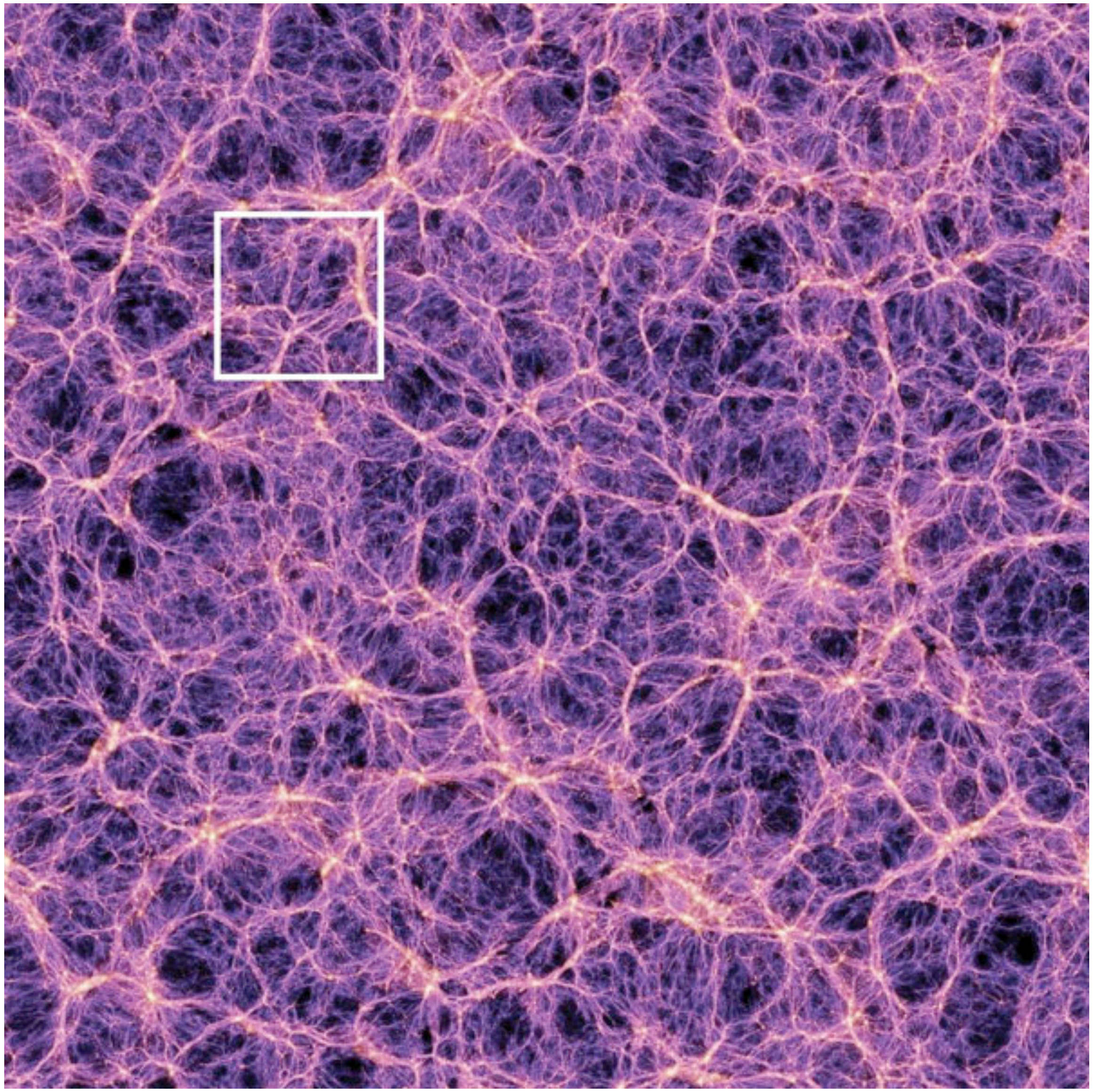
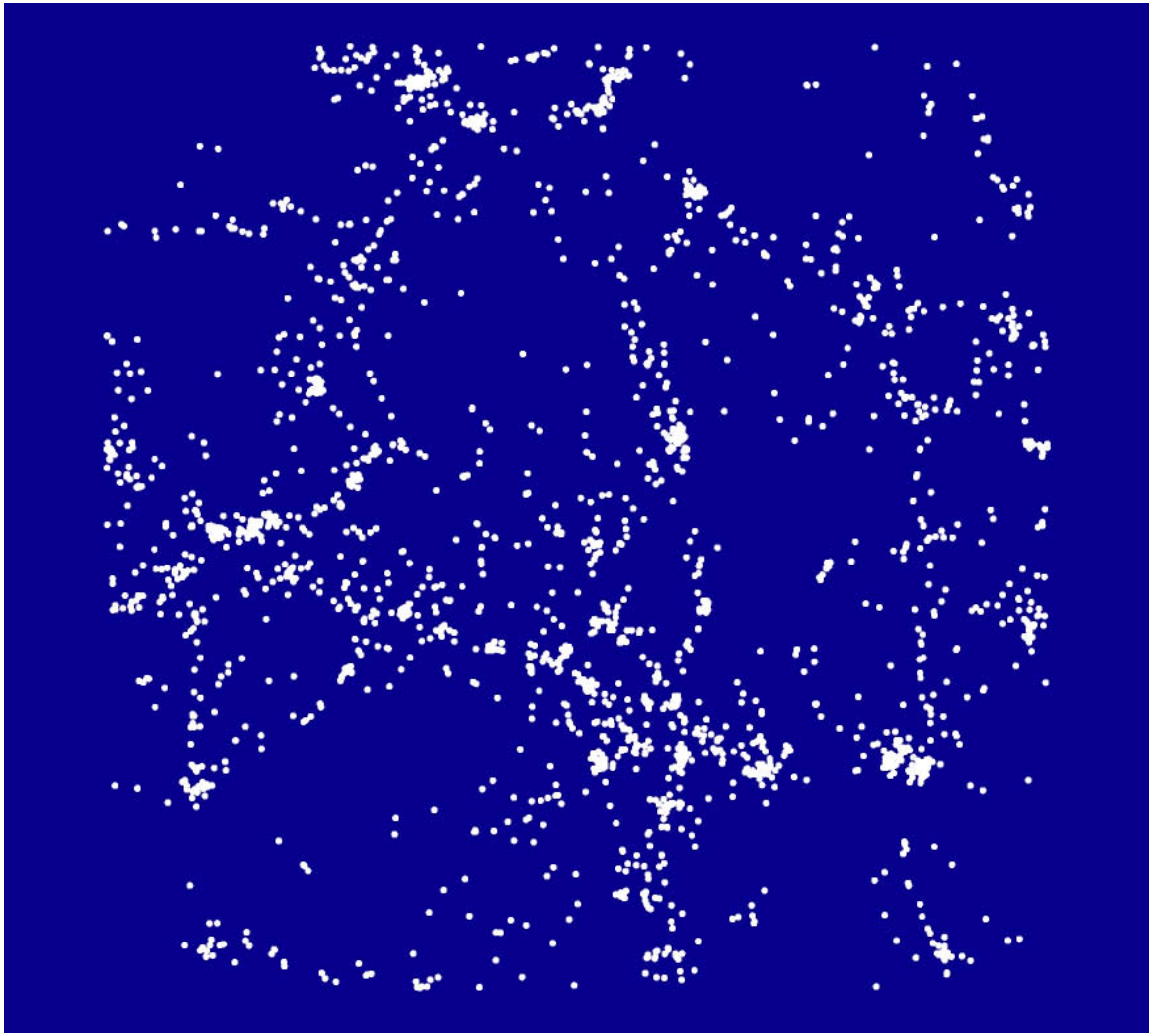
Estimating Filaments

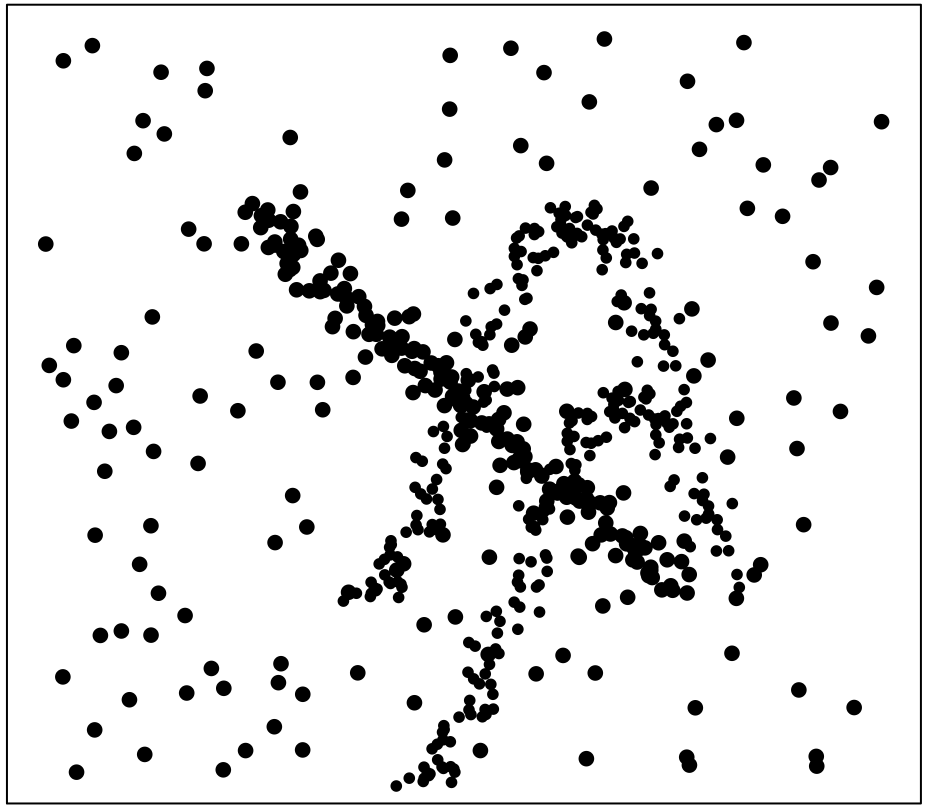
Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli and Larry Wasserman

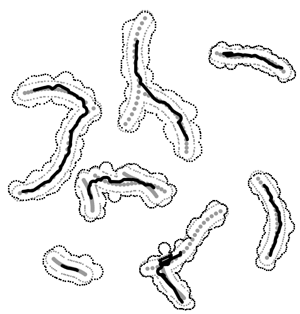
arxiv.org/abs/1003.5536

Toronto, April 2010









Related Problems

- Estimating blood vessel networks in neuroimaging
- Estimating seismic faults from earthquake epicenters
- Detecting minefields in aerial reconaissance images
- Identifying object boundaries in images
- Principal curves
- Manifold learning

Outline

- The Model
- Geometric Background
- Existing Methods
- New Methods
- Asymptotics
- Minimax Theory
- Examples

The Model

$$Y_i = f(U_i) + \epsilon_i$$

where

$$U_1, \ldots, U_n \sim H$$

are unobserved variables on $[0,1]$ and $f : [0,1] \to \mathbb{R}^2$.

Noise model for ϵ_i:

$$\epsilon_i \sim F \quad \text{is supported on a Disc of radius } \sigma.$$

Later, we include background clutter:

$$Y_i = \begin{cases} f(U_i) + \epsilon_i & \text{with prob } \pi \\ \text{Uniform} & \text{with prob } 1 - \pi \end{cases}$$

In general, f can be: open, closed, simple, self-intersecting, discontinuous (multiple curves).

For now, ignore the background clutter.

The Model

We don't use a Normal noise model since then:

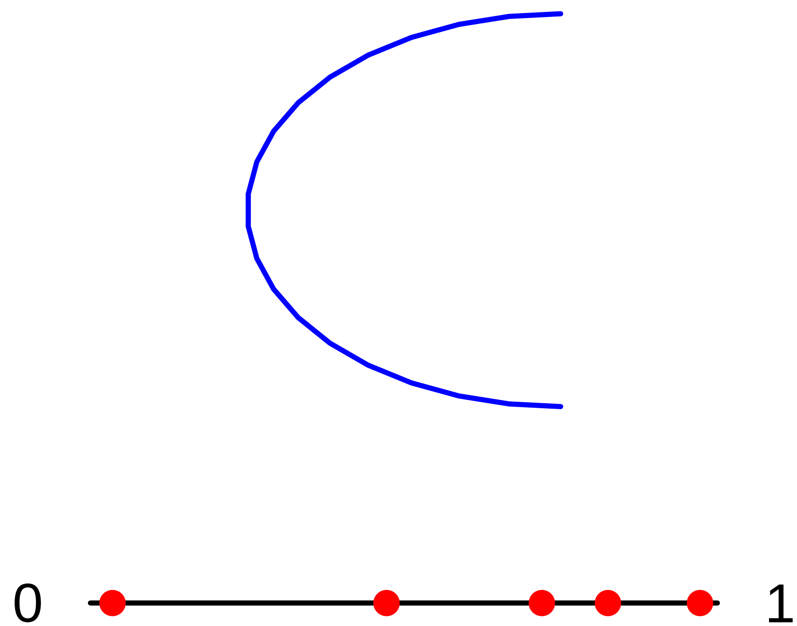
$$\max_i ||Y_i - f(U_i)|| \to \infty$$

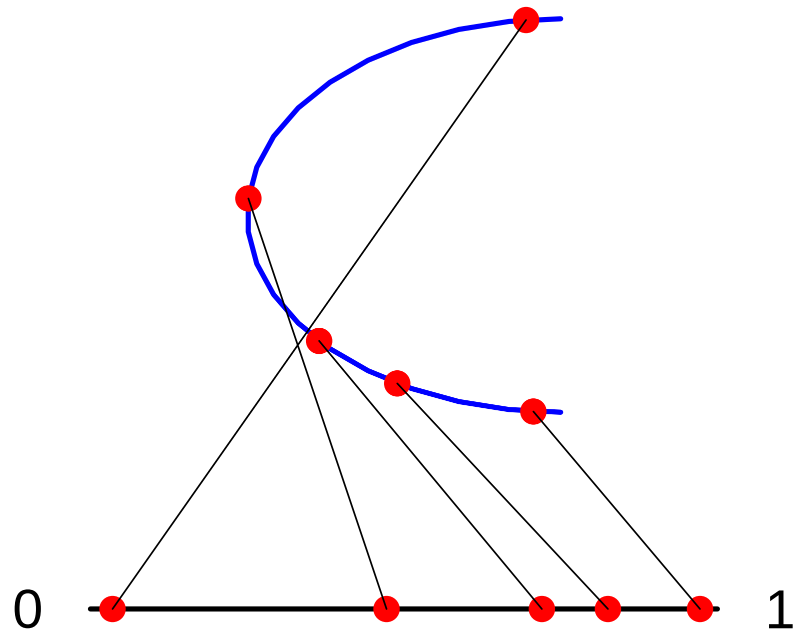
as $n \to \infty$.

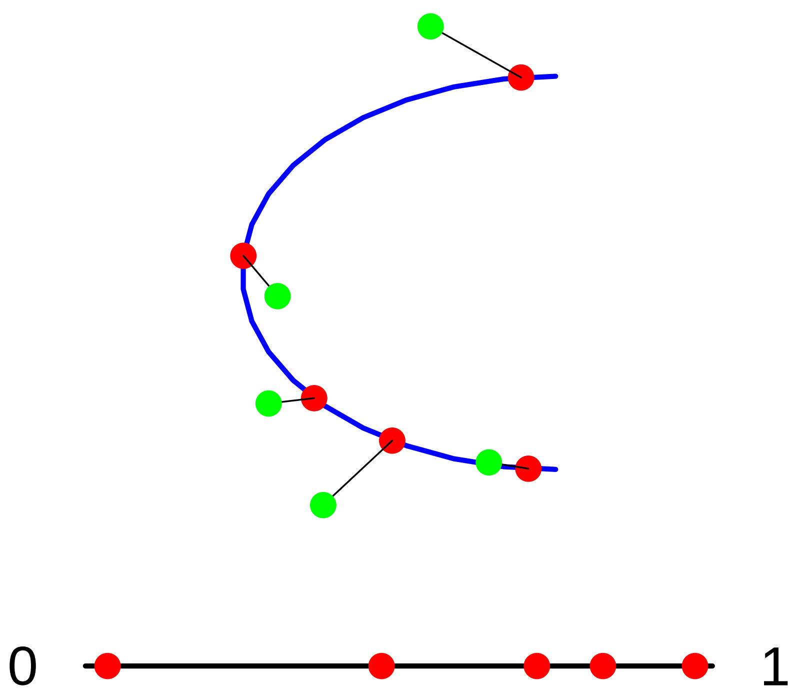
But we expect the points to cluster around the filaments.

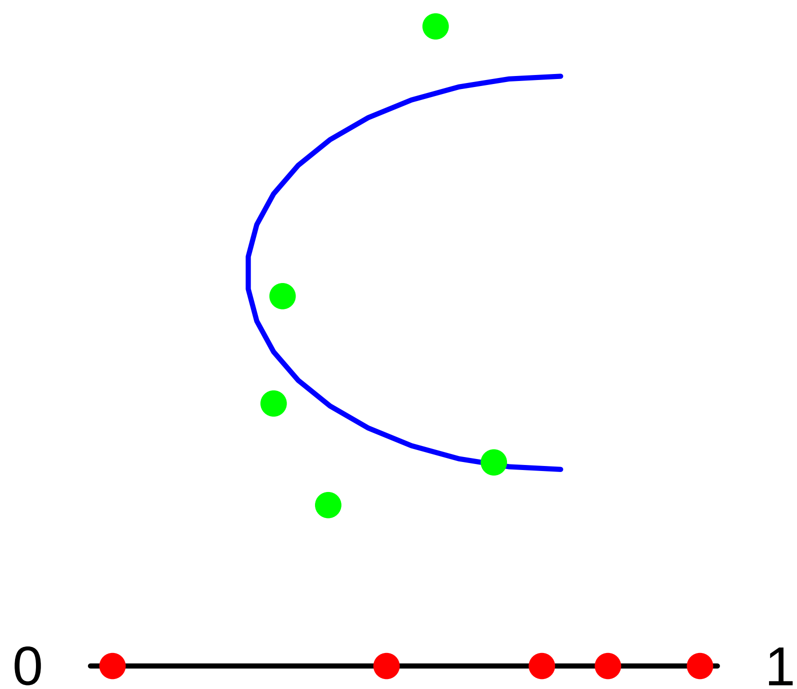
Hence, the compactly supported noise model is better.

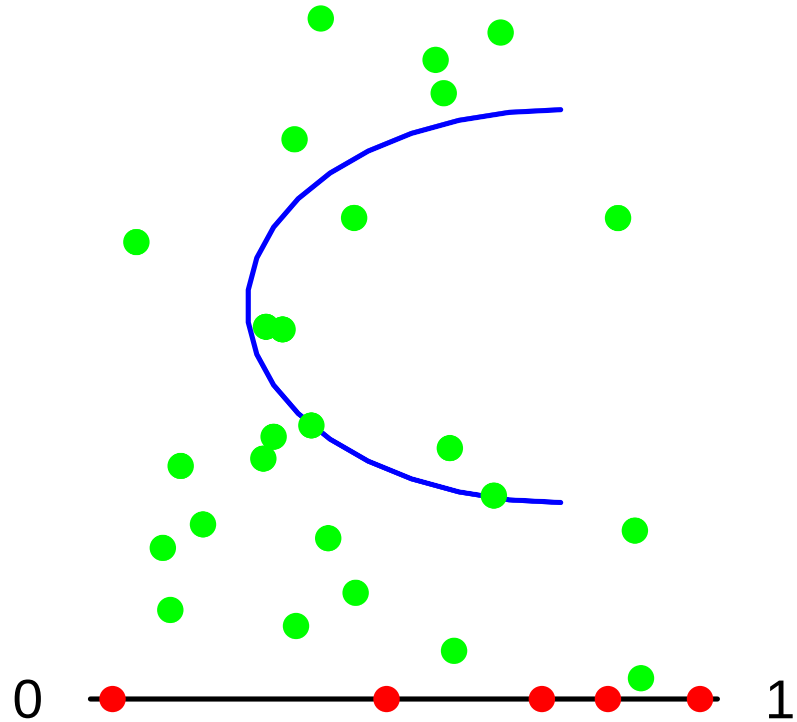
Also, with Normal noise one gets rates of the form $O(1/(\log n)^\alpha)$.

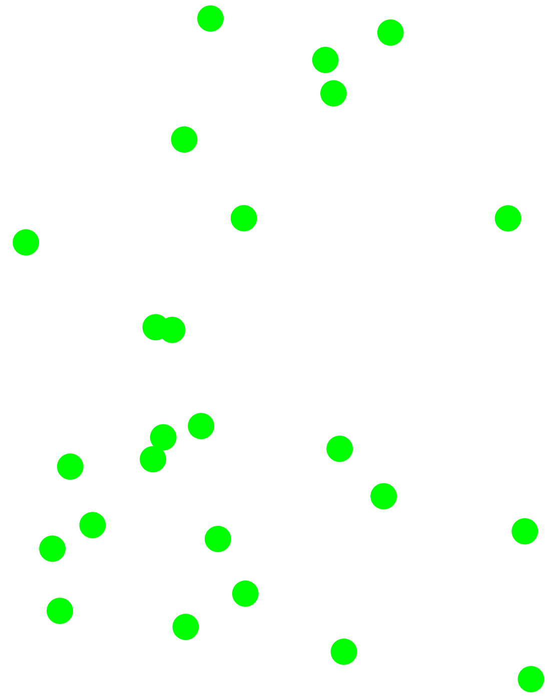

0
1


0
1


0
1


0
1


0
1



Available Methods

- Principal curves (Hastie and Stuetzle 1989)
- Second generation principal curves (Kegl et al. 2000, Smola et al. 1999)
- Penalized nonparametric likelihood (Tibshirani 1992)
- Manifold learning (ISOMAP, LLE, LLP)
- Deconvolution
- Beamlets (Xuo and Donoho)
- Combinatorial curve reconstruction (computational geometry)
- Gradient-based (GPVW 2009)
- Geometric Smoothing(Today)

Unanswered Questions

- When are these methods consistent?
- What is the rate of convergence?
- How do we choose the tuning parameters?
- What is the minimax risk?

GEOMETRIC BACKGROUND

Hausdorff Distance

For any set A define the enlargement

$$A \oplus \delta = \bigcup_{x \in A} B(x, \delta)$$

where $B(x, \delta)$ is a ball centered at x with radius δ.

The Hausdorff distance between two sets A and B is

$$d_H(A, B) = \inf\{\delta : A \subset B \oplus \delta \ \text{ and } \ B \subset A \oplus \delta\}.$$

Loss function:

$$d_H(\Gamma_f, \widehat{\Gamma})$$

$$\Gamma_f = \{f(u) : \ 0 \le u \le 1\}$$

is the filament set.

Relation to Regression

• If $U_1, \ldots, U_n$ were observed, this reduces to ordinary nonparametric regression.

• If only the order of the Y_i's were known, this is related to nonparametric regression with measurement error.

The Noise Free Case

(An Aside)

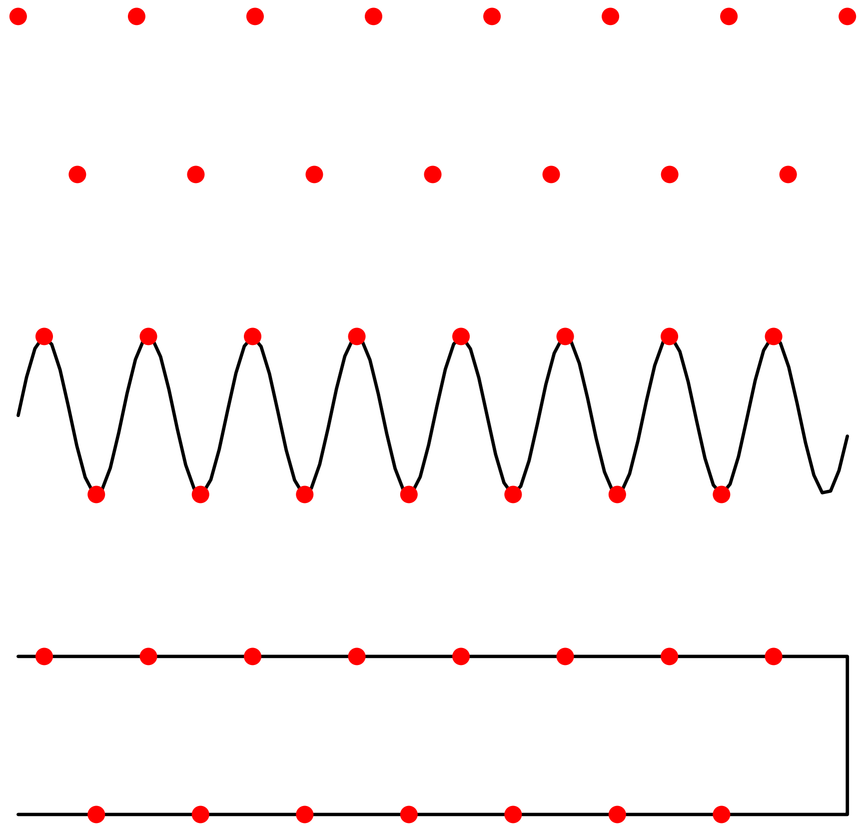
Unlike regression, even if $\epsilon_i = 0$ for all i, we are not done. Suppose that

$$Y_i = f(U_i) \quad i = 1, \ldots, n$$

There is no error but you only observe $Y_1, \ldots, Y_n$. How do you estimate f?

You need to order the Y_i's.

Ordering



Three Relevant Orderings

• The true order π_f is the permutation such that

$$\pi_f(i) < \pi_f(j) \quad \text{iff} \quad f^{-1}(\mu_{\pi(i)}) < f^{-1}(\mu_{\pi(j)}).$$

• Travelling Salesman ordering: π_{TS} = permutation that gives the shortest path through the points.

• Nearest Neighbor ordering: π_{NN} = permutation obtained by consecutively connecting each point to its nearest neighbor.

Theorem(Giesen 1999) Assume no noise. Then

$$\pi_f = \pi_{TS} = \pi_{NN} \quad a.s.$$

for all large n. Also, the linear interpolant based on any of these orderings converges to f. In fact, $d_H(\Gamma_f, \widehat{\Gamma}) = O_P(1/n)$.

But the main problem is the noise.

Medial Axis

Let S be a set. Let ∂S be the boundary of S. A ball $B \subset S$ is medial if

$$\text{interior}(B) \cap \partial S = \emptyset$$

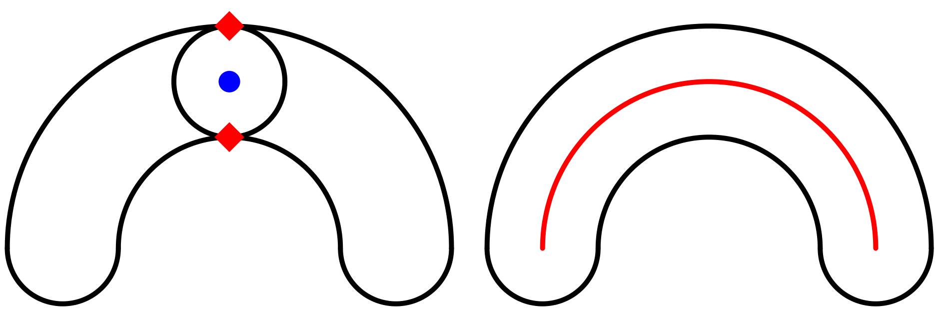
and

$$|B \cap \partial S| \geq 2.$$

The medial axis $M(S)$ is

$$M(S) = \text{closure}\{\textit{centers of the medial balls}\}.$$

Medial Axis



Medial Axis

Let $q(y)$ be the density of Y. Let

$$S = \mathrm{support}(q) = \{y :\ q(y) > 0\}.$$

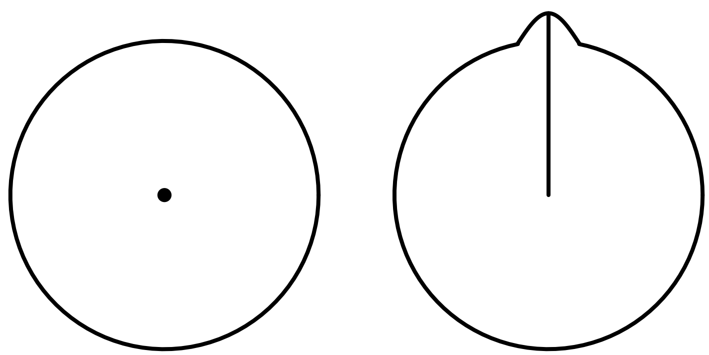
Under regularity conditions we have

$$M(S) = \Gamma_f$$

that is, the filament is the medial axis of the support of q.

Medial Axis

However, the medial axis is not continuous in Hausdorff distance. Small perturbations to S give a completely different medial axis.



Thickness

Let $r(x, y, z)$ be the radius of a ball passing through x, y, z. Define the thickness $\Delta(f)$ (global radius of curvature, or normal injectivity radius) by

$$\Delta = \min_{x,y,z} r(x, y, z).$$

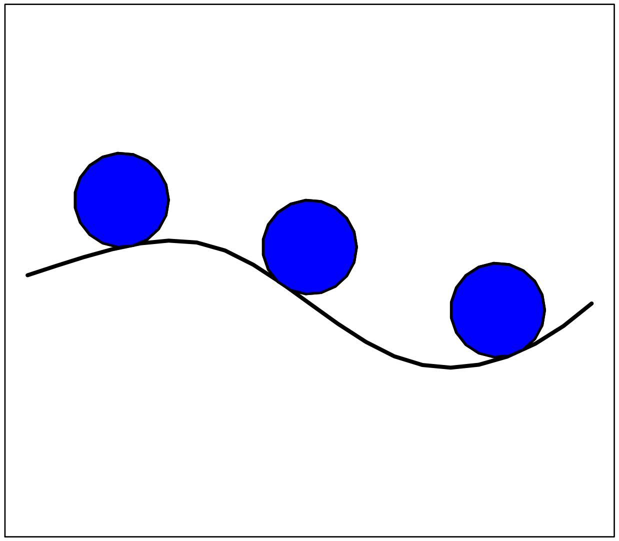
(See Gonzalez and Maddocks 1999.)

This measures local curvature as well as "closeness of approach."

A ball of radius Δ can roll freely around the curve. So Δ large means that f is smooth and not too close to being self-intersecting.

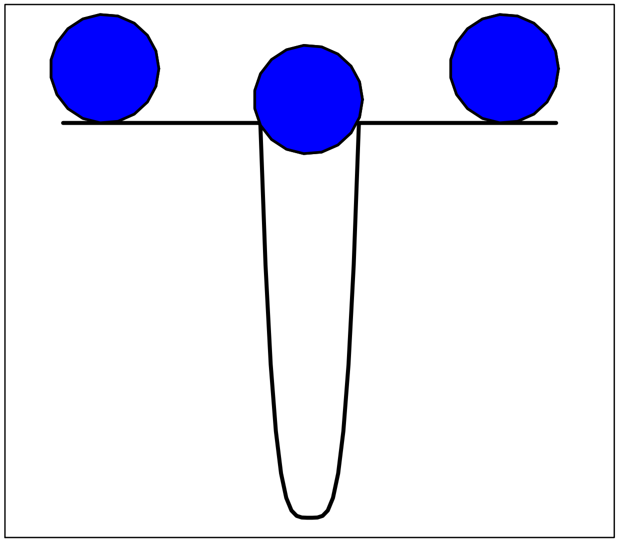
Thickness

If a ball B has radius Δ then it can roll freely:



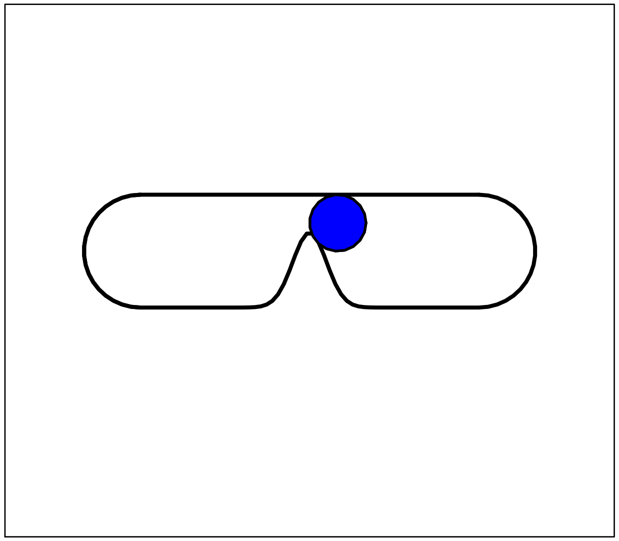
Thickness

If B has radius larger than Δ then one of these two things happen. It can't roll because of curvature:



Thickness

... or it can't roll because of a "close approach" of the curve:



EDT

The Euclidean Distance Transform (EDT) is

$$\Lambda(y) = d(y, \partial S) = \min_{x \in \partial S} ||y - x||$$

for $y \in S$. Thus, $\Lambda(y)$ is the distance from y to the boundary.

$\Lambda(y) = 0$ for $y \in \partial S$. Otherwise, $\Lambda(y) > 0$.

Nice Sets

S is standard if there are $\delta, \lambda > 0$ such that

$$\text{Lebesgue}(B(y,\epsilon) \cap S) \geq \delta \ \text{Lebesgue}(B(y,\epsilon))$$

for all $y \in S$ and all $0 < \epsilon \leq \lambda$. This means that S has no pointy parts.

S is expandable if there are $r > 0$ and $R \geq 1$ such that

$$d_H(\partial S, \partial S^{\epsilon}) \leq R\epsilon$$

for all $0 \leq \epsilon < r$.

Medial Axis = Filament

Let $S = \{y : \; q(y) > 0\}$ be the support.

Theorem:

If $\sigma < \Delta(f)$ then

- $\Gamma_f = M(S)$.
- S is standard.
- S is expandable.
- $y \in M(S)$ iff $\Lambda(y) = \sigma$.
- $y \notin M(S)$ iff $\Lambda(y) < \sigma$.

ESTIMATING THE FILAMENT

Estimation

For now, assume no background clutter and a single filament. First we estimate S and ∂S. Let

$$\widehat{S} = \bigcup_{i=1}^{n} B(Y_i, \epsilon_n)$$

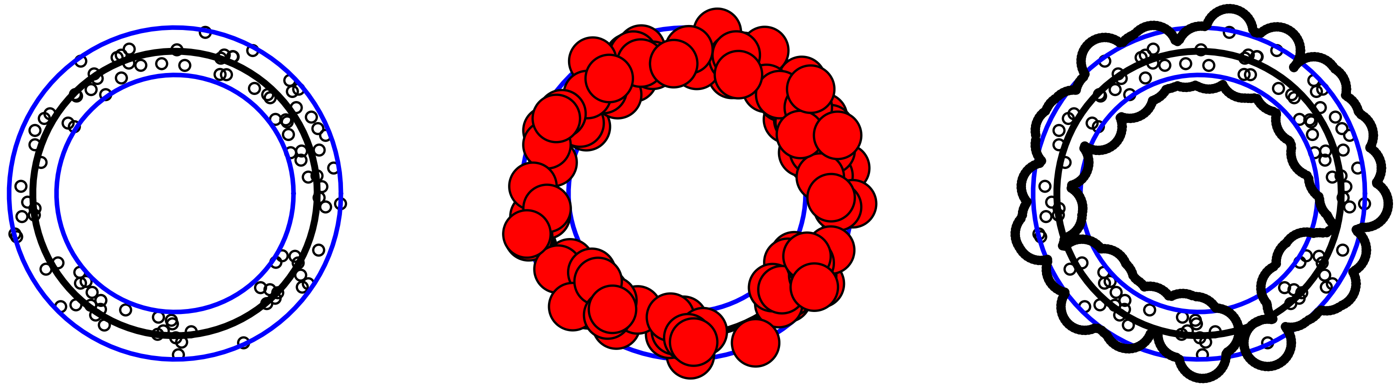
where $\epsilon_n = O(\sqrt{\log n/n})$. Then, almost surely, for all large n,

$$d_H(S, \widehat{S}) \leq C\sqrt{\frac{\log n}{n}} \quad \text{and} \quad d_H(\partial S, \partial\widehat{S}) \leq C\sqrt{\frac{\log n}{n}}.$$

(Cuevas and Ridriguez-Casal 2004.)

Later, we will discuss improved estimators. But note that $\widehat{S}$ is very simple.

Estimation



Estimation

Next we construct two estimators: the EDT estimator and the medial estmator.

The EDT Estimator. Let

$$\widehat{\Lambda}(y) = d(y, \widehat{\partial S}).$$

Let

$$\widehat{\sigma} = \max_{y \in \widehat{S}} \widehat{\Lambda}(y).$$

Let

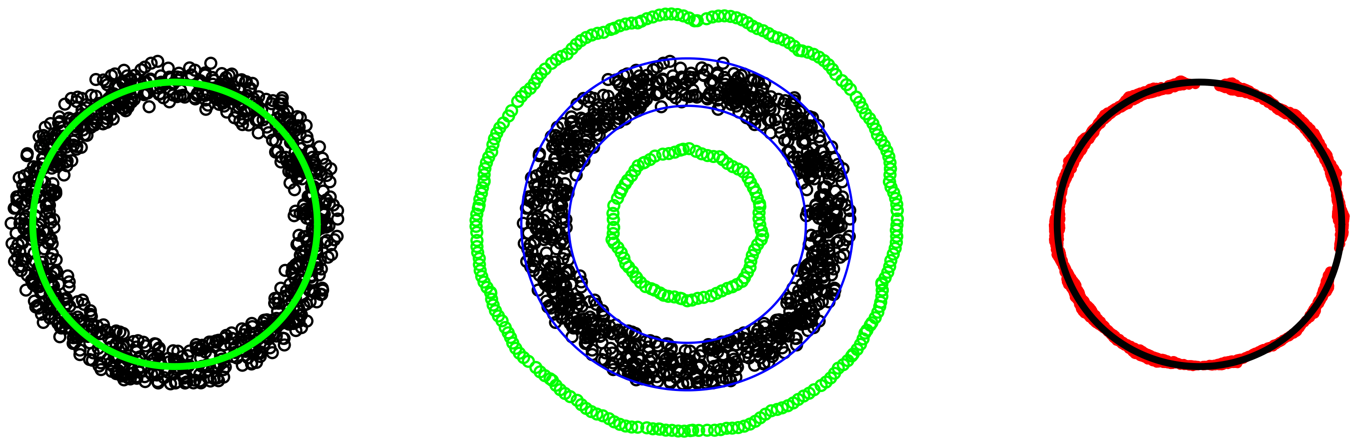
$$\widehat{M} = \{y \in \widehat{S} : \ \widehat{\Lambda}(y) \geq \widehat{\sigma} - 2\epsilon_n\}.$$

Theorem:

$$d_H(\widehat{M}, \Gamma_f) = O_P\left(\sqrt{\frac{\log n}{n}}\right).$$

Note that $\widehat{M}$ is a set not a curve.

Estimation



Estimation

The Medial Estimator. (For closed curves.)

• Decompose $\widehat{\partial S} = \widehat{\partial S}_0 \cup \widehat{\partial S}_1$.

• For each $y \in \widehat{\partial S}_0$, find closest $x \in \widehat{\partial S}_1$ and let $\widehat{\mu}(y)$ be the midpoint of the line joining y and x.

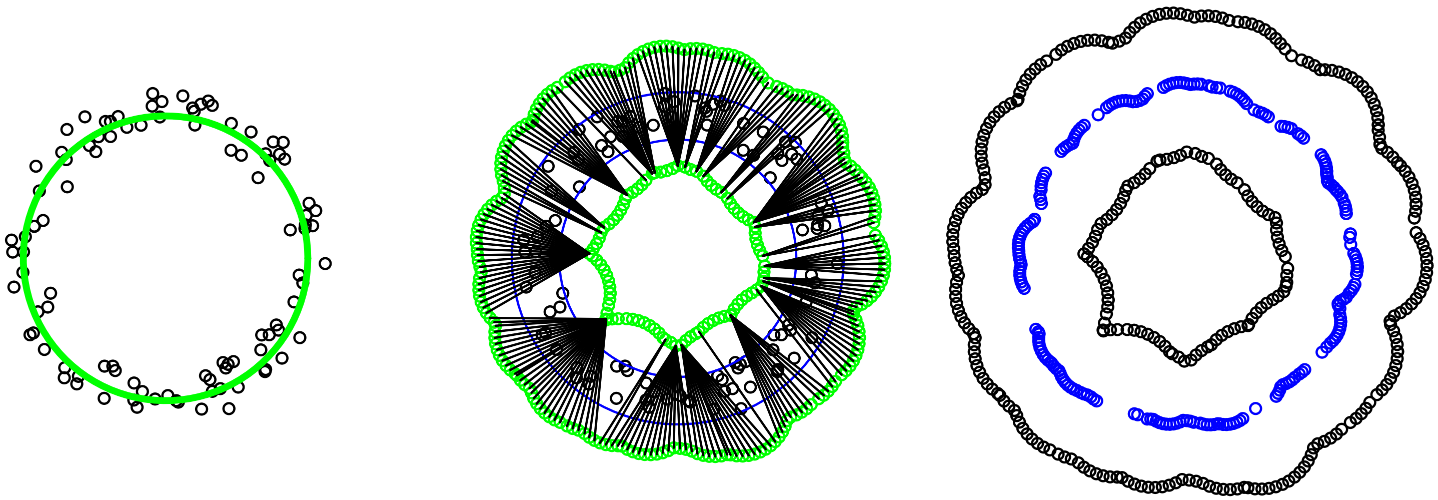
• Set $\widehat{M} = \{\widehat{\mu}(y) : \ y \in \widehat{\partial S}_0\}$.

Theorem:

$$d_H(\widehat{M}, \Gamma_f) = O_P\left(\frac{\log n}{n}\right)^{1/4}.$$

We will improve these rates shortly.

Estimation



Curve Extraction

EDT. (Open curves.)

1. Find two furthest points y_0 and y_1 in $\widehat{M}$ (in arc length.)

2. Connect y_0 and y_1 with shortest path $\widehat{\Gamma}$.

These steps can be approximated by sampling from $\widehat{M}$ and using a minimal spanning tree. Then

$$d_H(\Gamma_f, \widehat{\Gamma}) = O_P\left(\sqrt{\frac{\log n}{n}}\right).$$

Any smoothing procedure can be applied to $\widehat{\Gamma}$. As long as the fitted value stay in $\widehat{M}$, the rate of convergence is preserved.

Curve Extraction

Medial estimator. The set $\widehat{M}$ consists of a union of disconnected curves. Complete the estimator by linearly interpolating the disconnected components.

Theorem The completed estimator is a simple closed curve and

$$d_H(\widehat{M}, \Gamma_f) = O_P\left(\sqrt{\frac{\log n}{n}}\right).$$

Note the faster rate.

The differences of the fitted values also provide an estimate of the gradient with rate $(\log n/n)^{1/4}$.

Multiple Curves

• We have similar results for multiple curves that are sufficiently separated.

• For self-interecting curves, the same results apply to the parts of the curve not too close to the intersections.

MINIMAX ESTIMATION

Minimax Estimation

Let

$$\Theta = \{(f, h, \sigma) : \; 0 \leq \sigma \leq \Delta(f) - a, \quad \Delta(f) \geq d, h \in \mathcal{H}\}$$

where h is the density of U_i and

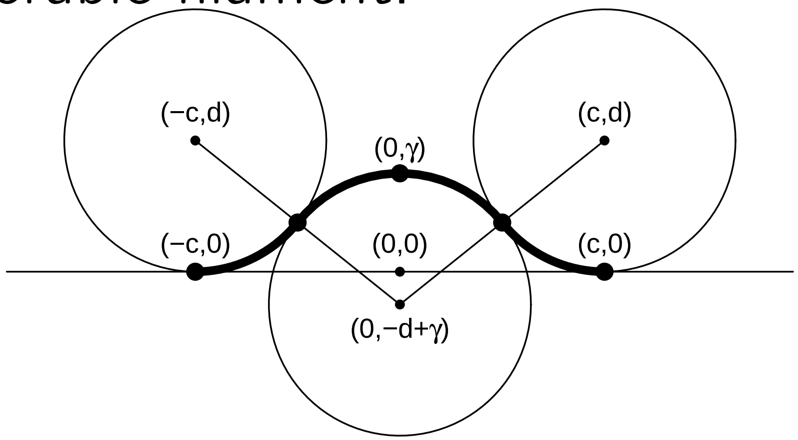
$$\mathcal{H} = \{h : \; c_1 \leq h \leq c_2\}.$$

Theorem

$$\inf_{\widehat{\Gamma}} \sup_{f,\sigma,h} \mathbb{E}(d_H(\Gamma_f, \widehat{\Gamma})) \geq \frac{C}{n^{2/3}}.$$

Minimax Estimation

Proof uses Assoaud's lemma. The hypercube is built from the following least favorable filament:



Push the middle ball up. Roll in balls from left and right.

Minimax Estimation

- To achieve the minimax rate, replace $\widehat{S}$ with a smoother estimator as in Mammen and Tsybakov (1995). If we do this then both estimators are minimax.

- Create a finite net of sets $\mathcal{G} = \{S_1, \ldots, S_N\}$.

- Take

$$\widehat{S} = \text{argmin}\{\text{Lebesgue}(S) : \ \{Y_1, \ldots, Y_n\} \subset S\}.$$

- Take $\widehat{\partial S} = \partial \widehat{S}$. Then

$$\sup_{(f,\sigma,h)\in\Theta} E_{f,\sigma,h} d_H(\partial S, \widehat{\partial S}) \leq \frac{C}{n^{2/3}}.$$

Minimax Estimation

- However, this estimator is mainly of theoretical interest. Can't really compute this.

- The Hall-Park-Turlach (2002) "rolling ball" estimator may be feasible and appears to achieve the same rate of convergence.

- Currently, we use the (suboptimal) union of balls estimator because it is extremely simple and only requires one tuning parameter.

Decluttering

$$Y_1, \ldots, Y_n \sim m(y) = (1-\eta)q_0 + \eta q$$

where q_0 is uniform. Bayes rule:

$$c(y) = I(m(y) \geq 2(1-\eta)q_0(y))$$

conservative rule:

$$c(y) = I(m(y) \geq 2q_0(y))$$

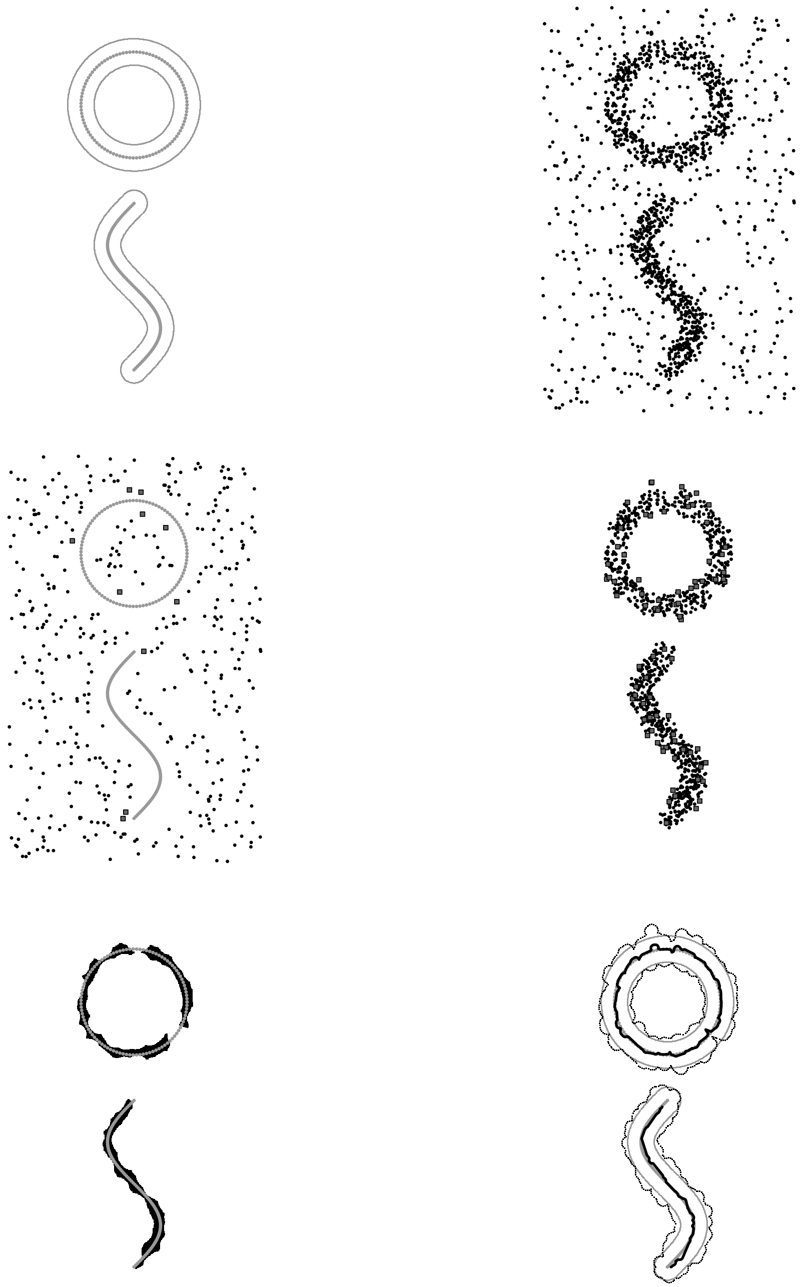
estimate:

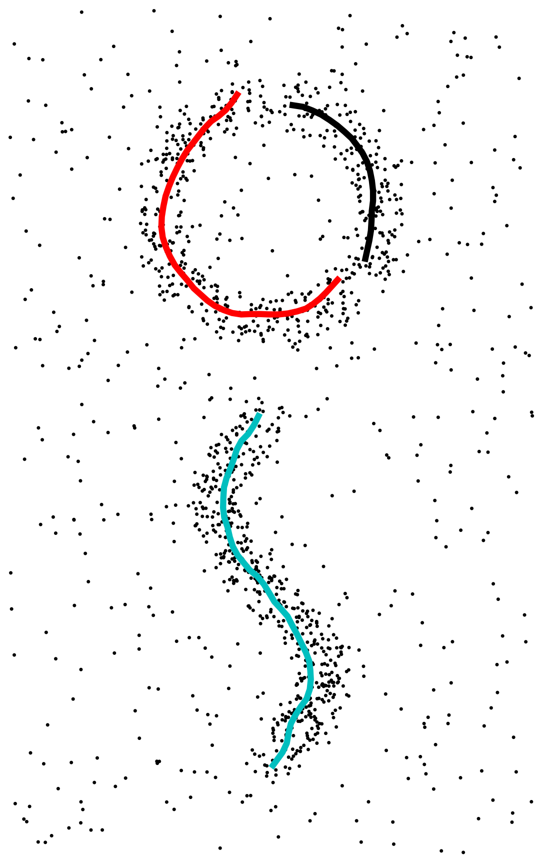
$$\widehat{c}(y) = I(\widehat{m}(y) \geq 2q_0(y))$$

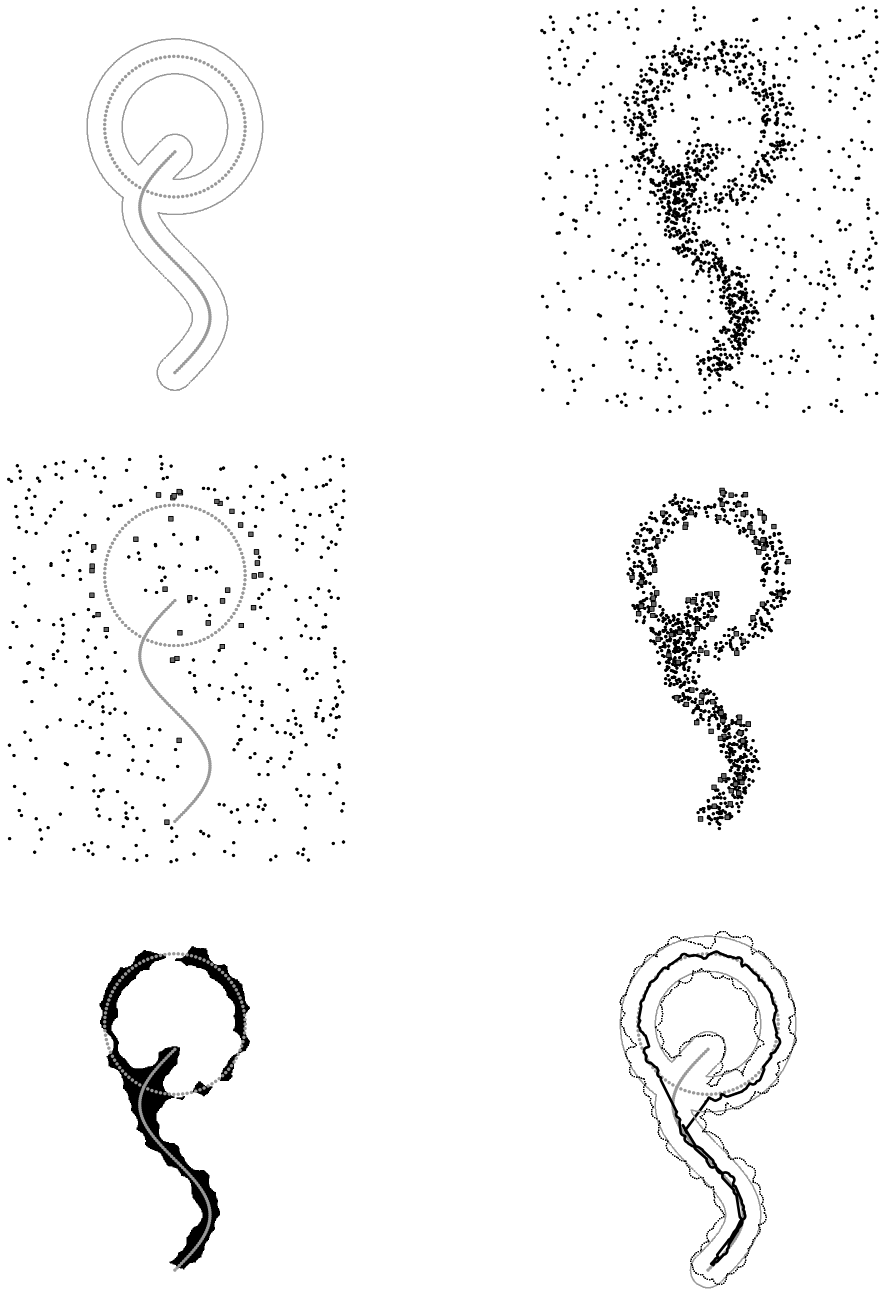
Use:

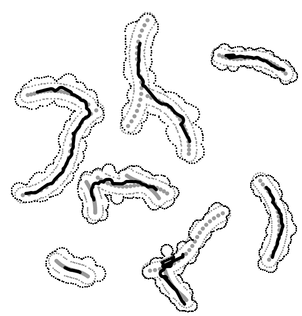
$$\mathcal{Y} = \{Y_i : \ \widehat{c}(Y_i) = 1\}.$$

EXAMPLES









Conclusion

Currently we are working on the following extensions:

- apply to astro data (SDSS and n-body simulations)
- extends readily to higher dimensions
- can allow σ to vary
- smoother methods
- other noise models
- tuning parameters
- compare to beamlets

THE END

OTHER METHODS

Principal Curves

Original version (Hastie and Stuetzle 1989).

f_* is the self-consistent smooth curve:

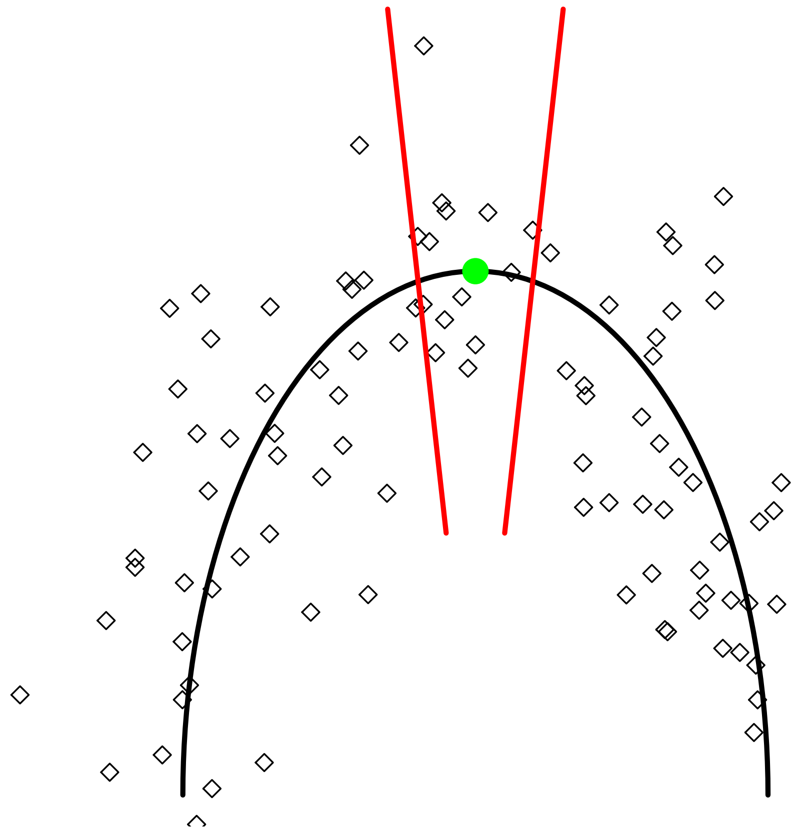
$f_*(x) = \mathbb{E}(Y | \Pi_f Y = x)$.

Algorithm: iterate these two steps:

(1) Project data onto curve

(2) Regress (smooth) data given the projections.

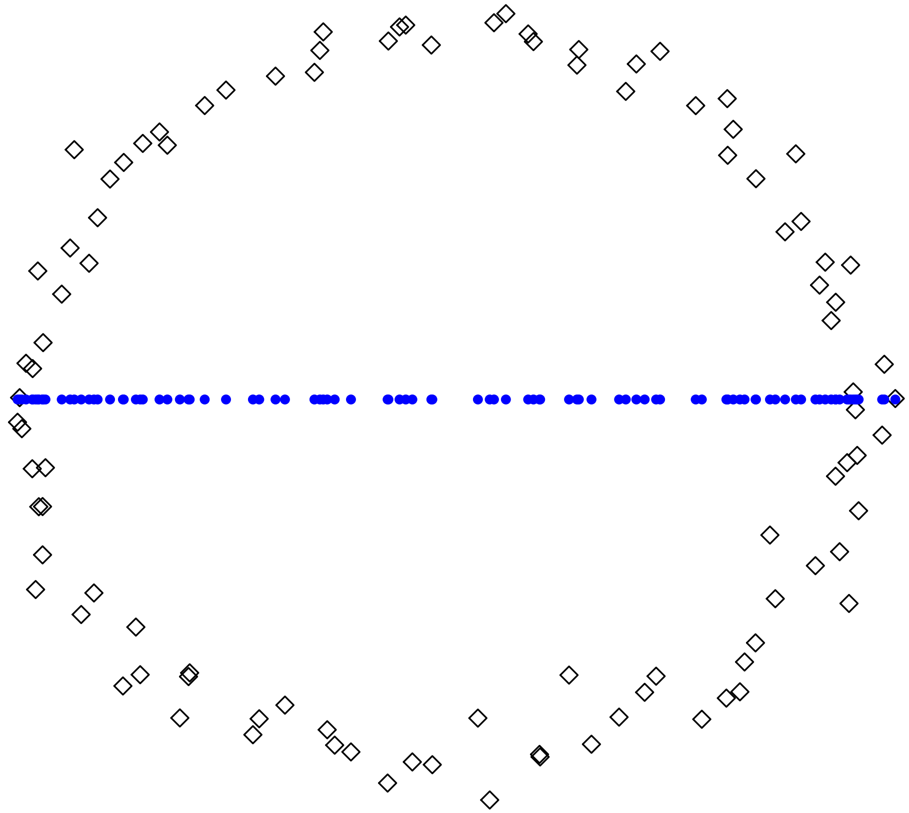
Principal Curves



Principal Curves

- f_* need not exist
- $f \neq f_*$ but close: $f_* = f + O(\sigma^2 \text{ Curvature})$
- not much theory
- very sensitive to starting values.
- doesn't handle multiple curves well

Principal Curves



Second Generation Principal Curves

The principal curve f_* is

$$f_* = \mathrm{argmin}_{f\in\mathcal{F}} \mathbb{E}||Y - \Pi_f Y||^2$$

where $\mathcal{F}$ is a class of functions. If

$$\widehat{f} = \mathrm{argmin}_{f\in\mathcal{F}} \ \ \frac{1}{n}\sum_{i=1}^{n} ||Y_i - \Pi_f Y_i||^2$$

then, under conditions on $\mathcal{F}$,

$$\sup_u ||\widehat{f}(u) - f_*(u)|| \xrightarrow{P} 0.$$

Problems:

(i)difficult algorithms and

(ii) $f_* \neq f$.

Spin and Smooth

Suppose that

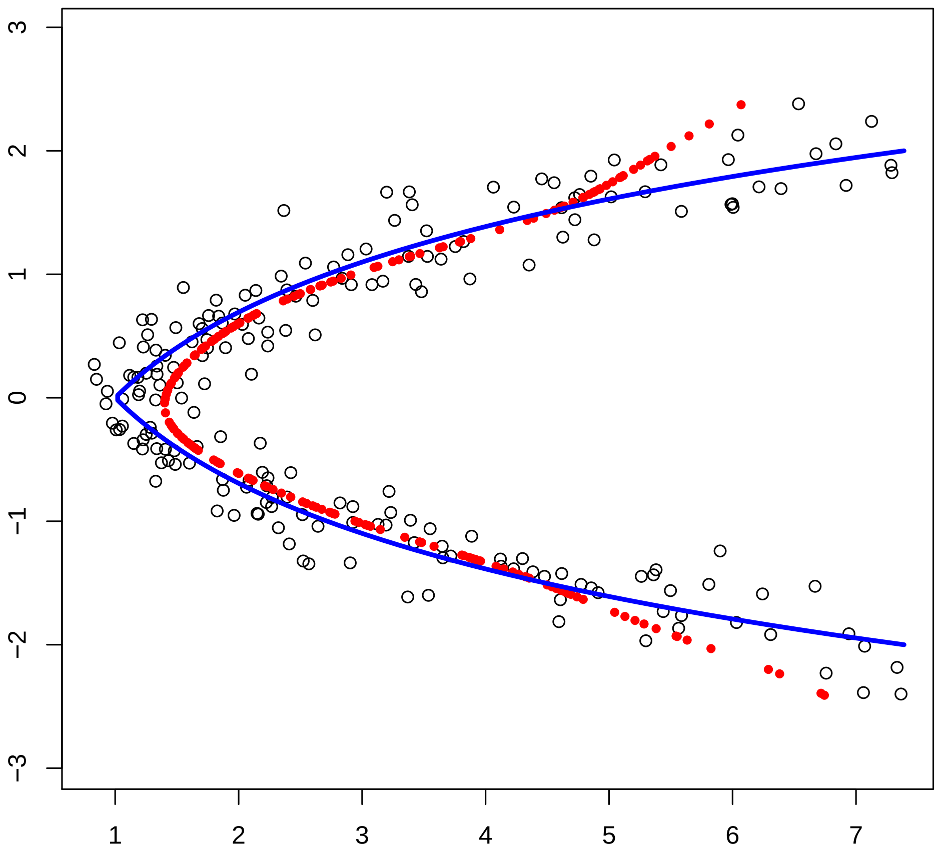
$$R_\theta \Gamma_f = \{(z, g(z)) : \; a \le z \le b\}$$

for some rotation R_θ and some function g. In other words, f is a function after some rotation.

- Rotate by R_θ
- apply smoother
- minimize RSS over θ

Then $\widehat{f}$ has the same asymptotic behavior as nonparametric regresson with measurement error.

Example



3
2
1
0
−1
−2
−3
1
2
3
4
5
6
7

For Multiple Filaments: Quantization

A *codebook* is a finite set of vectors $C = \{c_1, \ldots, c_k\}$.

A codebook induces a quantization function $Q(x) = \text{argmin}_j ||x - c_j||$ with risk $R(Q) \equiv R(C) = \mathbb{E}||X - Q(X)||^2$.

The minimal risk for codebooks of size k is $R_k = \inf_{Q \in \mathcal{Q}_k} R(Q)$.

Given data $X_1, \ldots, X_n$, the empirical risk is

$$\widehat{R}(Q) = \frac{1}{n} \sum_{i=1}^{n} ||X_i - Q(X_i)||^2,$$

which is minimized at some $\widehat{Q}$.

With high probability, $R(\widehat{Q}) \leq R_k + O(\sqrt{k \log k / n})$.

For Multiple Filaments: Quantization

Extend quantization algorithm and theory to codebooks of curves $C = \{f_1, \ldots, f_k\}$ (cf. Kegl et al. 2000;Smola et al. 2001).

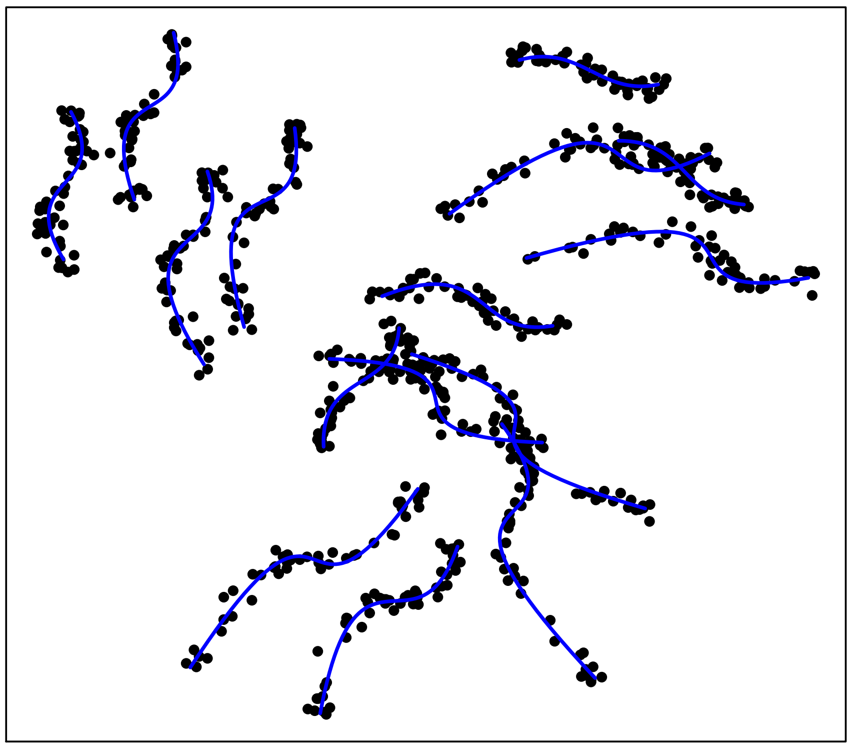
Use k-means clustering but apply spin-and-smooth within each cluster.

With high probability,

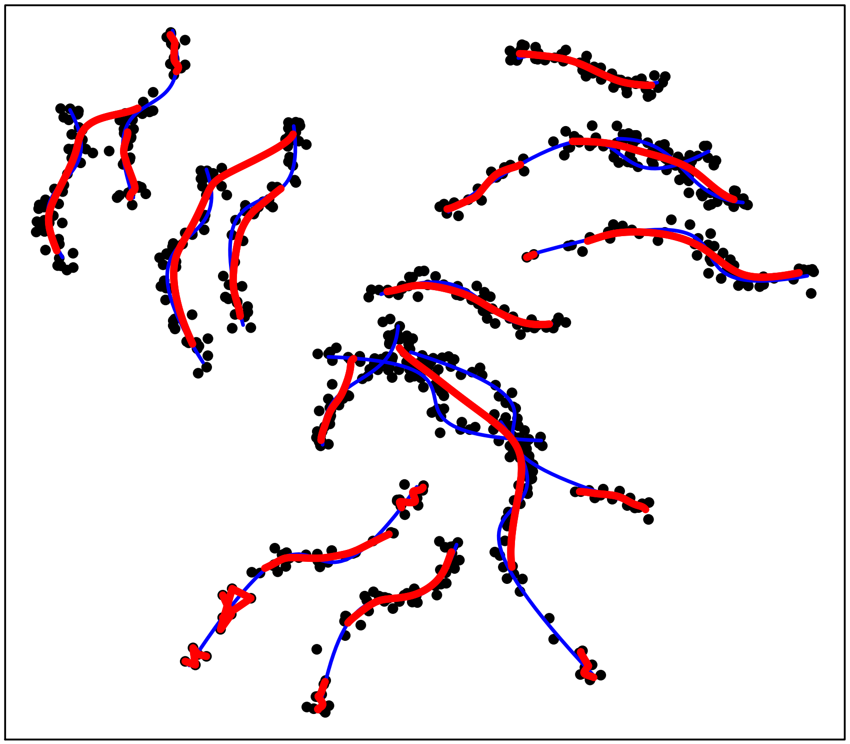
$R(\widehat{Q}) \leq R_k + O(\sqrt{k \log k/n})$.

But this inherits all the problems of clustering: chosing k, starting values etc. (See also Stanford and Raftery 2000).

For Multiple Filaments: Quantization



For Multiple Filaments: Quantization



Local Smoothing

Called *moving least squares* in computational geometry and *local linear projection LLP* in manifold learning.

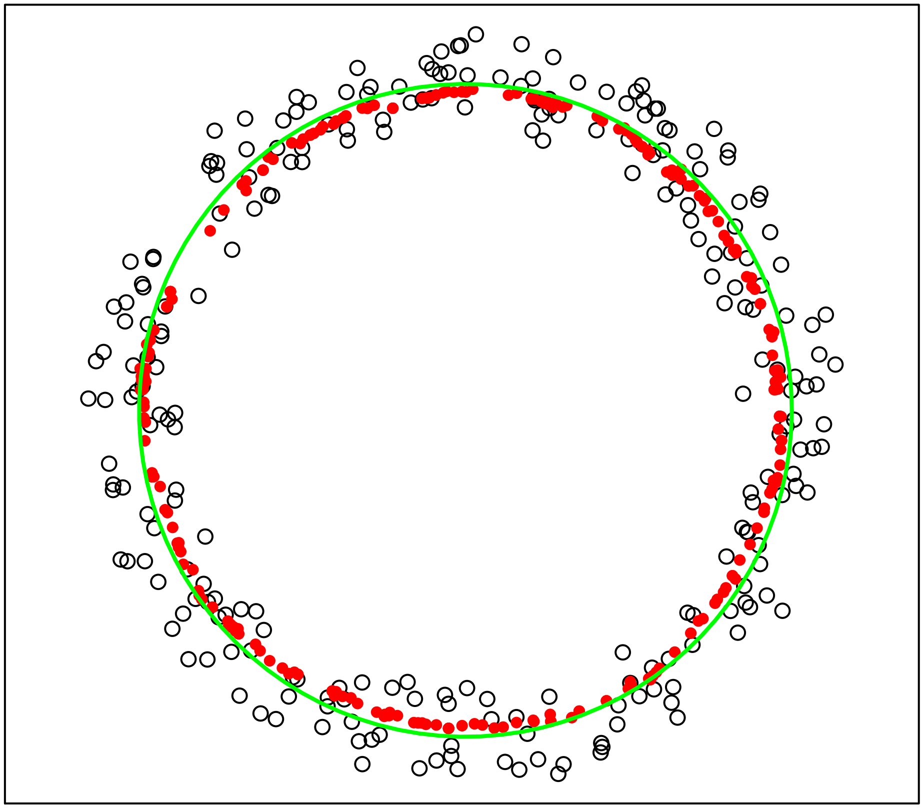
- For each Y_i fit PCA line to all points in a neighborhood of size h.
- $\widehat{\mu}_i =$ projection of Y_i onto the line.

A simpler (and essentially equivalent) version is to set $\widehat{\mu}_i =$ to the local average:

$$\widehat{\mu}_i = \frac{\sum_j Y_j K_h(||Y_j - Y_i||)}{\sum_j K_h(||Y_j - Y_i||)}.$$

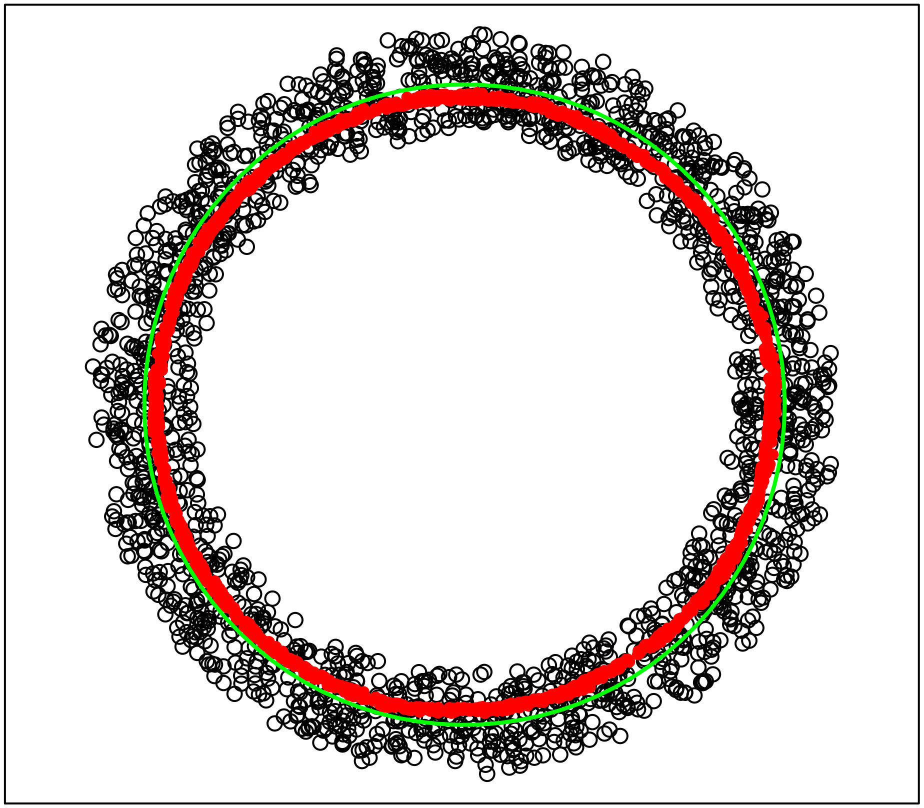
However, this method is not consistent. (Closely related to the mean shift algorithm.)

Example



$n = 250$

Example



$n = 2000$

We see the lack of consistency here.

Prune and Smooth

- Density estimate $\widehat{p}$
- Order the points by density:

$$\widehat{p}(Y_{(1)}) > \widehat{p}(Y_{(2)}) > \cdots > \widehat{p}(Y_{(n)})$$

Select the $k_n = n^{3/4}$ points with highest density
- Apply local smoother to these points with $h_n = n^{-1/8}$
- Decimate: $||\widehat{\mu}_i - \widehat{\mu}_{i-1}|| > \delta_n = n^{-1/4}$
- Apply NN ordering algorithm.

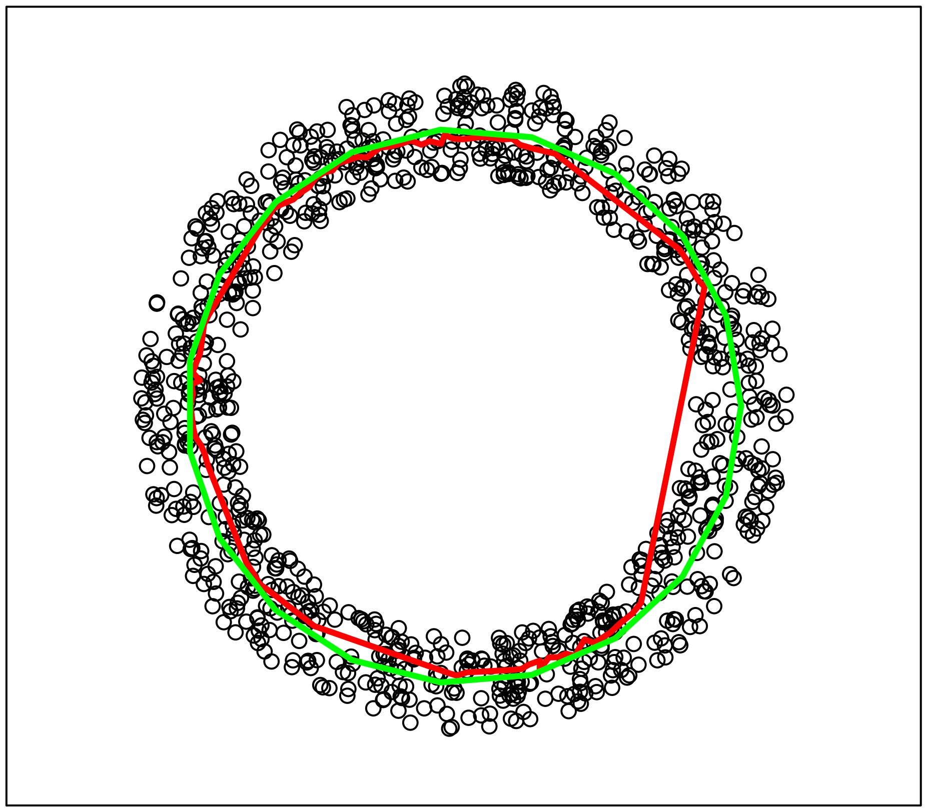
Then

$$\max_i ||\widehat{\mu}_i - \mu_i|| = O_P\left(\frac{\log n}{n^{1/4}}\right).$$

This is similar in spirit to the method in Cheng et al (2004) and Lee (2000).

$$Y \longrightarrow \text{prune} \longrightarrow \text{smooth} \longrightarrow \text{decimate} \longrightarrow \text{order} \quad \widehat{\mu}$$

Example



Normal Smoothing

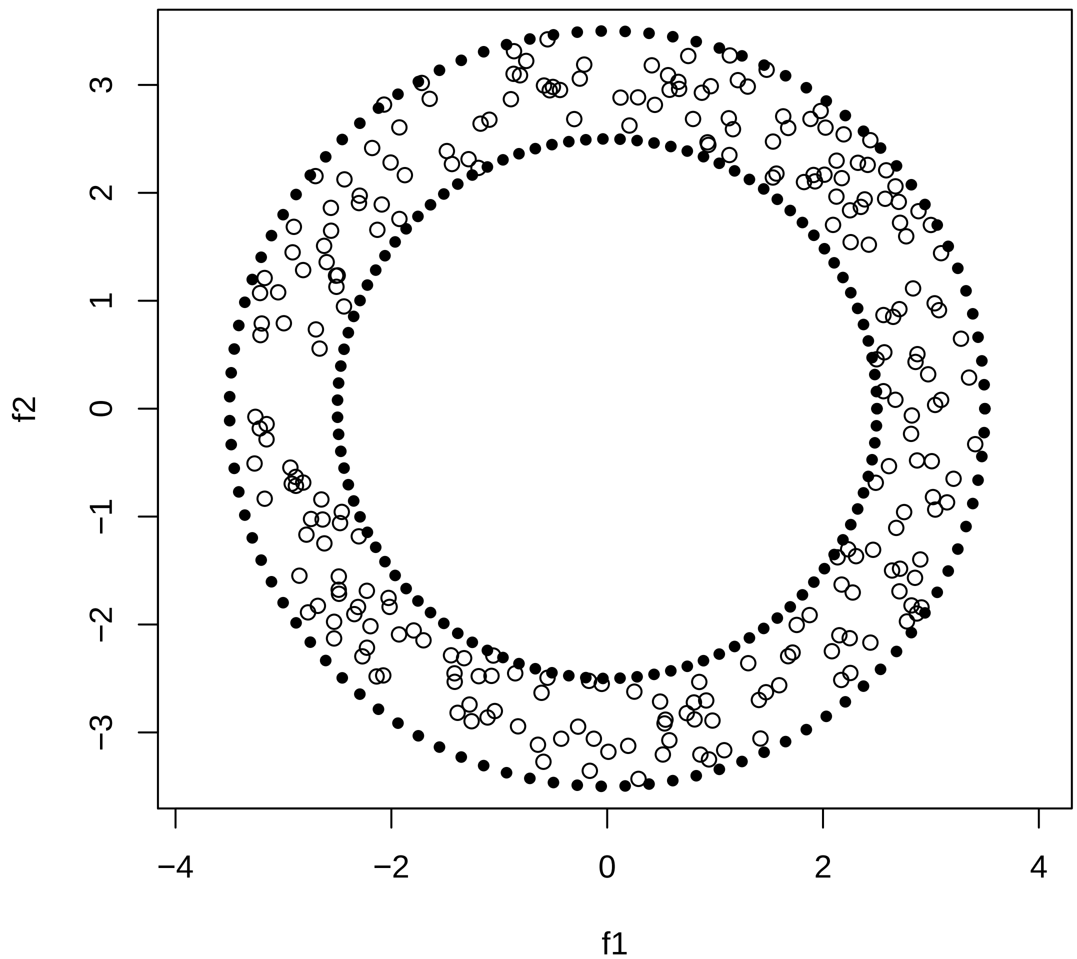
Estimate the gradient towards the medial axis (normal of the filament). Let

$$\text{line}_i = \{Y_i + t\ \nabla\widehat{p}(Y_i):\ \ t \in \mathbb{R}\}.$$

Let

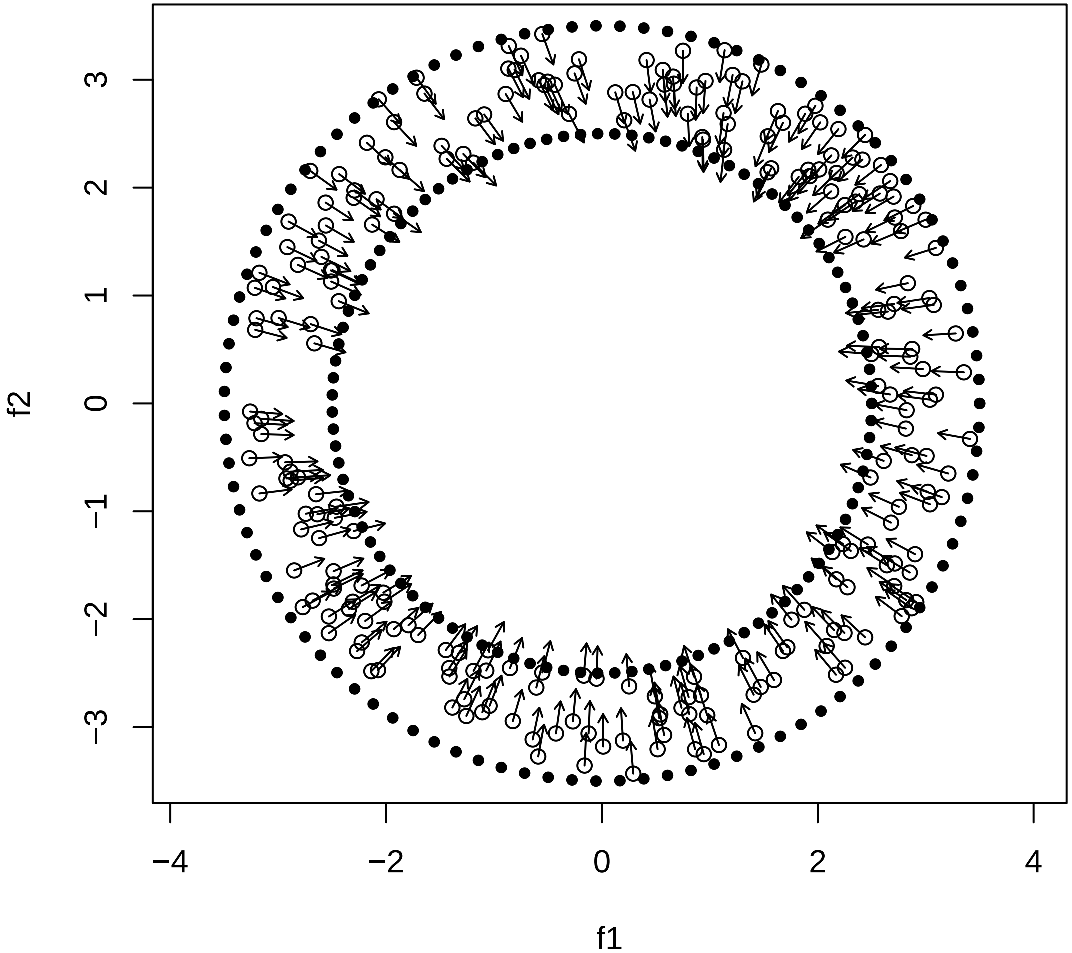
$$\widehat{\mu}_i = \text{midpoint}\left(\text{line}_i \cap \widehat{S}\right).$$

Normal Smoothing



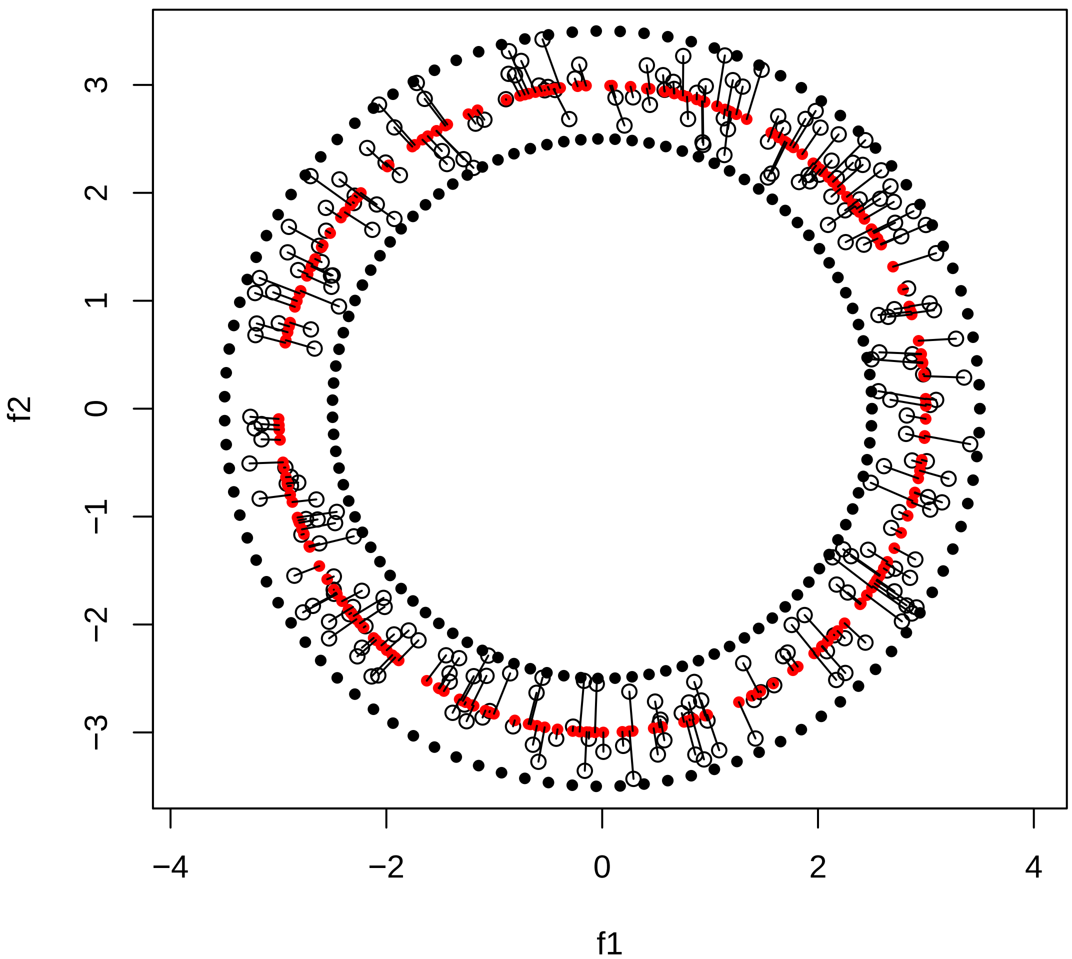
f1
f2
−4
−2
0
2
4
−3
−2
−1
0
1
2
3

Normal Smoothing



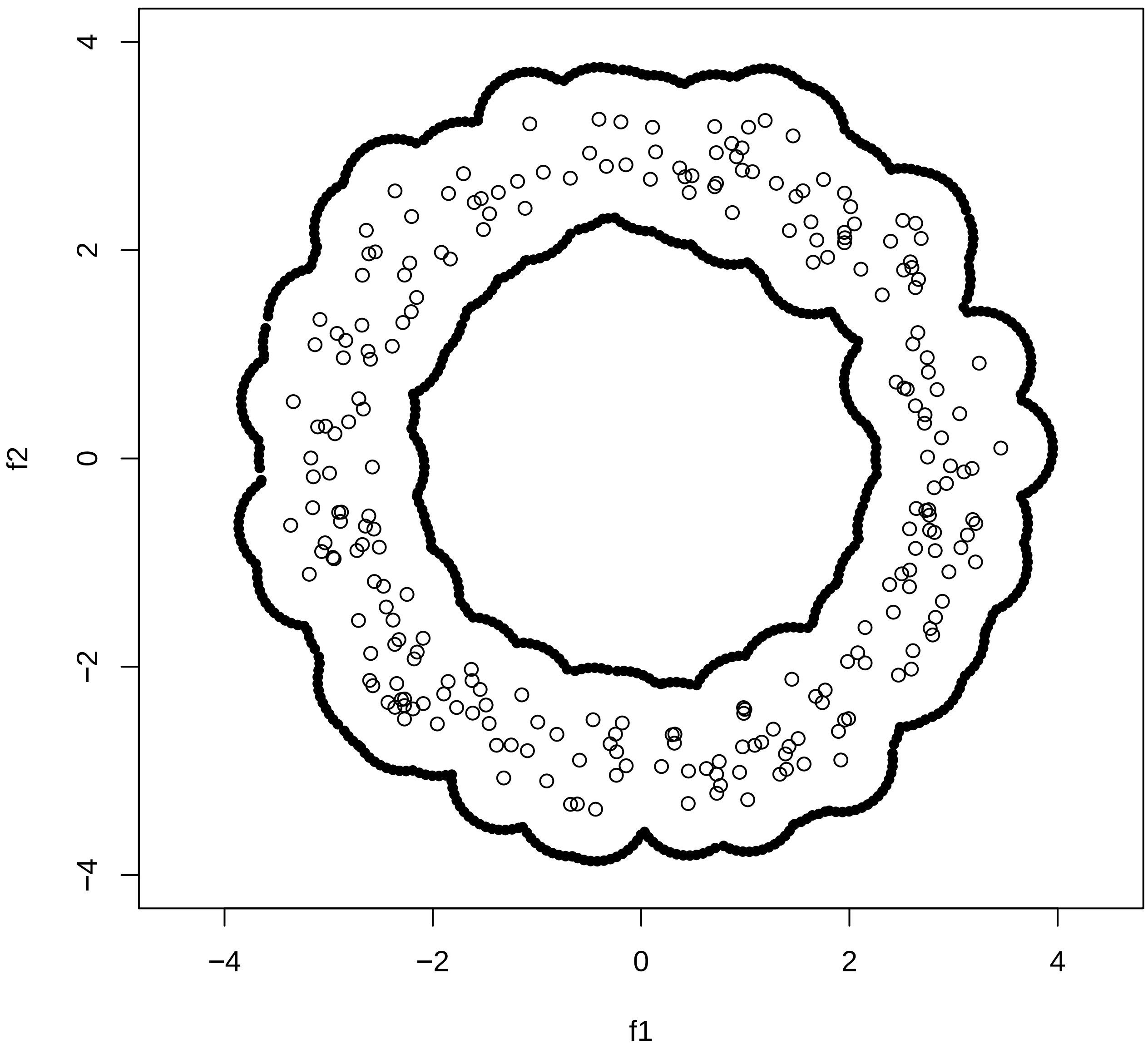
3
2
1
0
−1
−2
−3
f2
−4
−2
0
2
4
f1

Normal Smoothing



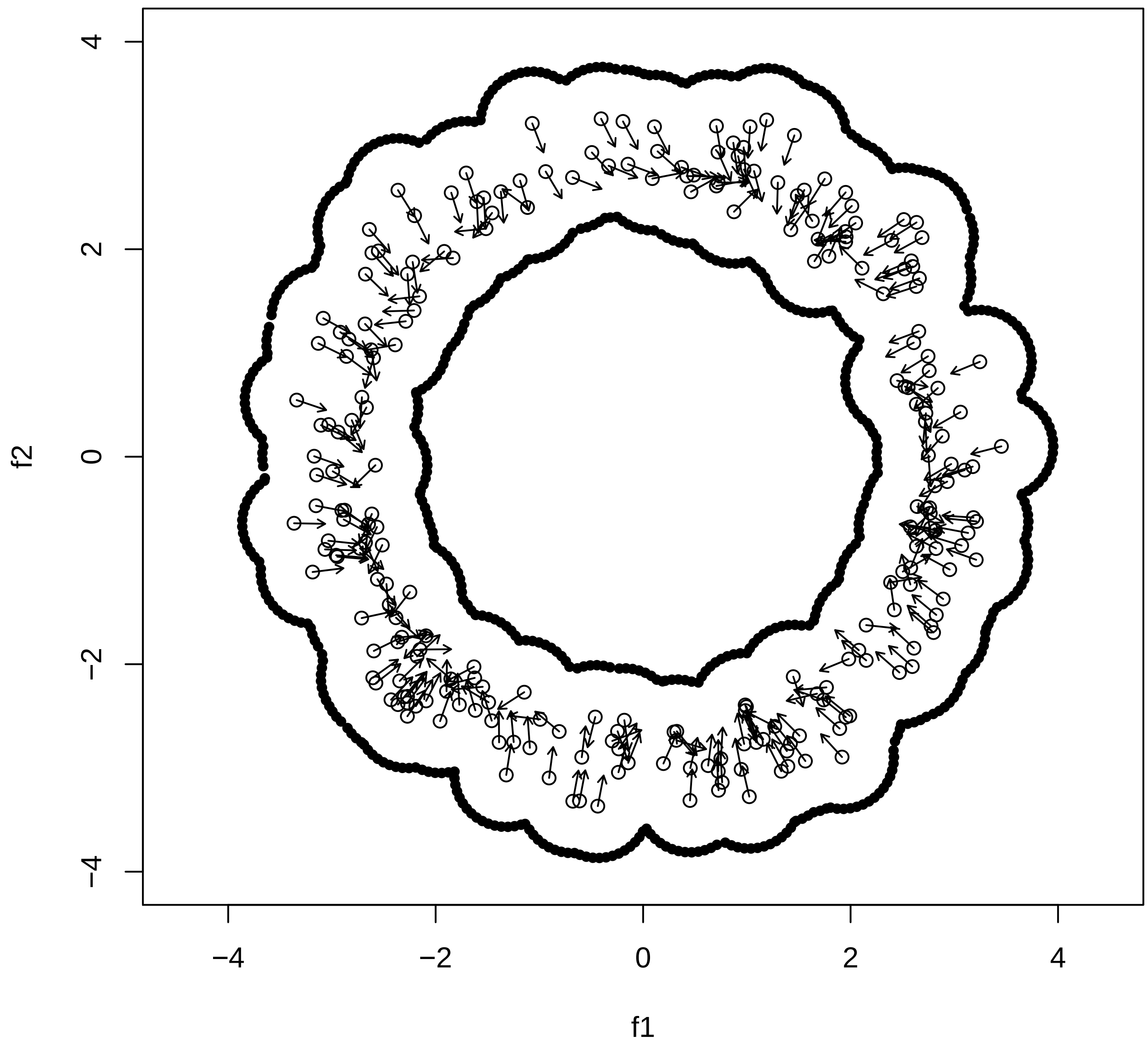
f2
3
2
1
0
−1
−2
−3
−4
−2
0
2
4
f1

Normal Smoothing



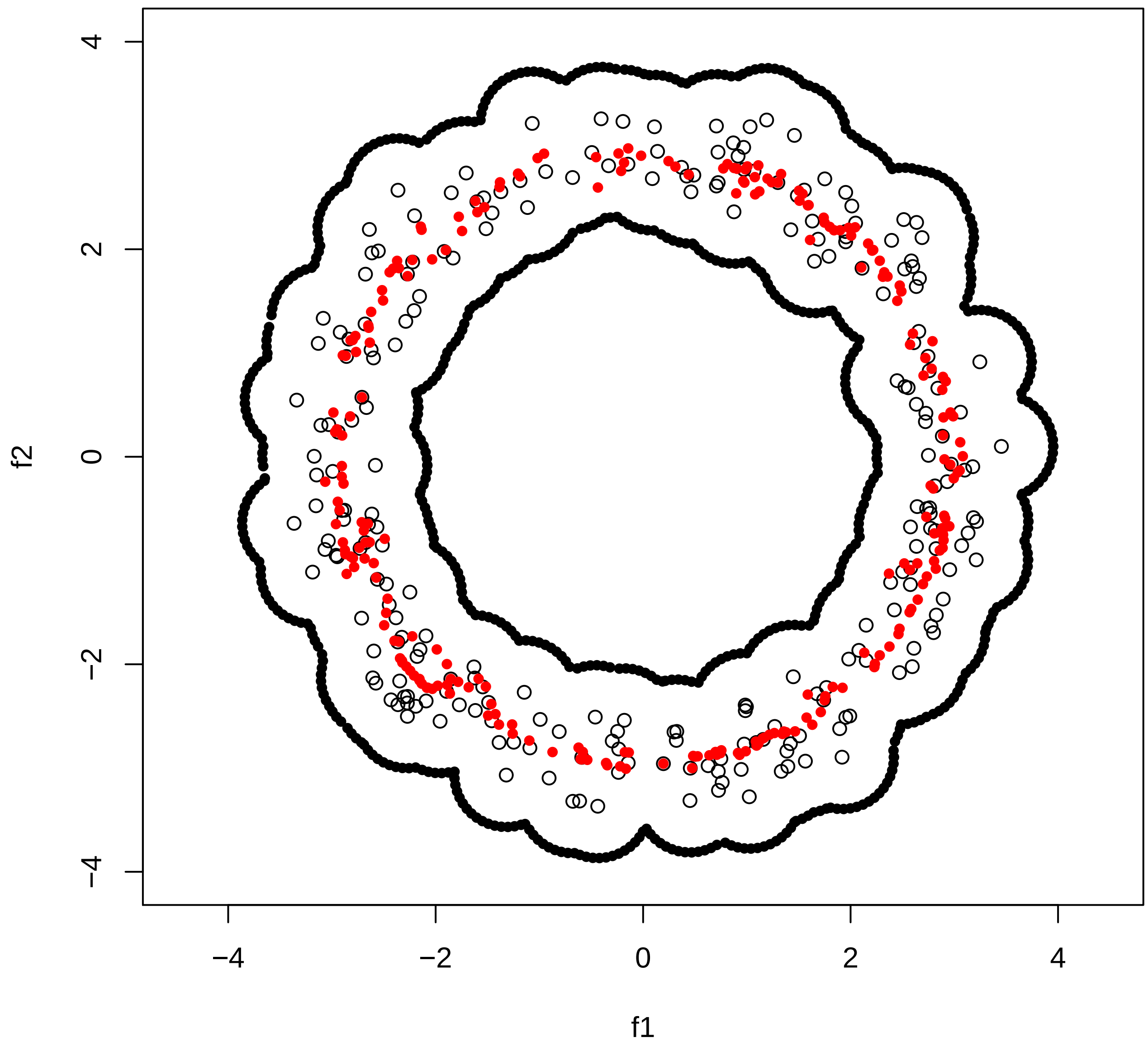
f2
4
2
0
−2
−4
−4
−2
0
2
4
f1

Normal Smoothing



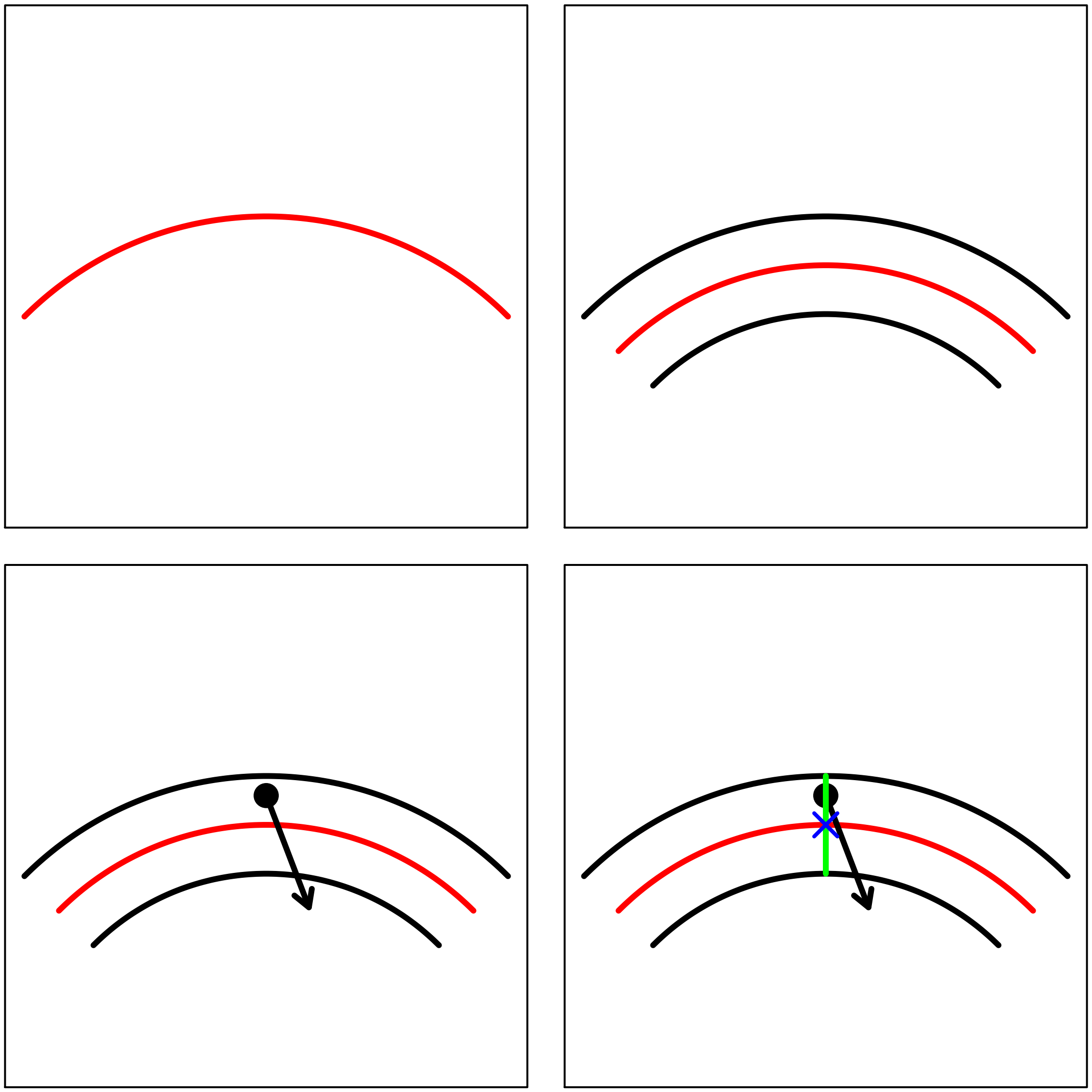
f2
4
2
0
−2
−4
−4
−2
0
2
4
f1

Normal Smoothing



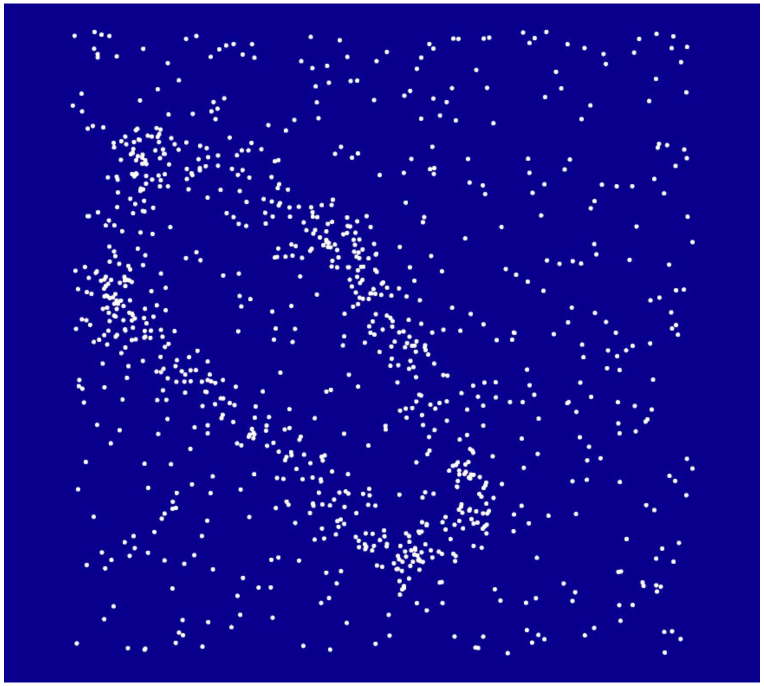
f1
f2
−4
−2
0
2
4

Bias Adjustment



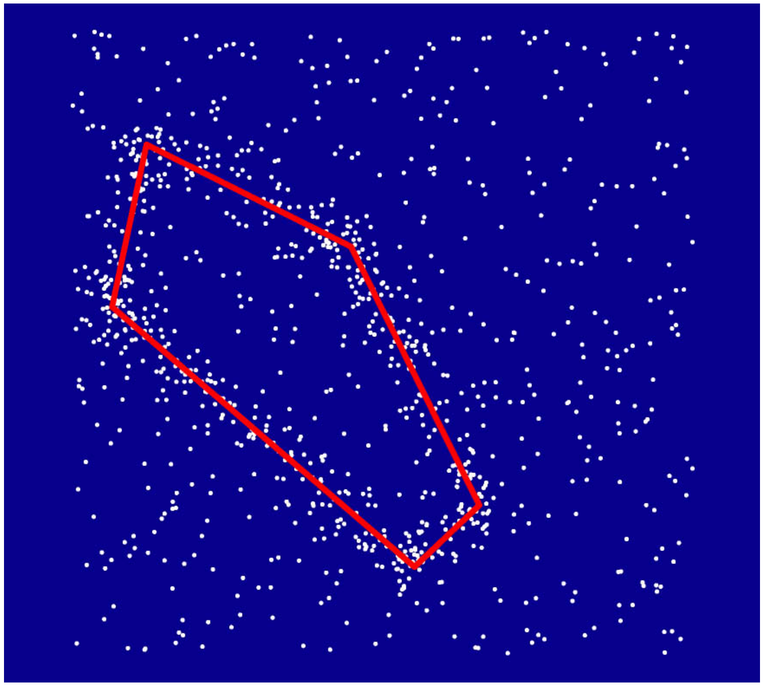
Gradient Method

Filaments correspond to ridges of the marginal density $p(y)$ of Y.



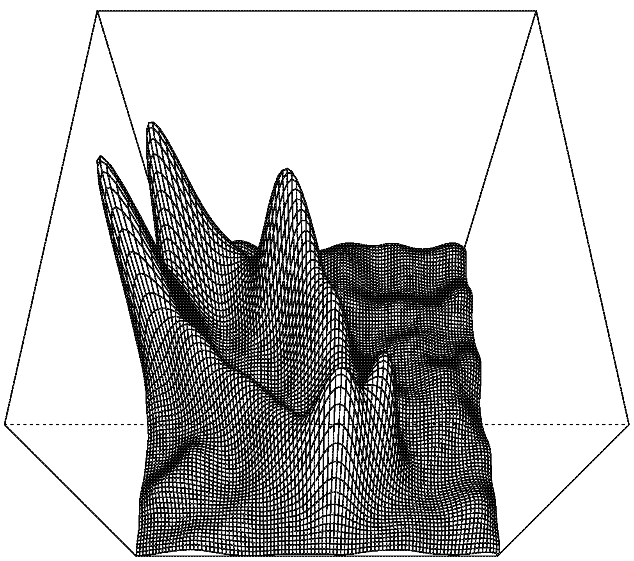
Genovese, Perone-Pacifico, Verdinelli and Wasserman (Annals, to appear).

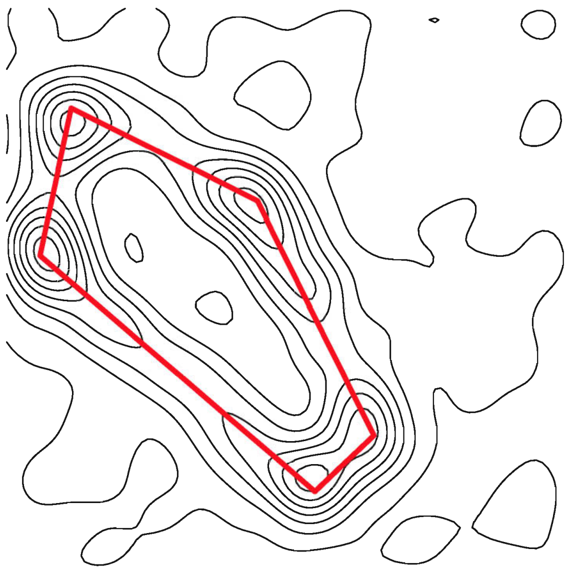
Gradient Method



Gradient Method

Filaments are ridges of the density





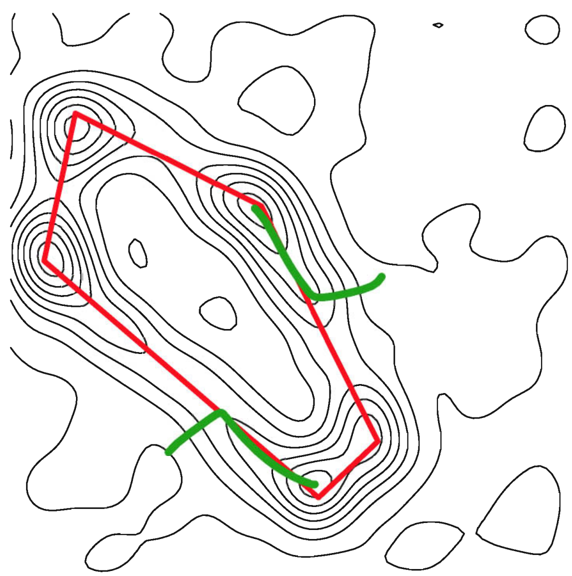
Mean Shift

The mean shift algorithm (Fukunaga and Hostetler 1975, Cheng 1995) is a mode-finding procedure that moves a point along the steepest-ascent paths of the kernel density estimate until a mode is reached.

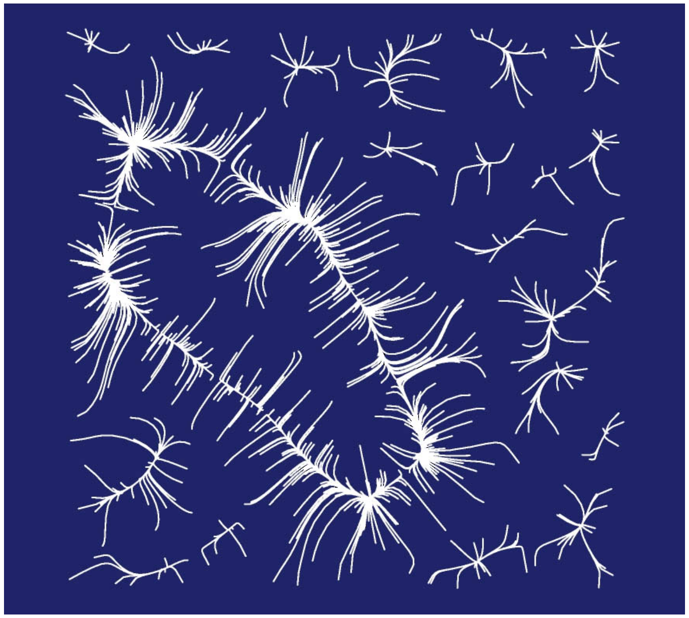
The path sa produced by the algorithm from any point approximates the steepest-ascent path for p.

Empirical observation: the mean-shift paths concentrate along filaments.

Mean-shift paths concentrate along filaments:



Mean-shift paths



Gradient Method

The concentration of the mean-shift paths suggests an approach to filament estimation: look for regions with a high concentration of paths.

We formalize this by studying the relationship between filaments and the steepest-ascent paths of the true density p.

We define the path density based on the probability that the steepest-ascent path starting at a random point X gets close to x:

$$p(y) = \lim_{r \to 0} \frac{\mathbb{P}(\mathrm{sa}(Y) \cap B(y, r) \neq \emptyset)}{r}$$

For any $\epsilon > 0$ and for $\lambda \geq \epsilon$,

$$\Gamma_f \subset \{p > \lambda\} \subset B(\Gamma_f, r(\lambda) + \epsilon),$$

for $r(\lambda)$ decreasing.

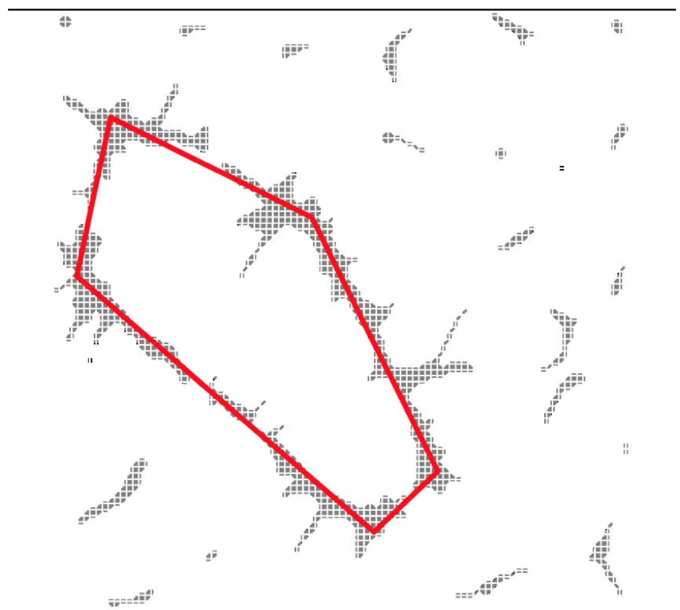
path density estimator

We define a kernel estimator for the path density based on the mean shift paths $\widehat{\mathsf{sa}}$

$$\widehat{p}_n(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\nu_n}K\left(\frac{\inf_{z\in\widehat{\mathsf{sa}}(Y_i)}||z-y||}{\nu_n}\right)$$

$$\sup_y |\widehat{p}_n(y) - p(y)| = O_P\left(\frac{\log n}{n^{1/4}}\right)$$

Example



levelset at 90-th percentile of density estimate

Example