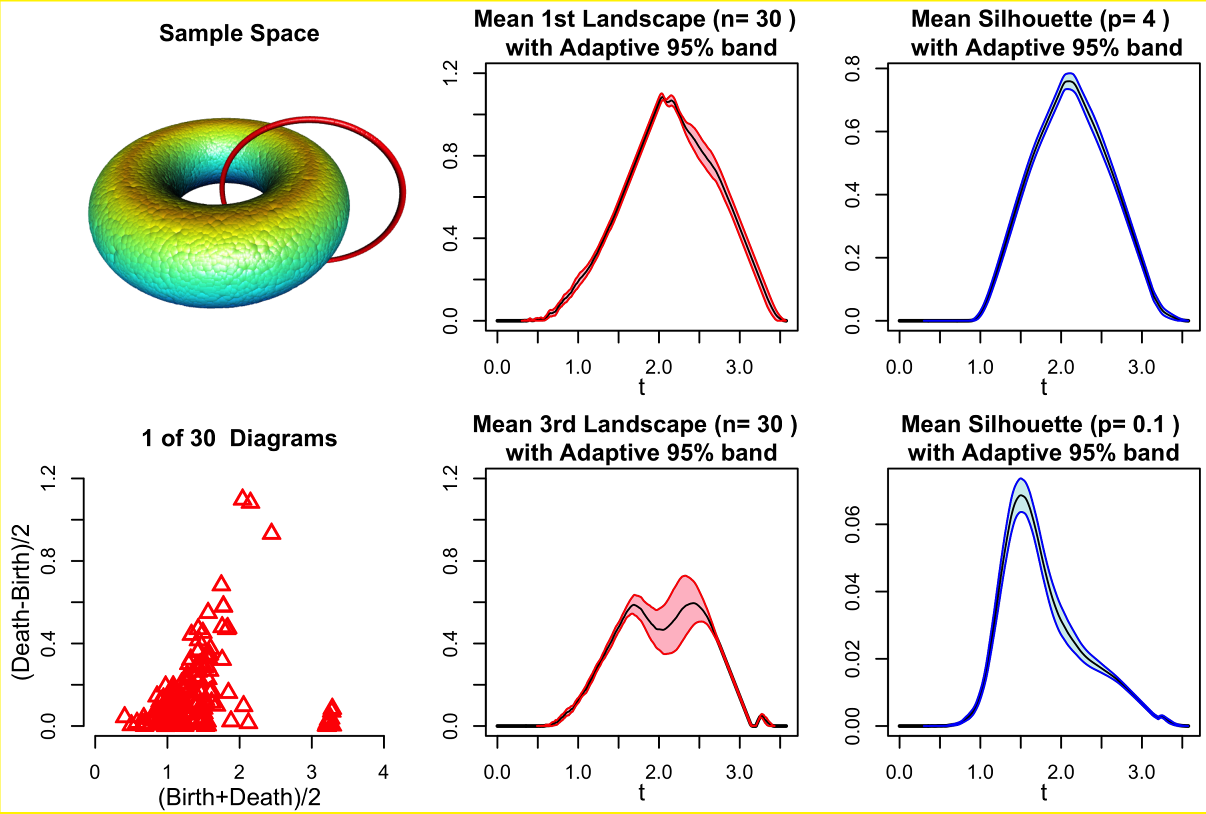
Persistence Diagram
Silhouette (p=0.1)
Silhouette (p=1)
Persistence Diagram
Silhouette (p=0.1)
Silhouette (p=3)
(Death−Birth)/2
(Birth+Death)/2
φ(t)
t

Earthquake data ($N = 8000$, Rips filtration, β_1, Dionysus program (Dmitriy Morozov))



sample $m = 400$ epicenters, 30 times.

Torus + circle. $N = 11,800$ points.



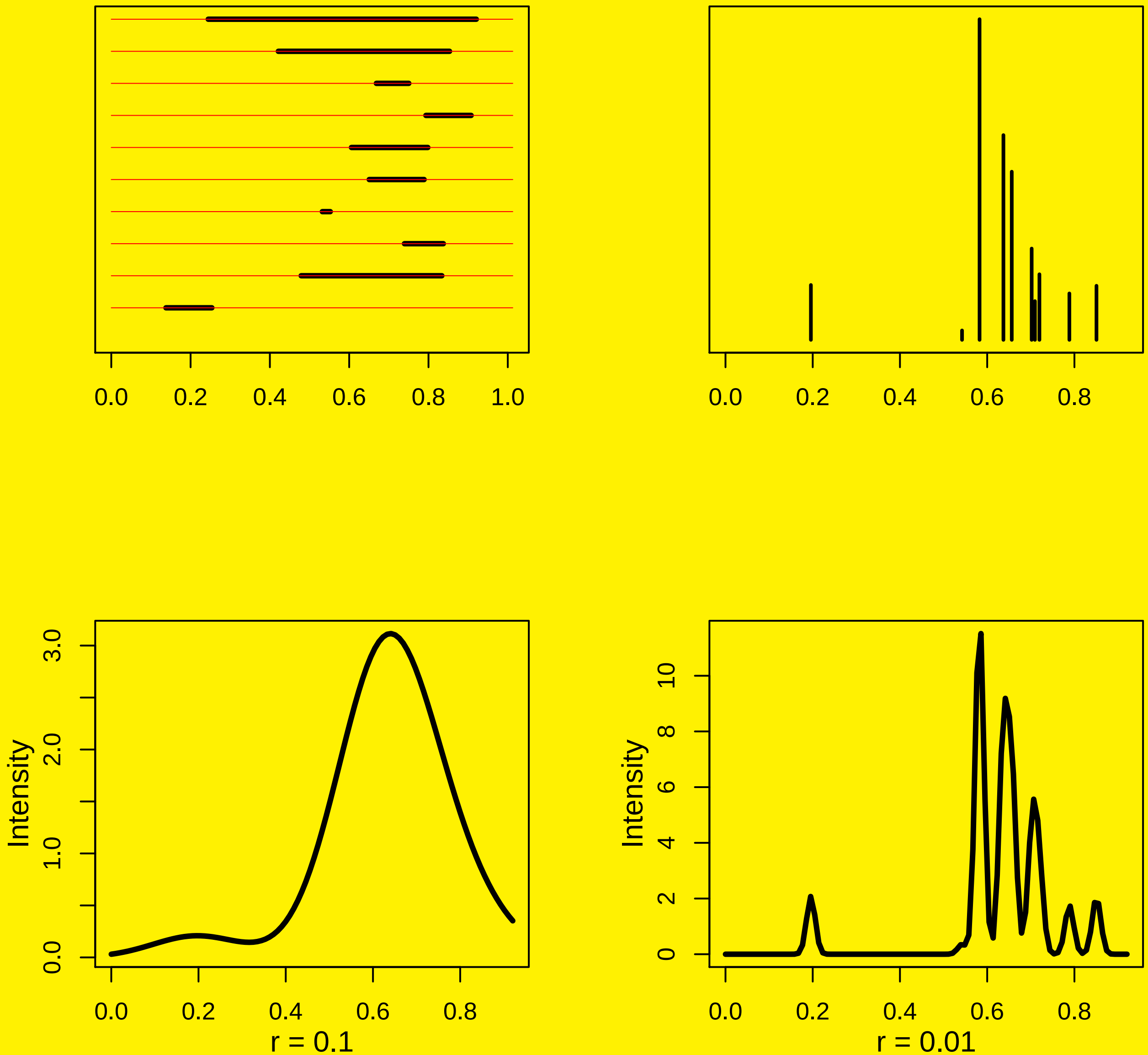
Barcode Intensity Function

(1) Turn barcode sideways, (2) drop onto the axis, (3) smooth. Equivalently: collapase the landscape triangles:

$$\iota_r(t) = \sum_j \pi_j \frac{1}{r} K\left(\frac{t-\delta_j}{r}\right)$$

$r > 0$ is a bandwidth, K is a kernel,
$\pi_j =$ normalized lifetime
δ_j is a point mass at $(b_j + d_j)/2$.

"Bias-Variance" tradeoff:
small r: low bias, but large confidence band
large r: high bias (obscures detail) but narrow band.


0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
Intensity
0.0
1.0
2.0
3.0
0.0
0.2
0.4
0.6
0.8
r = 0.1
Intensity
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
r = 0.01

Persistence Intensity Function (Weygaert, Edelsbrunner, Pranav et al.)

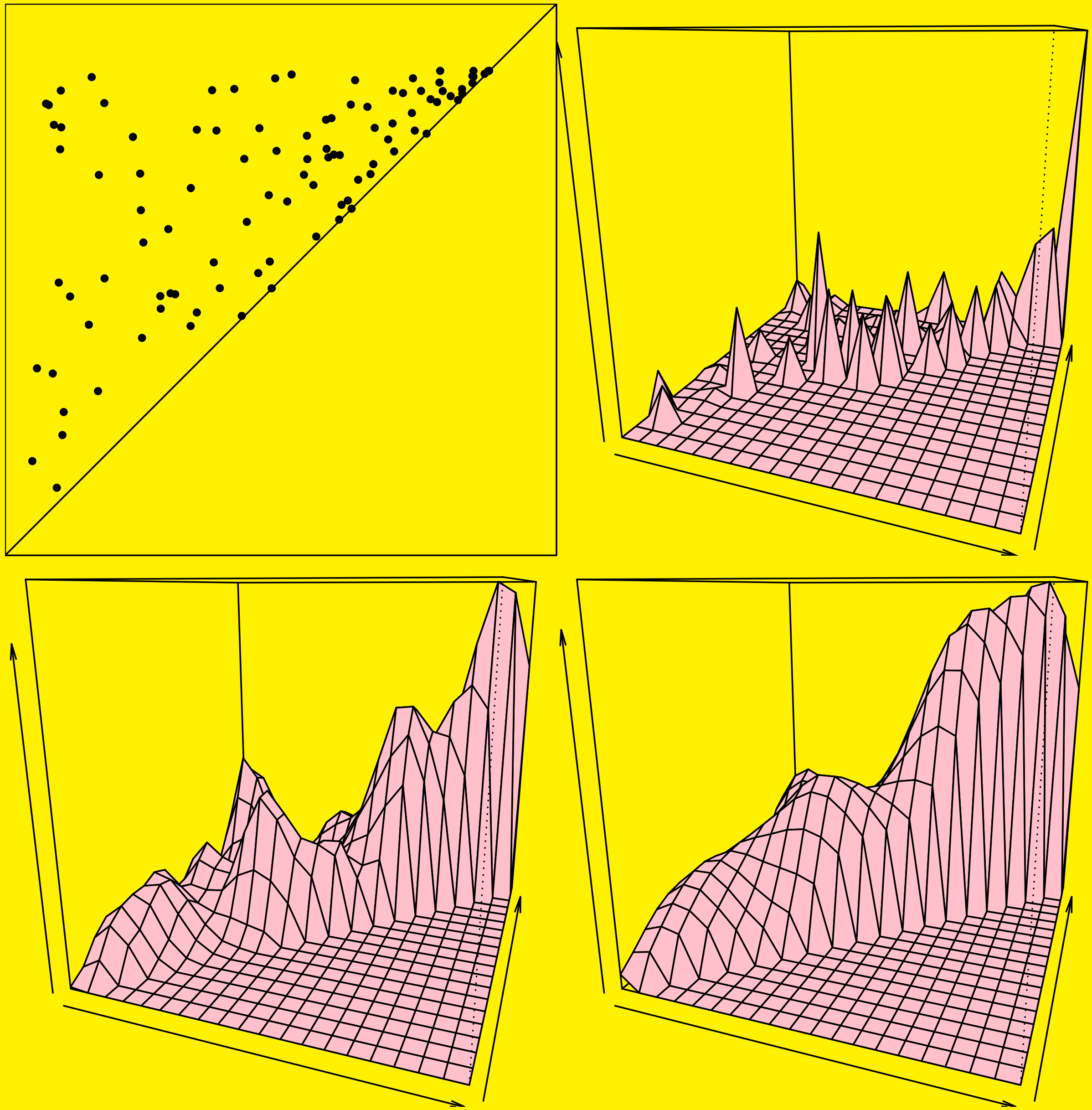
Treat points $z_j = (b_j, d_j)$ in persistence diagram as apoint process then smooth it. They use a histogram but we can use a kernel:

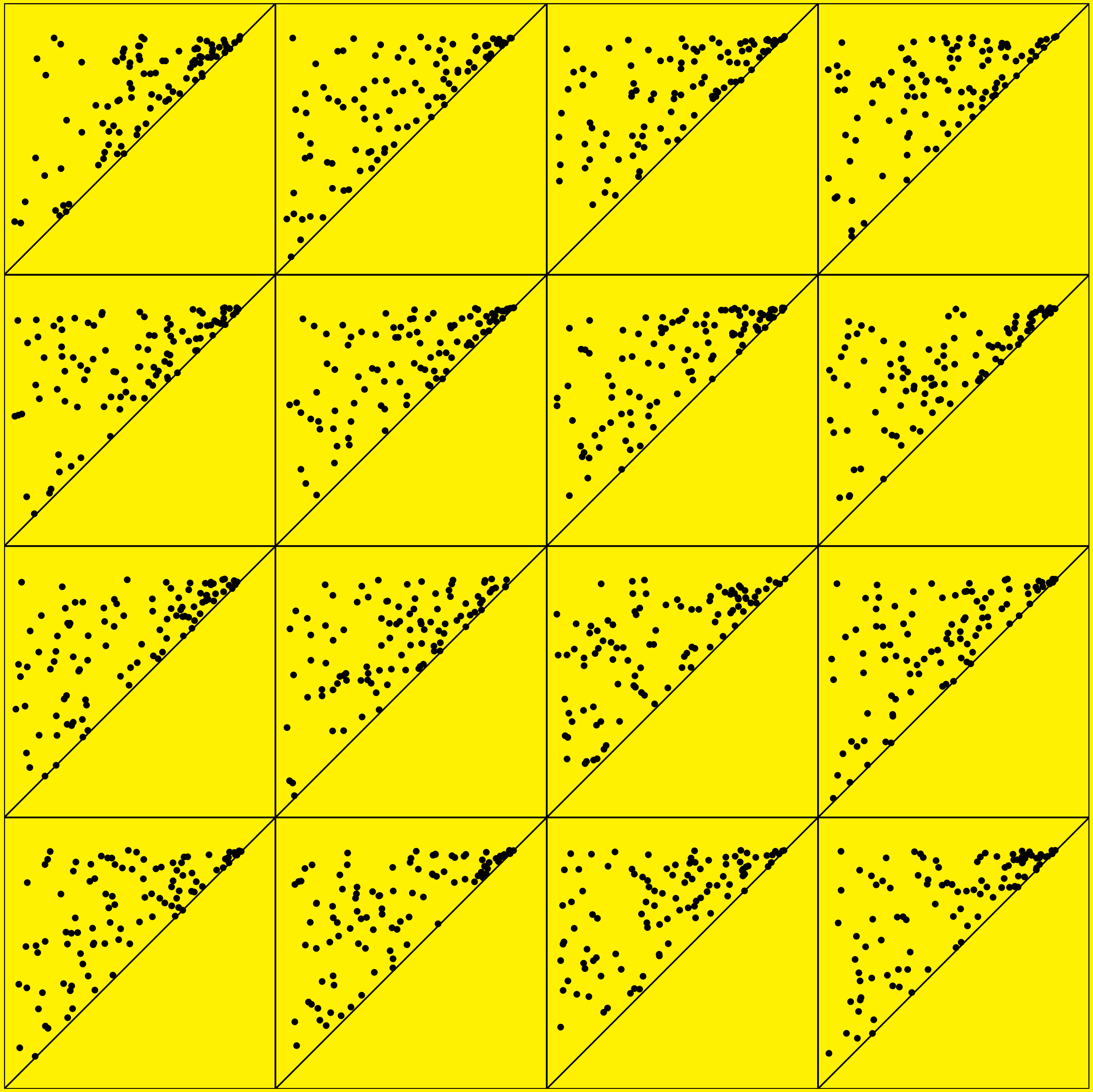
$$\iota_r(t) = \frac{1}{m} \sum_j \frac{1}{r^2} K\left(\frac{||t - z_j||}{r}\right).$$

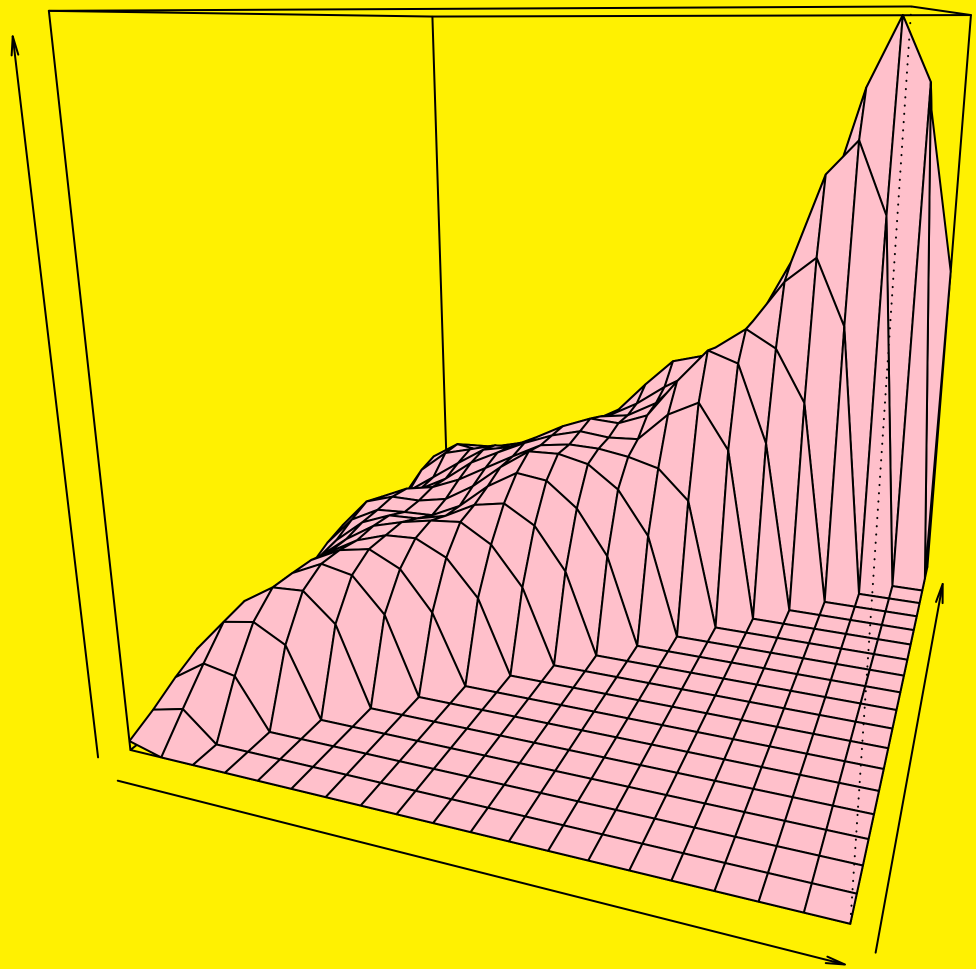
Again, there is a quasi bias-variance tradeoff.

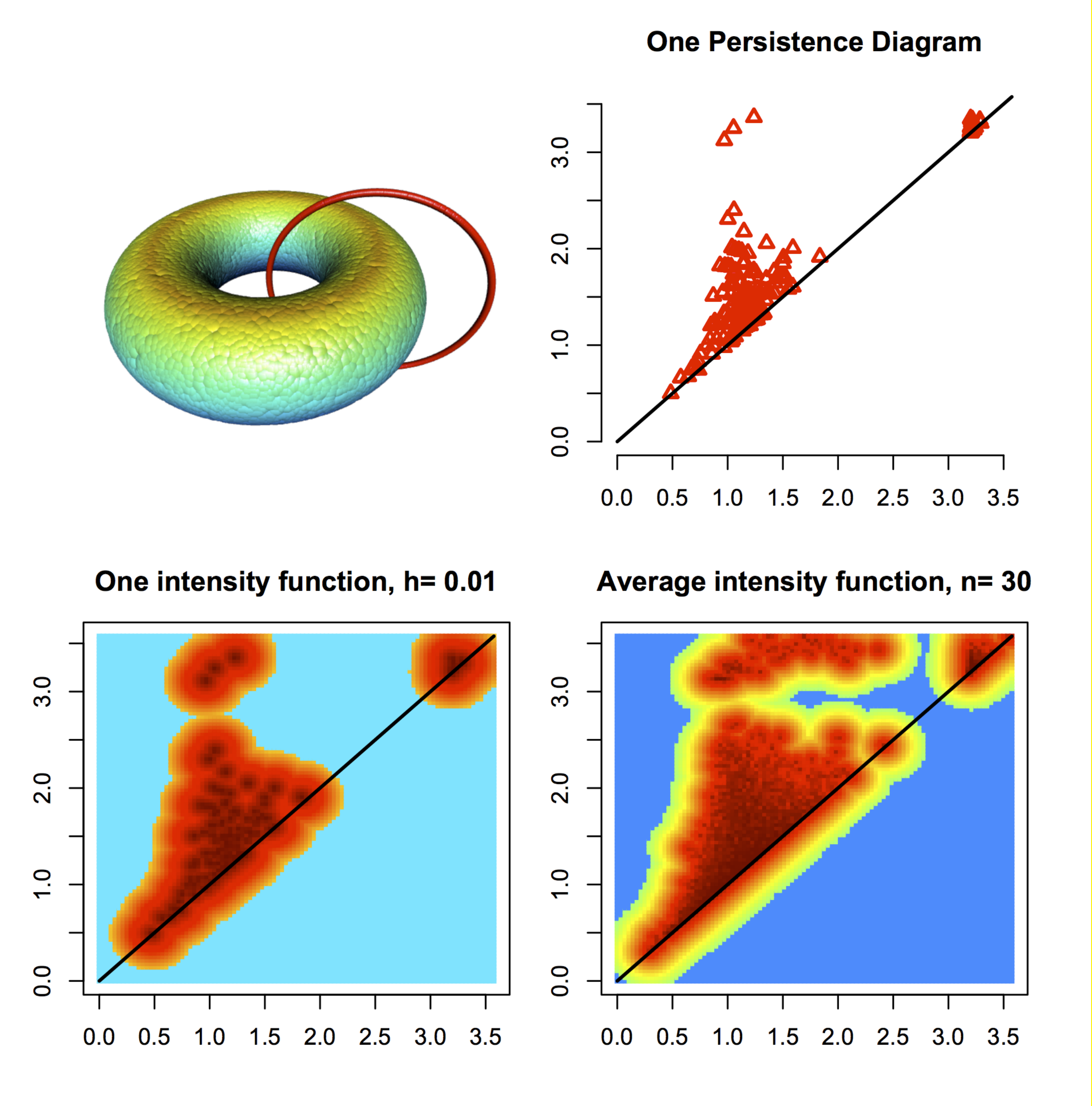
For many diagrams $D_1, \ldots, D_n$ we can simply average

$$\iota(t) = \frac{1}{n} \sum_{i=1}^{n} \iota_i(t).$$









One Persistence Diagram
0.0
1.0
2.0
3.0
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
One intensity function, h= 0.01
Average intensity function, n= 30

Work in progress:

1. Optimal choice of r.
2. Different r for each diagram.
3. Spatially varying r.
4. Convergence theory.
5. Bootstrap.
6. Invertibility.

Bias? Note that when $r = 0$ we can recover D. When $r > 0$, the map $D \to \iota$ is (apparently) not invertible. The "bias" should be related to the modulus of continuity:

$$m_r(\epsilon) = \sup\Big\{ d_\infty(D, D') : \ ||\iota_r(D) - \iota_r(D')||_\infty \leq \epsilon \Big\}.$$

We can estimate $m_r'(0)$.

Meta Persistent Homology

Given summary functions

$$F_1, \ldots, F_n \sim P$$

why should we summarize them with their mean?

Perhaps we should look for clusters in P. Now P does not have a density but it does have a pseudo-density

$$p_\epsilon(f) = \mathbb{P}(N_\epsilon(f))$$

where

$$N_\epsilon(f) = \{g : \ d(f, g) \leq \epsilon\}.$$

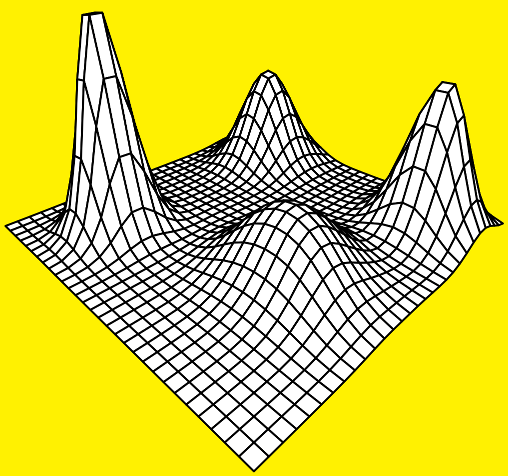
Estimate

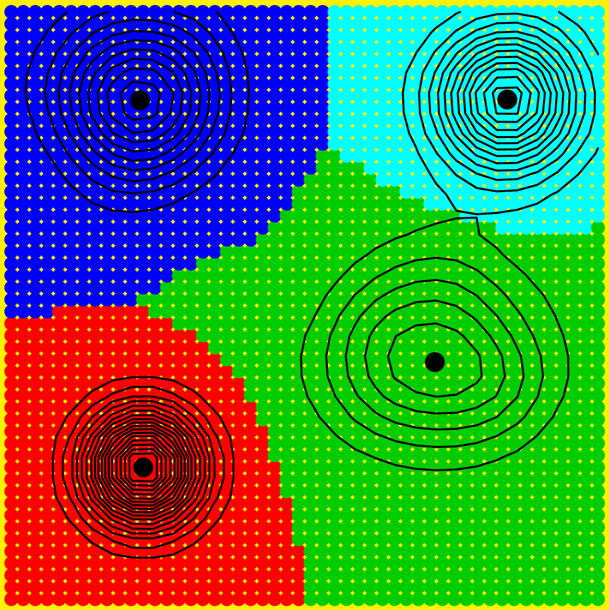
$$\widehat{p}_\epsilon(f) = \frac{1}{n} \sum_{i=1}^{n} I(F_i \in N_\epsilon(f)).$$

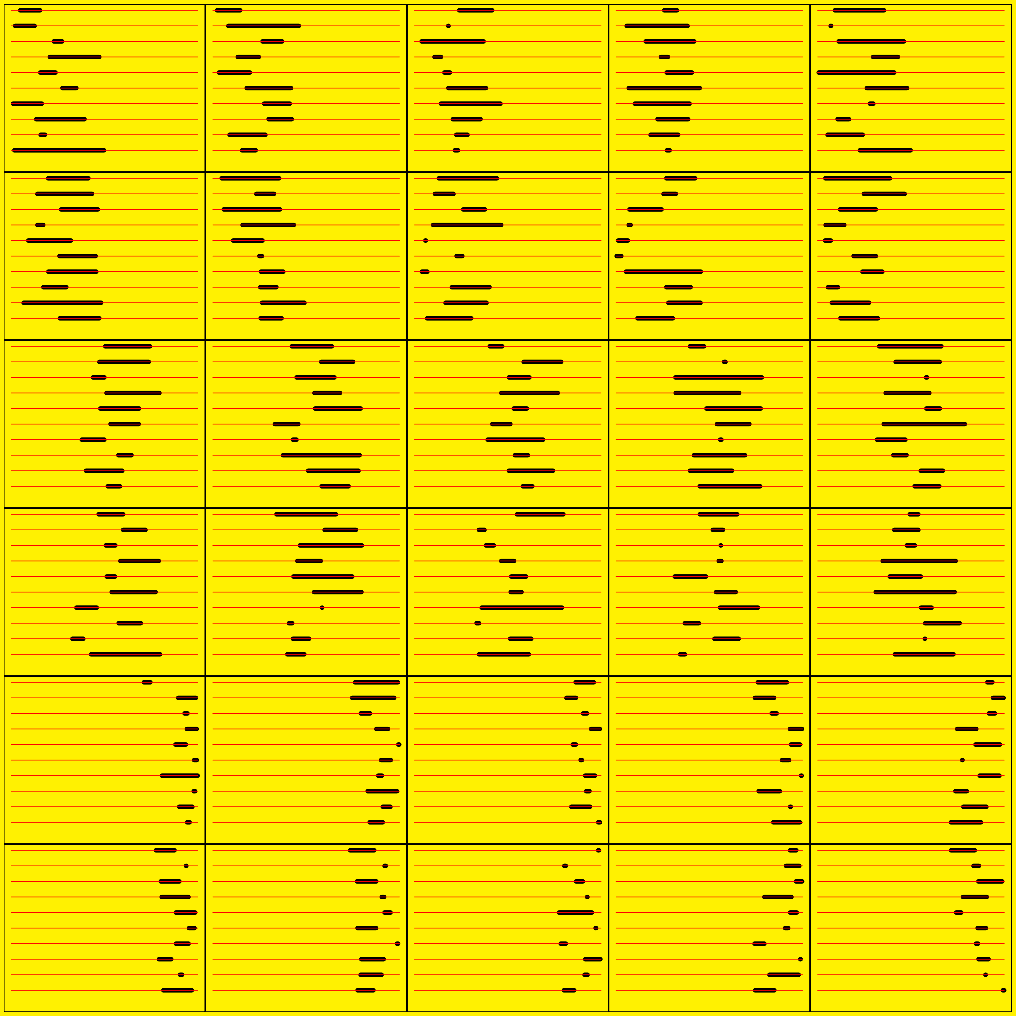
Now apply mode clustering (Morse clustering) to $\widehat{p}_\epsilon$.

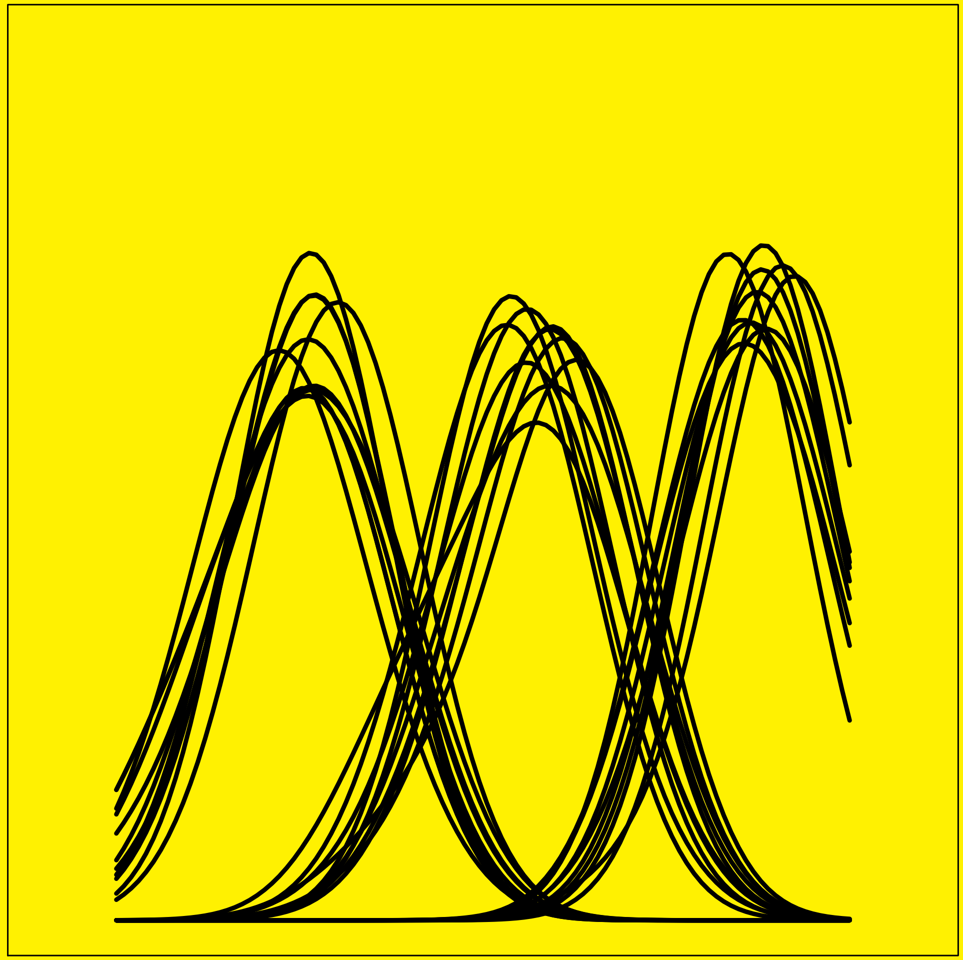
Locate the modes of $\widehat{p}_\epsilon$ using the mean-shift algorithm.

Each mode $\widehat{m}_j$ has a lifetime and basin of attraction which defines the clusters. (Chacon arxiv:1212.1385, Chazal, Guibas, Oudot and Skraba 2011).

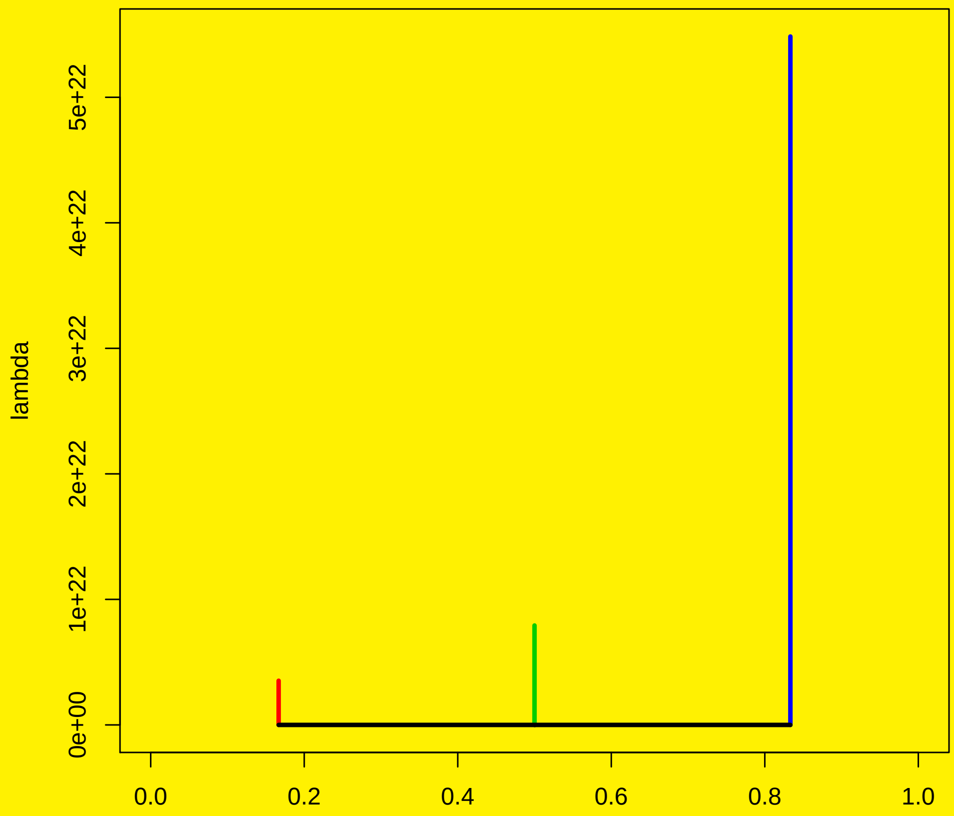








Met-Persistent Homology of the modes in function space (using DeBaClR: Brian Kent, Fabrizio Lecci).



THE TYRANNY OF TUNING PARAMETERS

(Warning: this is work in progress.)

Let $X_1, \ldots, X_n \sim G$ supported on K. Add noise:

$$Y_i = X_i + \epsilon_i$$

where $\epsilon \sim \Phi$. Add clutter: Let $U_1, \ldots, U_n \sim Q$.

$$Z_i = \begin{cases} Y_i & \text{with prob } \pi \\ U_i & \text{with prob } 1 - \pi. \end{cases}$$

Distribution of Z is $P = (1-\pi)Q + \pi(G \star \Phi)$ with density

$$p(z) = (1-\pi)q(z) + \pi \int \phi(z-u)dG(u).$$

p is concetrated near K and the persistent homology of the upper level sets is of interest. (See Fabrizio's talk later this week.)

Kernel density estimator:

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K\left(\frac{||x - X_i||}{h}\right).$$

K is any kernel (example: Gaussian).

$h > 0$, the bandwidth, is crucial.

How to choose h?

Usual method in statistics: cross-validation. No good for TDA.

(Similarly, distance-to-a-measure (Chazal, Cohen-Steiner and Merigot 2011) has a smoothing parameter m_0.)

FAILURE OF CROSS-VALIDATION FOR TDA

Cross-validation: Minimize

$$\int (\widehat{p}_h(x) - p(x))^2 dx = J(h) + \text{constant}$$

where

$$\begin{aligned} J(h) &= \int \widehat{p}_h^2(x)dx - 2\int \widehat{p}_h(x)p(x)dx \\ &\approx \int \widehat{p}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{k}\widehat{p}_h(Z_i) \quad \text{held out data} \\ &= \widehat{J}(h) \end{aligned}$$

and we minimize $\widehat{J}(h)$ over h.

But L_2 is the wrong loss function for TDA.

FAILURE OF CROSS-VALIDATION FOR TDA

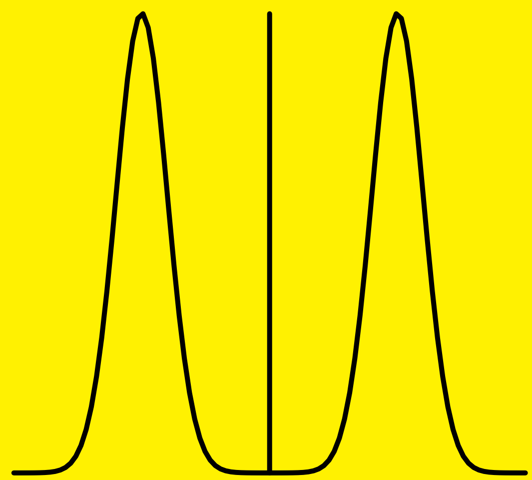
But in TDA, p might be singular or nearly singular. Consider

$$P = \frac{1}{3}N(-5,1) + \frac{1}{3}\delta_0 + \frac{1}{3}N(5,1)$$

where δ_0 is a point mass at 0.

P doesn't have a density but p_h does, where

$$p_h(x) = \mathbb{E}[\widehat{p}_h(x)] = \frac{d}{dx}(P \star K_h).$$

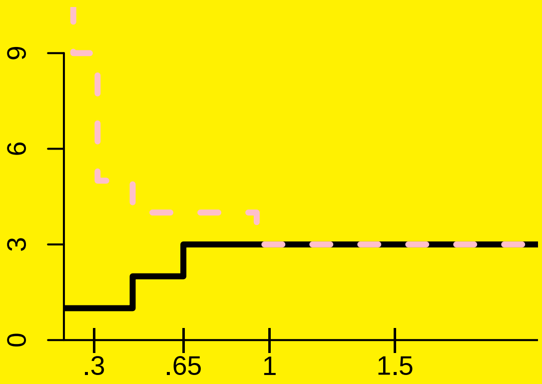


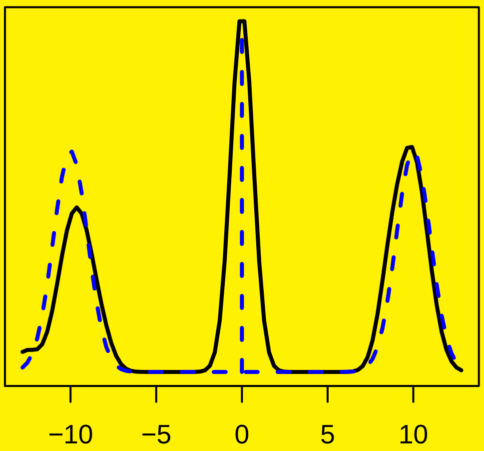
FAILURE OF CROSS-VALIDATION FOR TDA

Cross-validation gives $h = 0$ which is useless.

Genovese, Perone-Pacifico, Verdinelli and Wasserman (2013, arxiv:1312.7567) proposed the following:

-compute $\widehat{p}_h$ fo each h
-find modes
-test significance of modes
-chose h to maximize number of significant modes





Methods for choosing h (and other tuning parameters) in TDA.

(1) MTSS (Maximum Significant Topological Signal Strength)
Choose h to maximize significant topological signal:

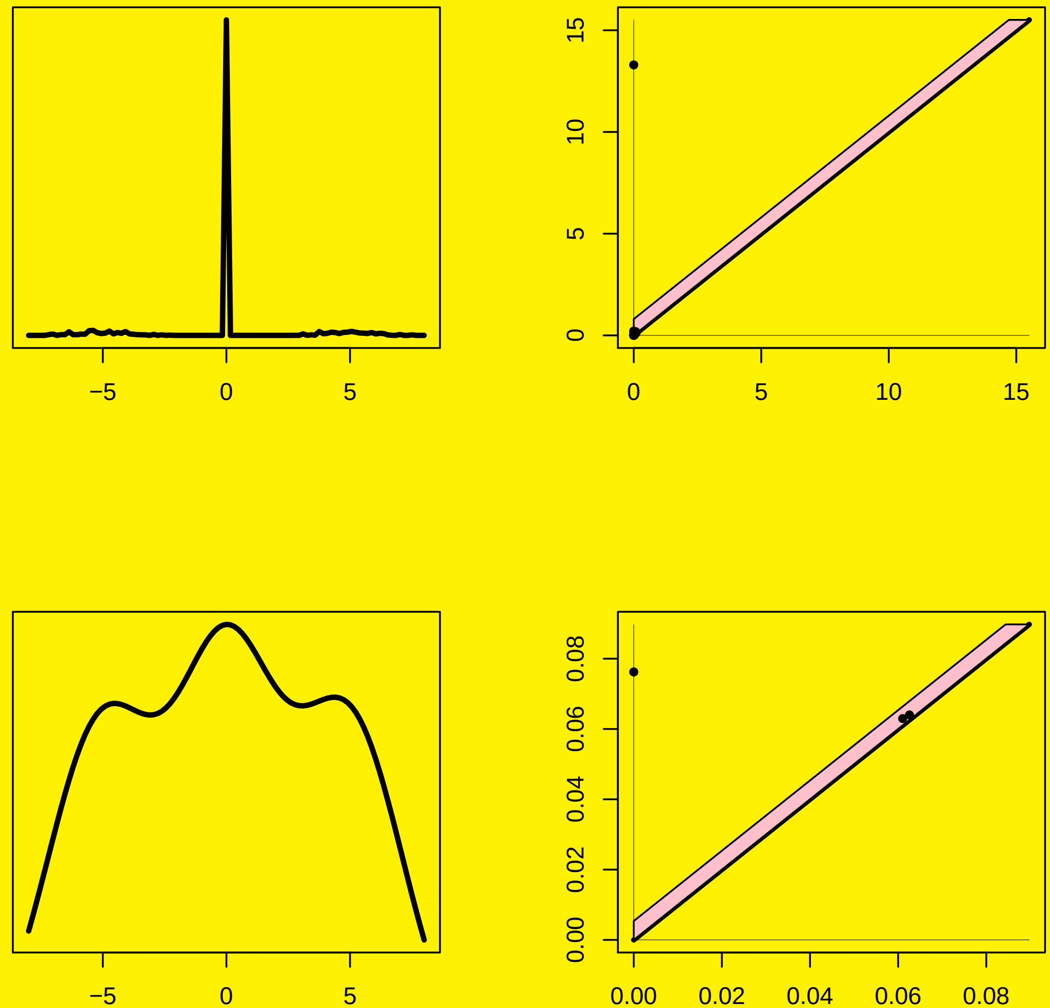
$$\xi(h) = \sum_j I(d_j - b_j > \epsilon(h))$$

where $\epsilon(h)$ comes from the bootstrap (Fabrizio's talk).

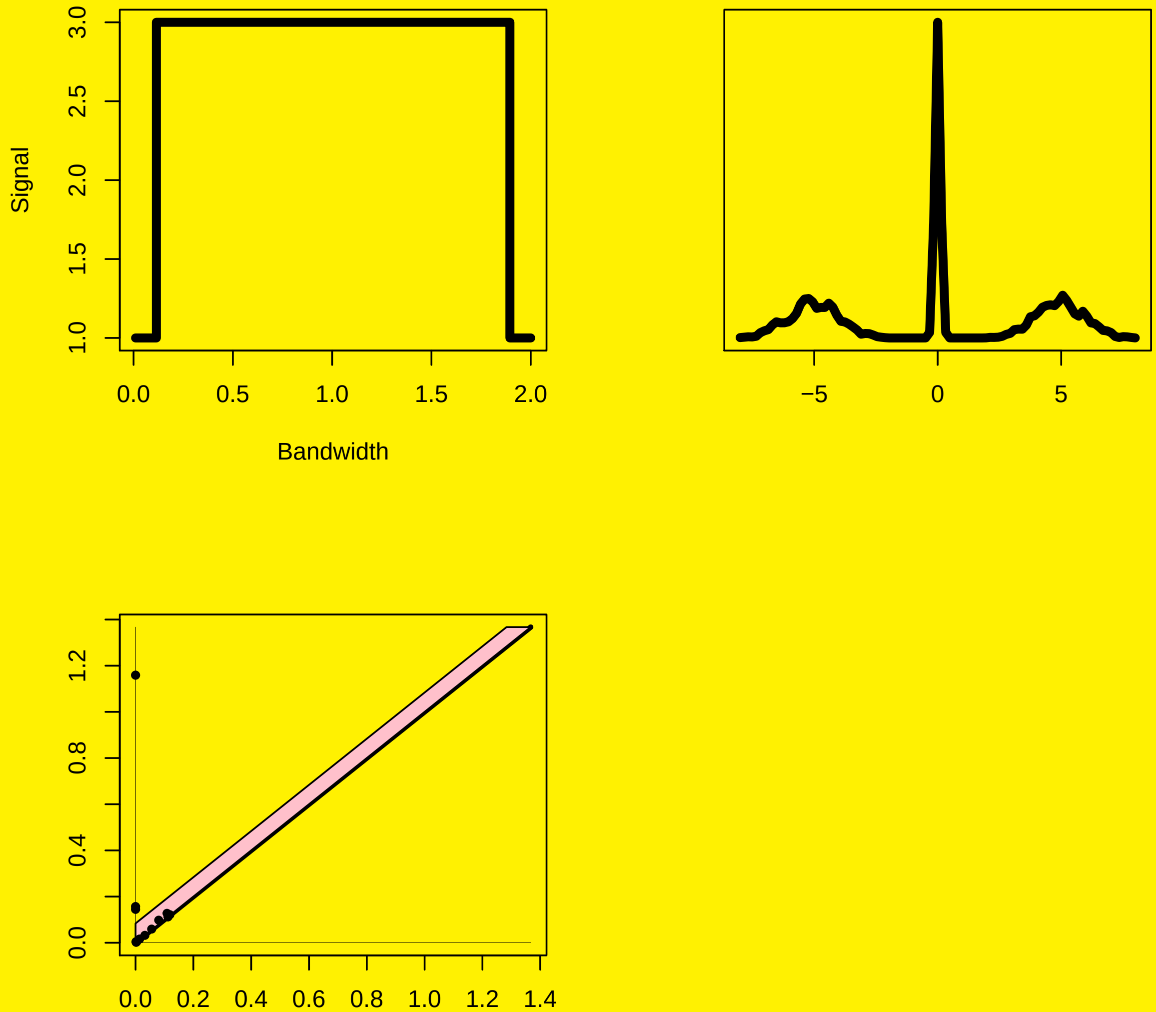
$\xi(h) = 0$ for small h and large h.

Example: mixture with singular component again ...

Small h and Large h:



Maximum significant topological signal:



(2) **Density Diversity** (adapted from an idea in Ferraty and Vieu 2000).

Let $Z = (Z_1, \ldots, Z_n)$ where

$$Z_i = \frac{1}{\widehat{p}_h(X_i)}.$$

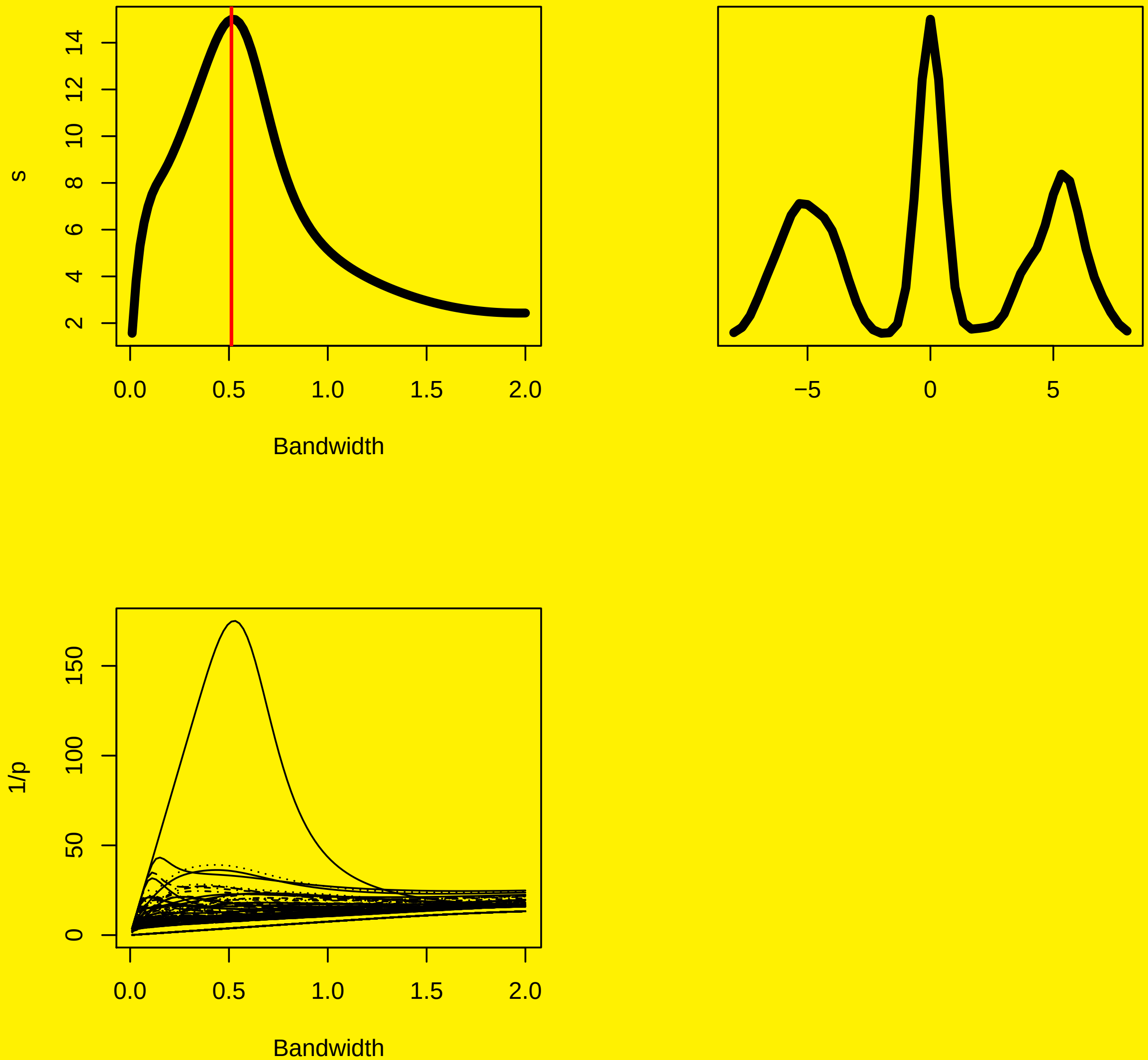
Let

$$S(h) = \text{empirical standard deviation of } Z_1, \ldots, Z_n.$$

$h = 0$ implies $S(h) = 0$.

$h = \infty$ implies $S(h) = 0$.

Choose $\widehat{h}$ to maximimize $S(h)$.


s
2
4
6
8
10
12
14
0.0
0.5
1.0
1.5
2.0
Bandwidth
−5
0
5
1/p
0
50
100
150
0.0
0.5
1.0
1.5
2.0
Bandwidth

(3) SKI-BOOT. LepSKI with BOOTstrap.

Oleg Lepski (and co-authors) have a series of papers on selecting tuning parameters. See, especially, arXiv:1210.7078. Essentially, it works like this.

1. Start with large h.
2. Test: is there a bandwidth $t < h$ with a significantly different fit?
$$T(h) = \sup_{t<h} \frac{||\widehat{p}_h - \widehat{p}_t||_\infty}{\widehat{\sigma}(h,t)}.$$
3. If $T(h)$ is big, reduce h and repeat. Else, stop.

The details of the procedure are actually very complicated and perhaps not practical. We are working on a bootstrap version:
$$\widehat{\sigma}(h,t) = \mathbb{E}_h||\widehat{p}_t^* - \widehat{p}_t||_\infty.$$

DTM

We can apply similar ideas to distance-to-a-measure (DTM). (Chazal, Cohen-Steiner, Merigot 2011).

$$\widehat{d}_\beta(x) = \frac{1}{k}\sum_{i=1}^{k} ||X_x(i) - x||^2$$

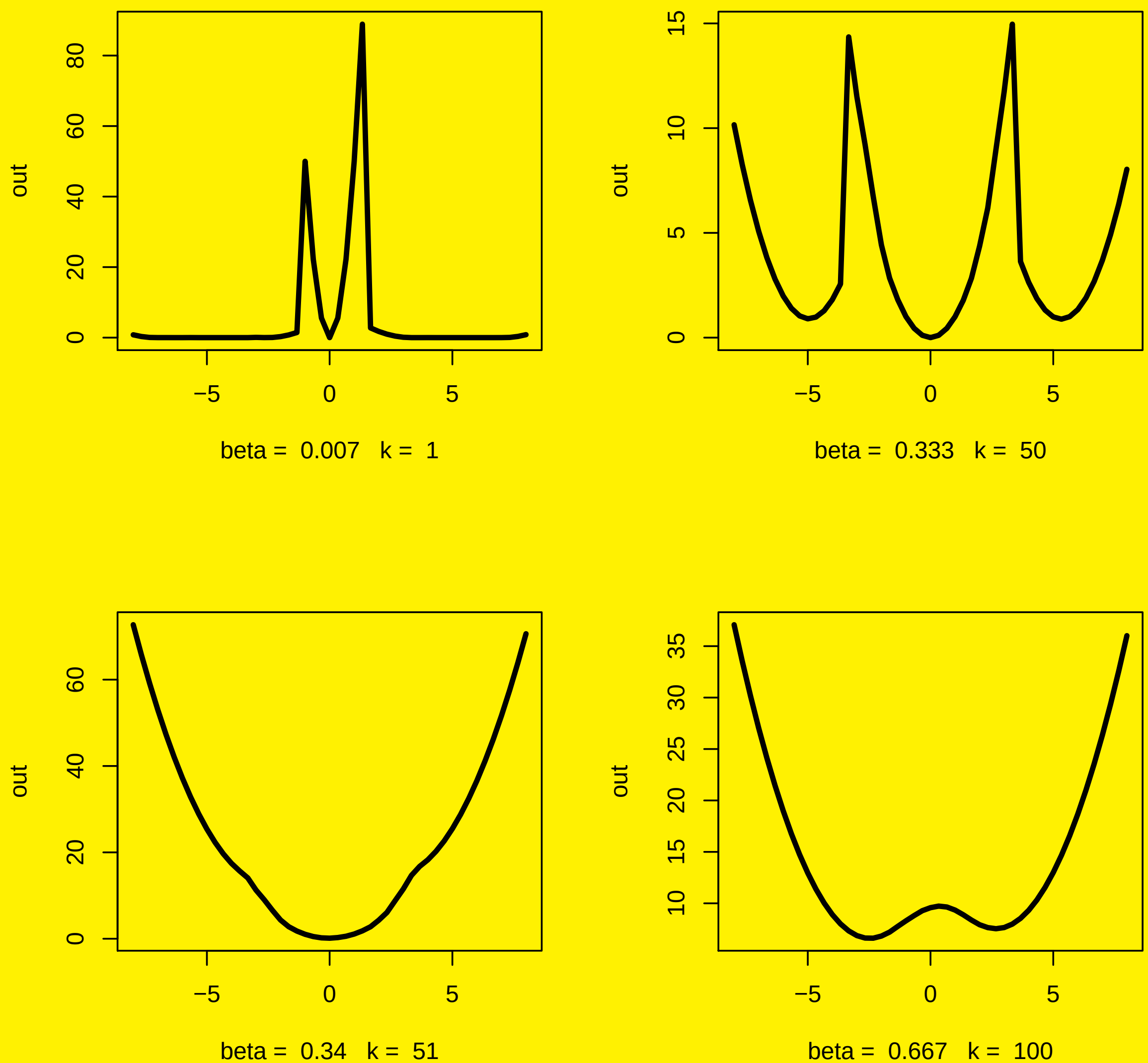
where $k = \beta n$. Here, $0 < \beta < 1$ is the bandwidth.

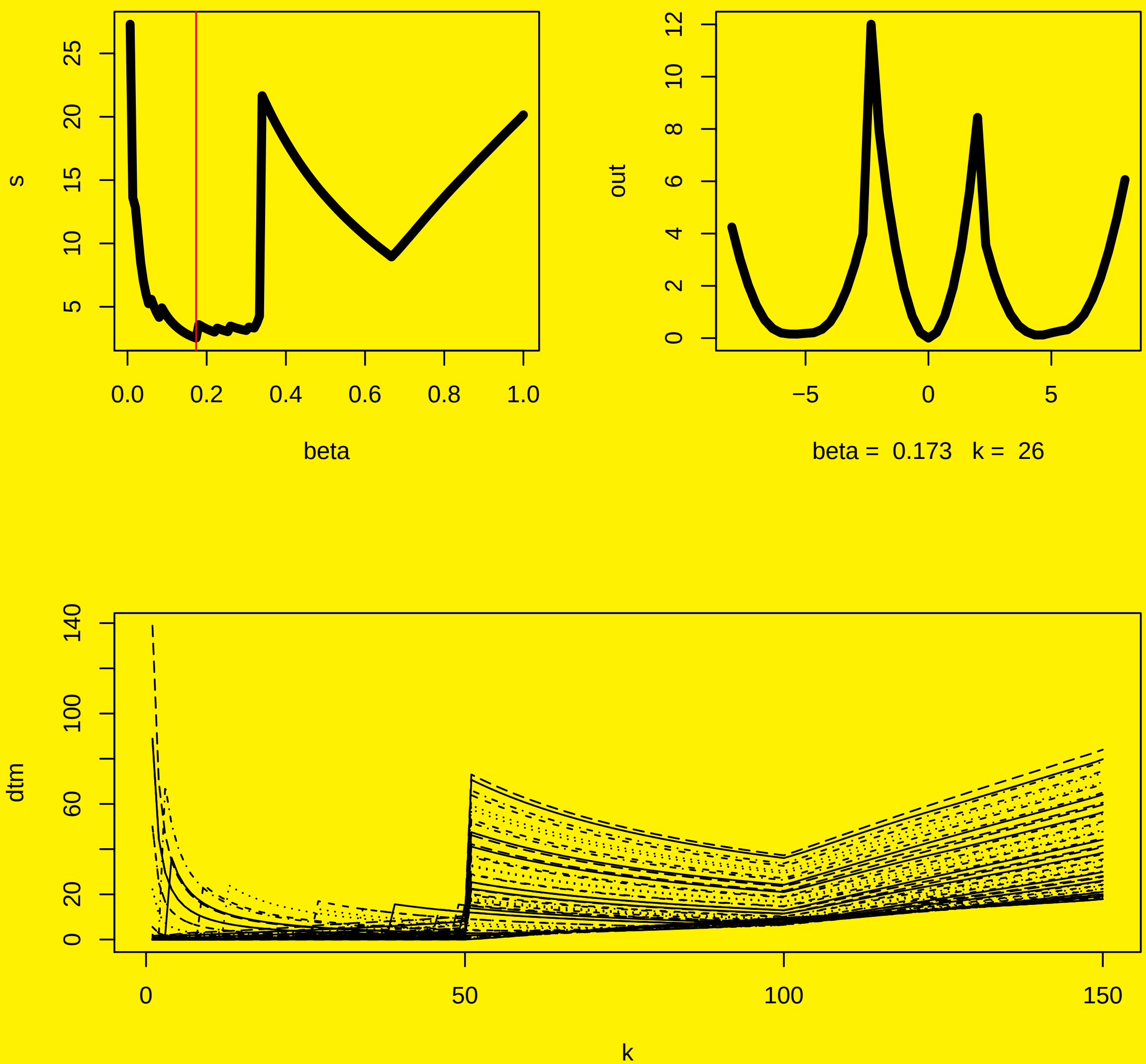
Let K be support and let d_K be distance function. Then

$$||d_K - \widehat{d}_\beta||_\infty \le ||d_K - d_\beta||_\infty + ||d_\beta - \widehat{d}_\beta||_\infty.$$

Here we use a minimum modified diversity:

$$s(\beta) = \int \widehat{(}d_\beta(x) - \overline{d})^2 dx.$$


out
beta = 0.007 k = 1
out
beta = 0.333 k = 50
out
beta = 0.34 k = 51
out
beta = 0.667 k = 100



s
beta
out
beta = 0.173 k = 26
dtm
k

CONCLUSION

1. Functional summaries: very useful. Still working on intensity functions.
2. Tuning parameters: this is very important and unsolved.
3. We should really be using locally adaptive tuning parameters which is even harder.

THE END