Simplicity and Inference

or

Some Case Studies From Statistics and Machine Learning

or

Why We Need a Theory of Oversmoothing

CFE Workshop, Carnegie Mellon, June 2012

Larry Wasserman
Dept of Statistics and Machine Learning Department
Carnegie Mellon University

Conclusion

Statistics needs a rigorous theory of oversmoothing (undefitting).

There are hints:

G. Terrell (the oversmoothing principle)
D. Donoho (one-sided inference)
L. Davies (simplest model consistent with the data).

But, as I'll show, we need a more general way to do this.

Plan

1. Regression
2. Graphical Models
3. Density Estimation (simplicity versus L_2)
4. Topological data analysis

Preview of the Examples

1. (High-Dimensional) Regression: Observe $(X_1, Y_1), \ldots, (X_n, Y_n)$.
Observe new X.
Predict new Y.

Here, $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$ with $d > n$.

2. (High-Dimensional) Undirected Graphs.
$X \sim P$. $G = G(P) = (V, E)$. $V = \{1, \ldots, d\}$. $E =$ edges.
No edge between j and k means $X_j \amalg X_k |$ rest.

Preview of the Examples

3. Density Estimation. $Y_1, \ldots, Y_n \sim P$ and P has density p.
Estimator:

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

Need to choose h.

4. Topological Data Analysis.
$X_1, \ldots, X_n \sim G$.
G is supported on a manifold M.
Observe $Y_i = X_i + \epsilon_i$.
Want to recover the homology of M.

Regression

Best predictor is

$$m(x) = \mathbb{E}(Y|X = x).$$

Assume only iid and bounded random variables.

There is no uniformly consistent (distribution free) estimator of m.

How about the best linear predictor? Excess risk:

$$\mathcal{E}(\widehat{\beta}) = \mathbb{E}(Y - \widehat{\beta}^T X)^2 - \inf_{\beta} \mathbb{E}(Y - \beta^T X)^2.$$

But, as $n \to \infty$ and $d = d(n) \to \infty$

$$\inf_{\widehat{\beta}} \sup_{P} \mathcal{E}(\widehat{\beta}) \to \infty.$$

Simplicity: Best Sparse Linear Predictor

Let

$$\mathcal{B}_k = \{\beta : \ ||\beta||_0 \leq k\}$$

where $||\beta||_0 = \#\{j : \ \beta_j \neq 0\}$. Small $||\beta||_0$ = simplicity.

Good news: If $\widehat{\beta}$ is best subset estimator then

$$\mathbb{E}(Y - \widehat{\beta}^T X)^2 - \inf_{\beta \in \mathcal{B}_k} \mathbb{E}(Y - \beta^T X)^2 \to 0.$$

Bad news: Computing $\widehat{\beta}$ is NP-hard.

Convex Relaxation: The Lasso

Let

$$\mathcal{B}_L = \{\beta : \ ||\beta||_1 \leq L\}$$

where $||\beta||_1 = \sum_j |\beta_j|$. Note that L controls sparsity (simplicity).

Oracle: β_* minimizes $R(\beta)$ over $\mathcal{B}_L$.

Lasso: $\widehat{\beta}$ minimizes $\frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T X_i)^2$ over $\mathcal{B}_L$.

In this case:

$$\sup_P P(R(\widehat{\beta}) > R(\beta_*) + \epsilon) \preceq \exp\left(-cn\epsilon^2\right).$$

But how to choose L?

Choosing L (estimating the simplicity)

Usually, we minimize risk estimator $\widehat{R}(L)$ (such as cross-validation). It is known that this overfits.

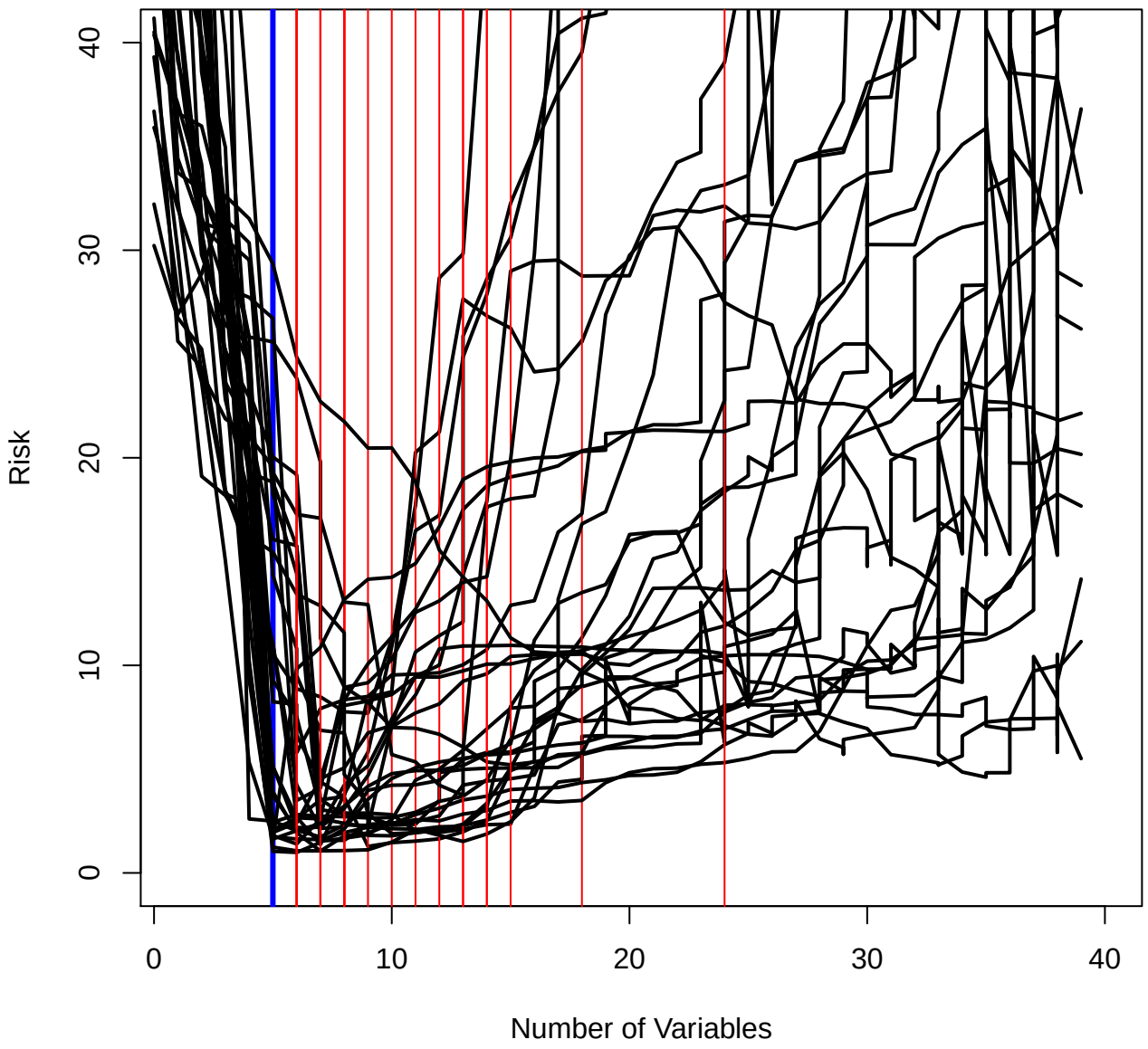
Theorem : (Meishausen and Buhlmann, Wasserman and Roeder):

$$P^n(\text{support}(\beta) \subset \text{support}(\widehat{\beta}_*)) \to 1$$

where β_* minimizes the true prediction loss.

But if we try to correct by moving to a simpler model, we risk huge losses since the risk function is asymmetric.

Here is a simulation: true model size is 5. ($d = 80$, $n = 40$).


Risk
0
10
20
30
40
0
10
20
30
40
Number of Variables

Corrected Risk Estimation

In other words:

simplicity $\neq$ accurate prediction

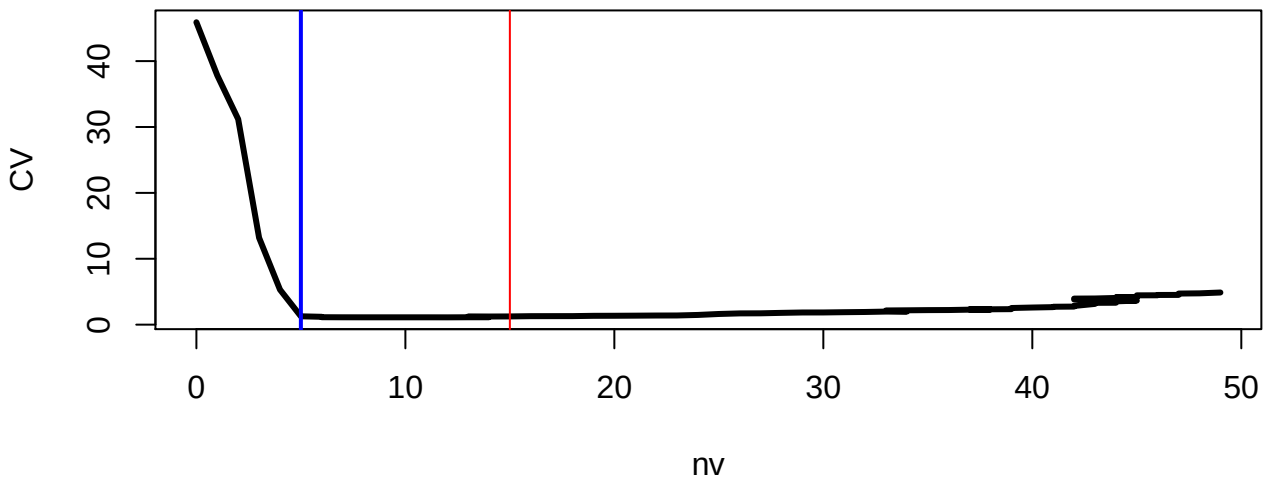
High predictive accuracy requires that we overfit.

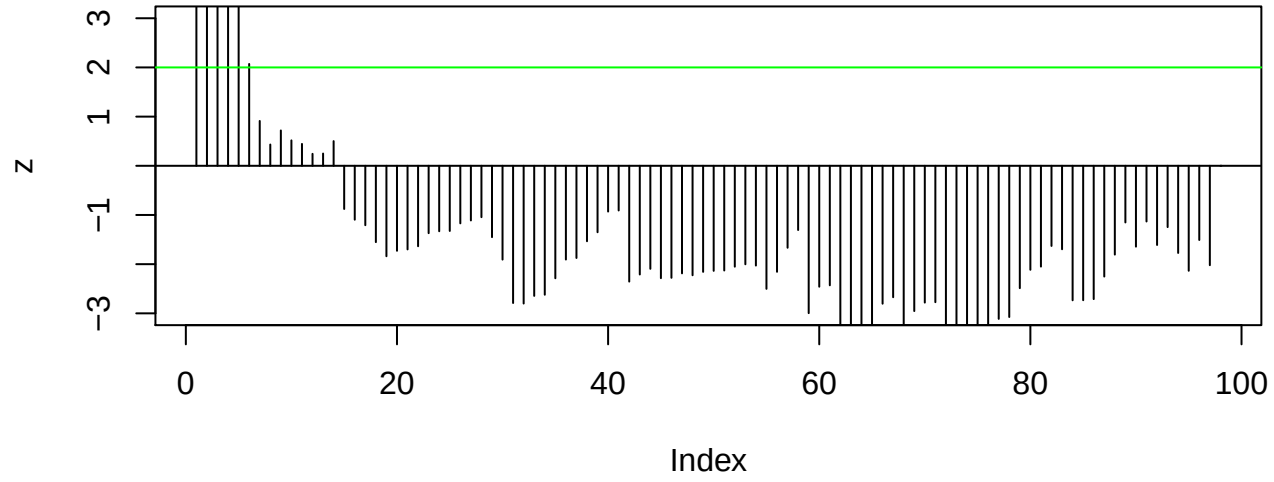
What if we want to force more simplicity? Can we correct the overfitting without incurring a disaster?

Safe simplicity:

$$Z(\Lambda) = \sup_{\ell \geq \Lambda} \frac{|\widehat{R}(\Lambda) - \widehat{R}(\ell)|}{s(\Lambda, \ell)}.$$

(This is Lepski's nonparametric method, adapted to the lasso.)


CV
0
10
20
30
40
0
10
20
30
40
50
nv

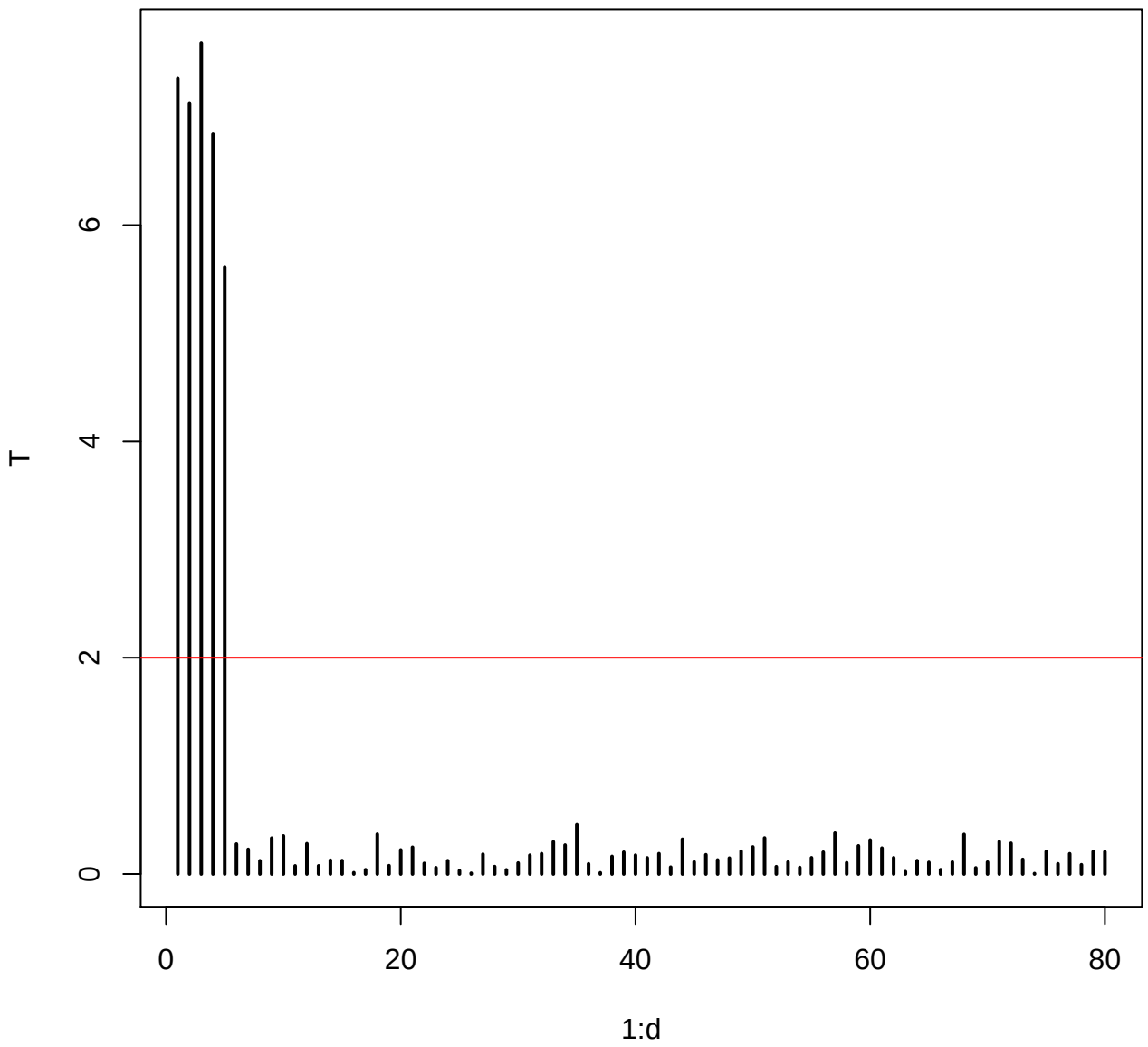

z
−3
−1
1
2
3
0
20
40
60
80
100
Index

Screen and Clean

Even better: Screen and Clean (Wasserman and Roeder, Annals 2009).

Split data into three parts:

Part 1: Fit lasso
Part 2: Variable selection by cross-validation
Part 3: Least squares on surviving variables followed by ordinary hypothesis testing.


T
0
2
4
6
0
20
40
60
80
1:d

Screen and Clean

However, this is getting complicated (and inefficient).

What happens when linearity is false, high correlations etc.?

Is there anything simpler?

Graphs

$X = (X_1, \ldots, X_d)$.
$G = (V, E)$.
$V = \{1, \ldots, d\}$.
$E =$ edges.
$(j, k) \notin E$ means that $X_j \amalg X_k |$rest.

$$X \text{ ——— } Z \text{ ——— } Y$$

means $X \amalg Y | Z$.

Observe: $X^{(1)}, \ldots, X^{(n)} \sim P$. Infer G.

Graphs

Common approach: assume $X^{(1)}, \ldots, X^{(n)} \sim N(\mu, \Sigma)$.

Find $\widehat{\mu}, \widehat{\Sigma}$ to maximize

$$\text{loglikelihood}(\mu, \Sigma) - \lambda \sum_{j \neq k} |\Omega_{jk}|$$

where $\Omega = \Sigma^{-1}$.

Omit an edge if $\widehat{\Omega}_{jk} = 0$.

Same problems as lasso: no good way to choose λ. In addition, non-Gaussianty seems to lead to overfitting.

The latter can be alleviated using forests.

Forests

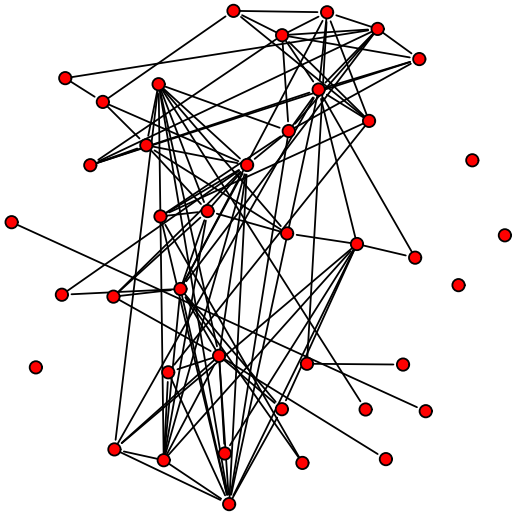
A forest is a graph with no cycles. In this case

$$p(x) = \prod_{j=1}^{d} p_j(x_j) \prod_{(j,k)\in E} \frac{p_{jk}(x_j, x_k)}{p_j(x_j)p_k(x_k)}.$$

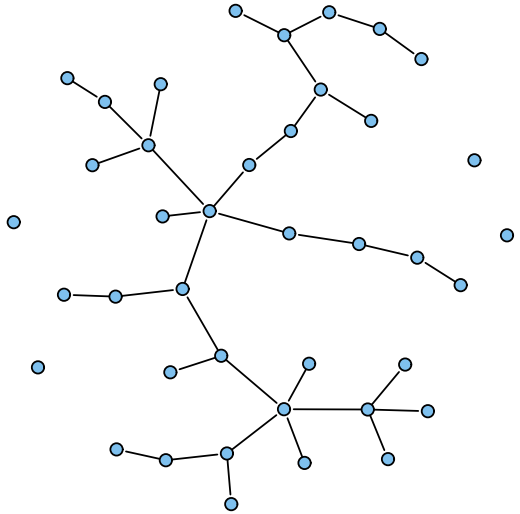
The densities can be estimated nonparametrically. The edge set can be estimated by the Chow-Liu algorithm based on nonparametric estimates of mutual information $I(X_j, X_k)$.

Han Liu, Min Xu, Haijie Gu, Anupam Gupta, John Lafferty, Larry Wasserman (JMLR 2010)

Gene Microarray:

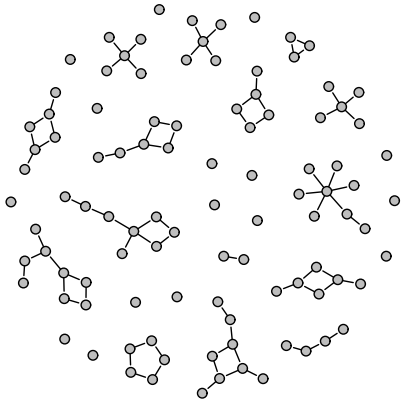


Glasso

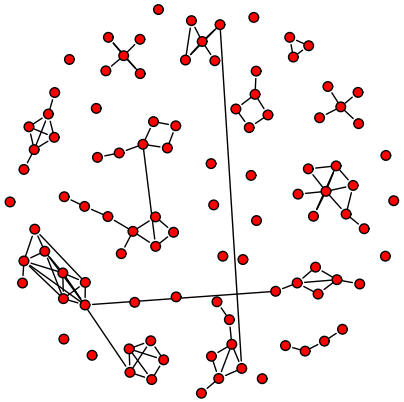


Nonparametric

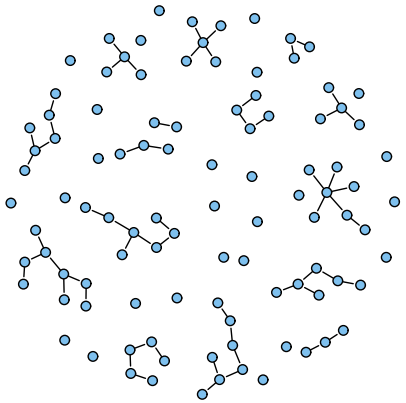
Synthetic Example:



True



Best Fit Glasso



Nonparametric

What We Really Want

Cannot estimate the truth. There is no universally consistent, distribution free test of

$$H_0 : X \amalg Y | Z.$$

We are better off asking: What is the simplest graphical model consistent with the data?

What We Really Want

Precedents for this are: Davies, Terrell, Donoho etc. Here is Davies idea (simplified).

Observe $(X_1, Y_1), \ldots, (X_n, Y_n)$. For any function m we can write

$$Y_i = m(X_i) + \epsilon_i = \text{signal} + \text{noise}$$

where $\epsilon_i = Y_i - m(X_i)$. He finds the "simplest" function m such that $\epsilon_1, \ldots, \epsilon_n$ look like "noise."

Davies, Kovac and Meise (2009) and Davies and Kovac (2001).

How do we do this for graphical models?

Density Estimation

Seemingly and old, solved problem.
$Y_1, \ldots, Y_n \sim P$ where $Y_i \in \mathbb{R}^d$ and P might have a density p.
kernel estimator

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

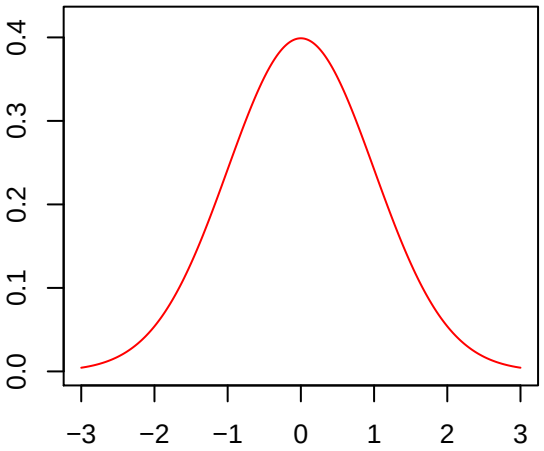
Here, K is a kernel and $h > 0$ is a bandwidth.

How do we choose h? Usually, we minimize an estimate of

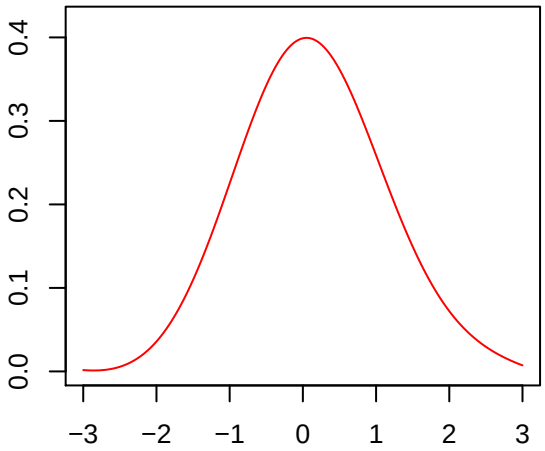
$$R(h) = \mathbb{E}\left(\int (\widehat{p}_h(x) - p(x))^2 dx\right).$$

But this is the wrong loss function ...

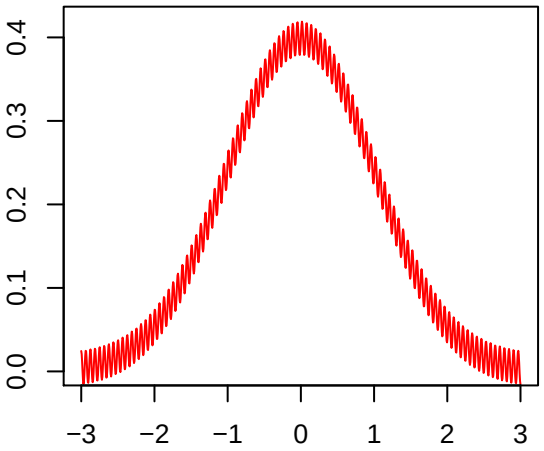
p_0



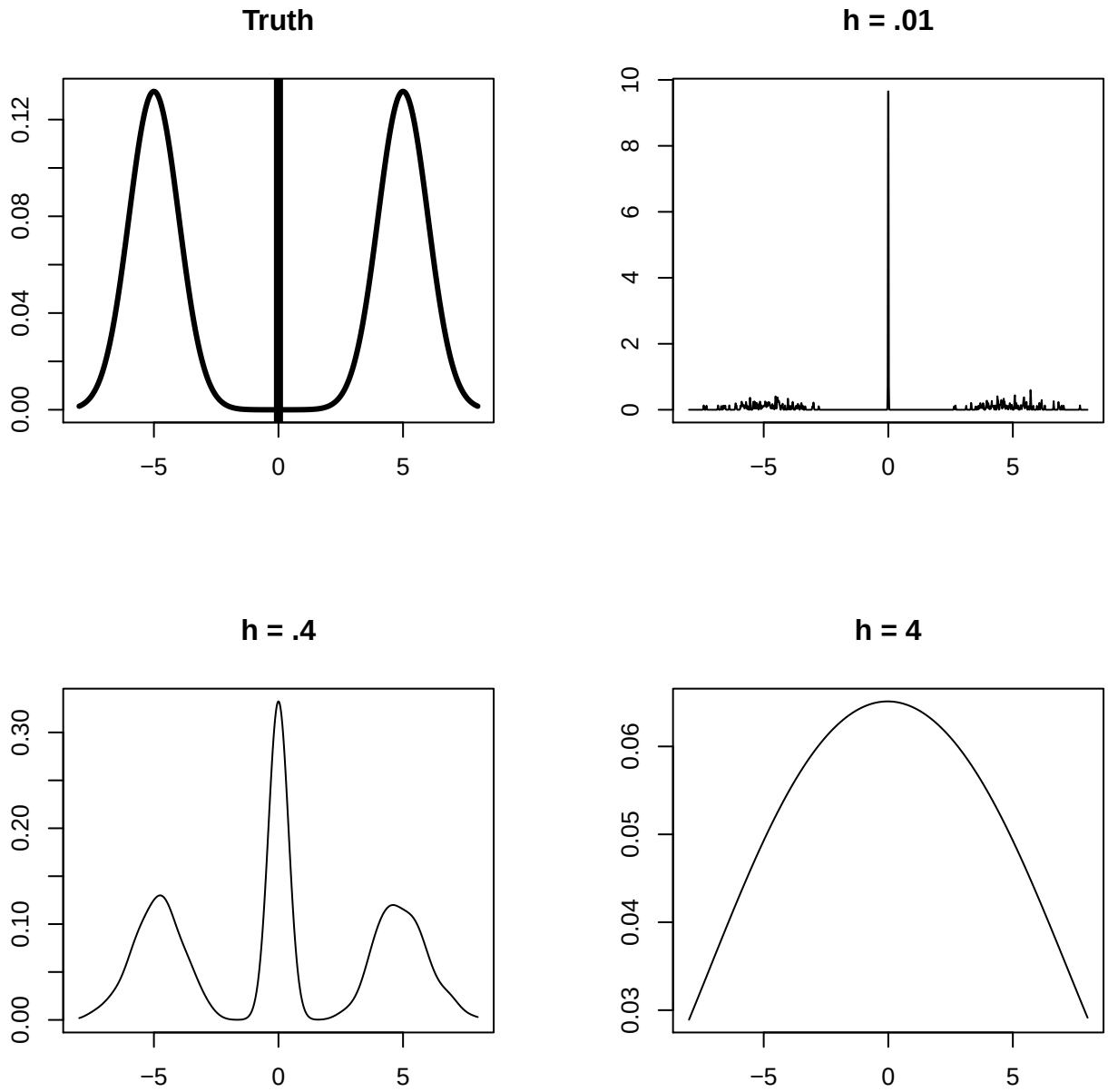
p_1



p_2



$$L_2(p_0, p_1) = L_2(p_0, p_2)$$


Truth
h = .01
h = .4
h = 4
0.00
0.04
0.08
0.12
0
2
4
6
8
10
0.00
0.10
0.20
0.30
0.03
0.04
0.05
0.06
−5
0
5

Density Estimation

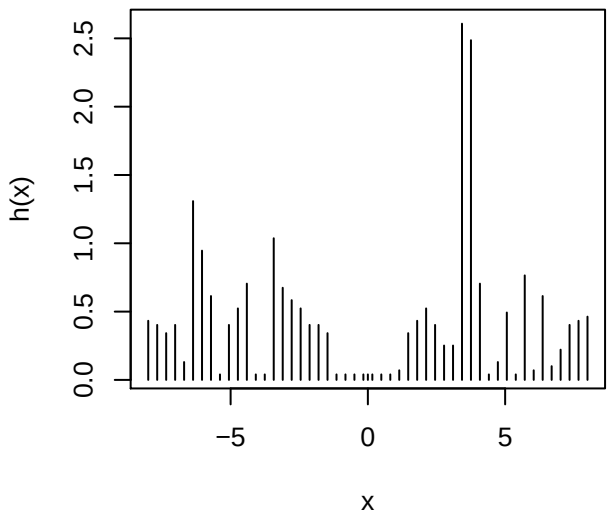
More generally p might have: smooth parts, singularities, near singularities (mass concentrated near manifolds) etc.

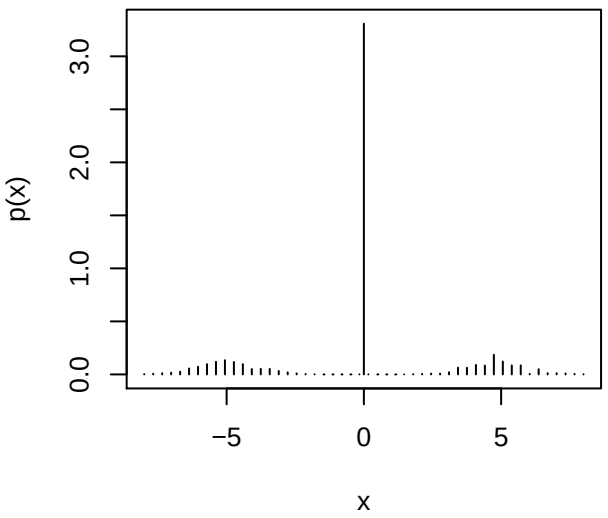
In principle we can use Lepski's method: choose a local bandwidth

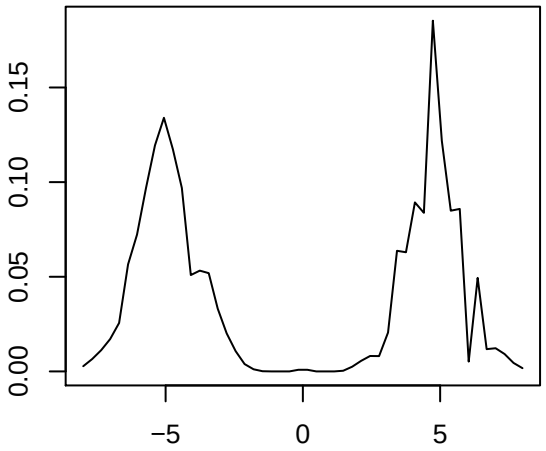
$$\widehat{h}(x) = \sup\Big\{h : \ |\widehat{p}_h(x) - \widehat{p}_t(x)| \leq \psi(t,h) \text{ for all } t < h\Big\}.$$

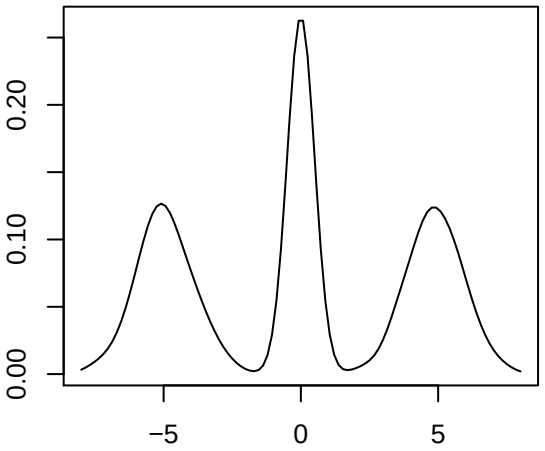
Lepski and Spokoiny 1997, Lepski, Mammen and Spokoiny 1997.

It leads to this ...


h(x)
0.0
0.5
1.0
1.5
2.0
2.5
−5
0
5
x


p(x)
0.0
1.0
2.0
3.0
−5
0
5
x


0.00
0.05
0.10
0.15
−5
0
5


0.00
0.10
0.20
−5
0
5

Oversmoothing

We really just want a principled way to oversmooth. Terrell and Scott (1985) and Terrell (1990) suggest the following: choose the largest amount of smoothing compatible with the scale of the density.

The asymptotically optimal bandwidth (with $d = 1$) is

$$h = \left(\frac{\int K^2(x)dx}{n\sigma_K^4 \, I(p)} \right)^{\frac{1}{5}}$$

where $I = \int (p'')^2$. Now: minimize $I = I(p)$ subject to:

$$T(P) = T(\widehat{P}_n)$$

where $T(\cdot)$ is the variance.

Oversmoothing

Solution:

$$h = \frac{1.47\, s \left(\int K^2\right)^{\frac{1}{5}}}{n^{\frac{1}{5}}}.$$

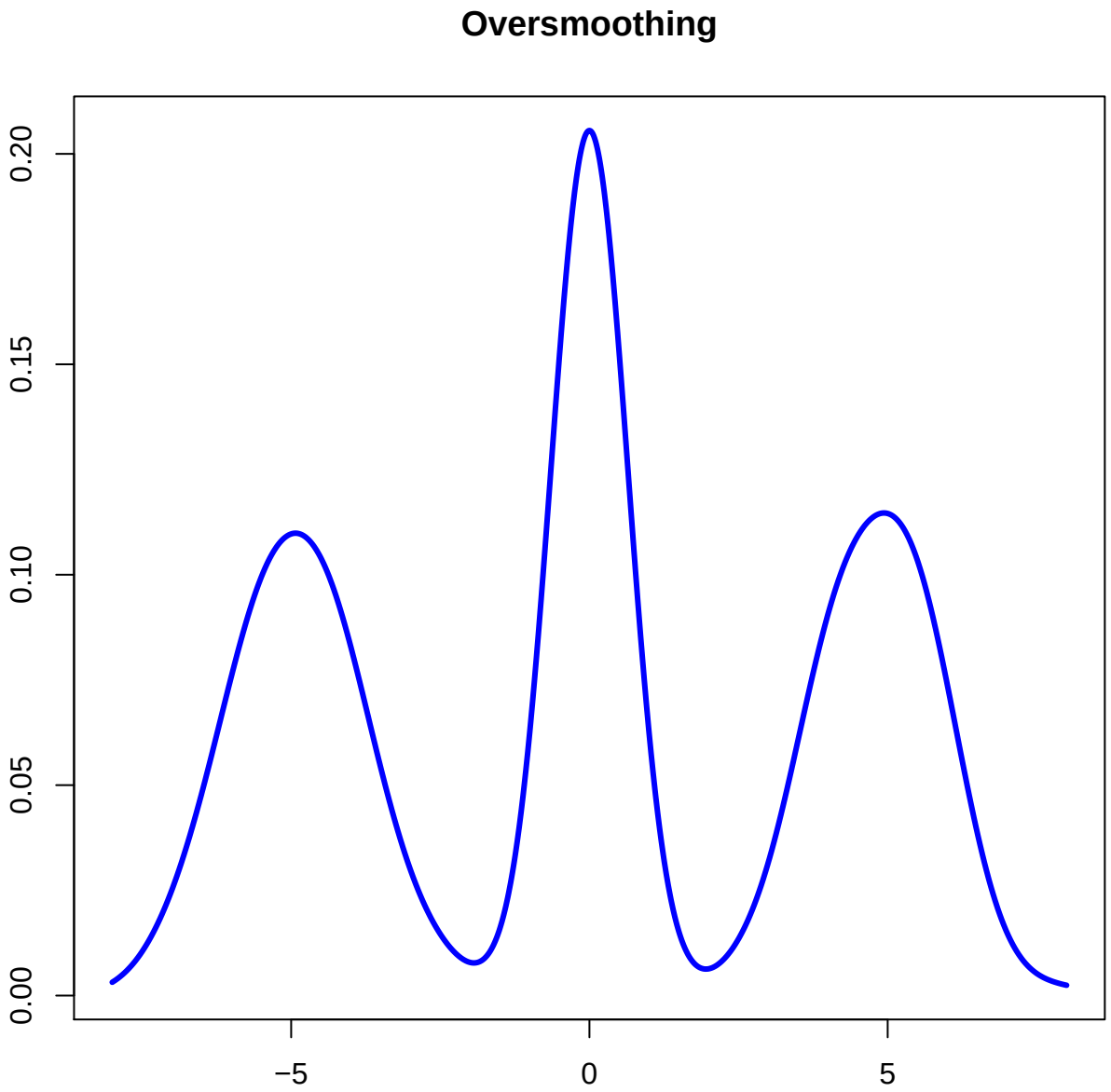
Good idea, but:

- it is still based on L_2 loss

- it is based on an asymptotic expression for optimal h.

We need a finite sample version with a more appropriate loss function.

Here it is on our example:


Oversmoothing
0.00
0.05
0.10
0.15
0.20
−5
0
5

Summary so far ...

Our current methods select models that are too complex.

Are there simple methods for choosing simple models?

Now, a more exotic application ...

Topological Data Analysis

$X_1, \ldots, X_n \sim G$ where G is supported on a manifold M.

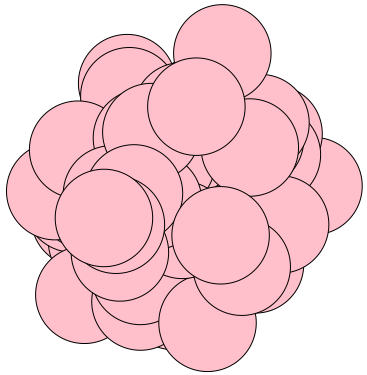
Here $X_i \in \mathbb{R}^D$ but dimension$(M) = d < D$.

Observe $Y_i = X_i + \epsilon_i$.

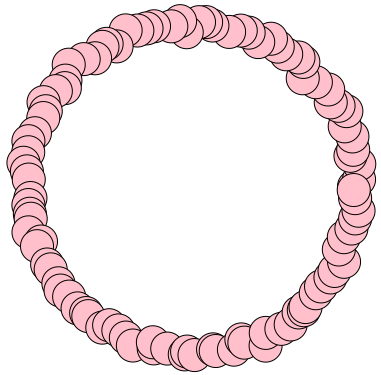
Goal: infer the homology of M.

Homology: clusters, holes, tunnels, etc.

One Cluster



One Cluster + One Hole



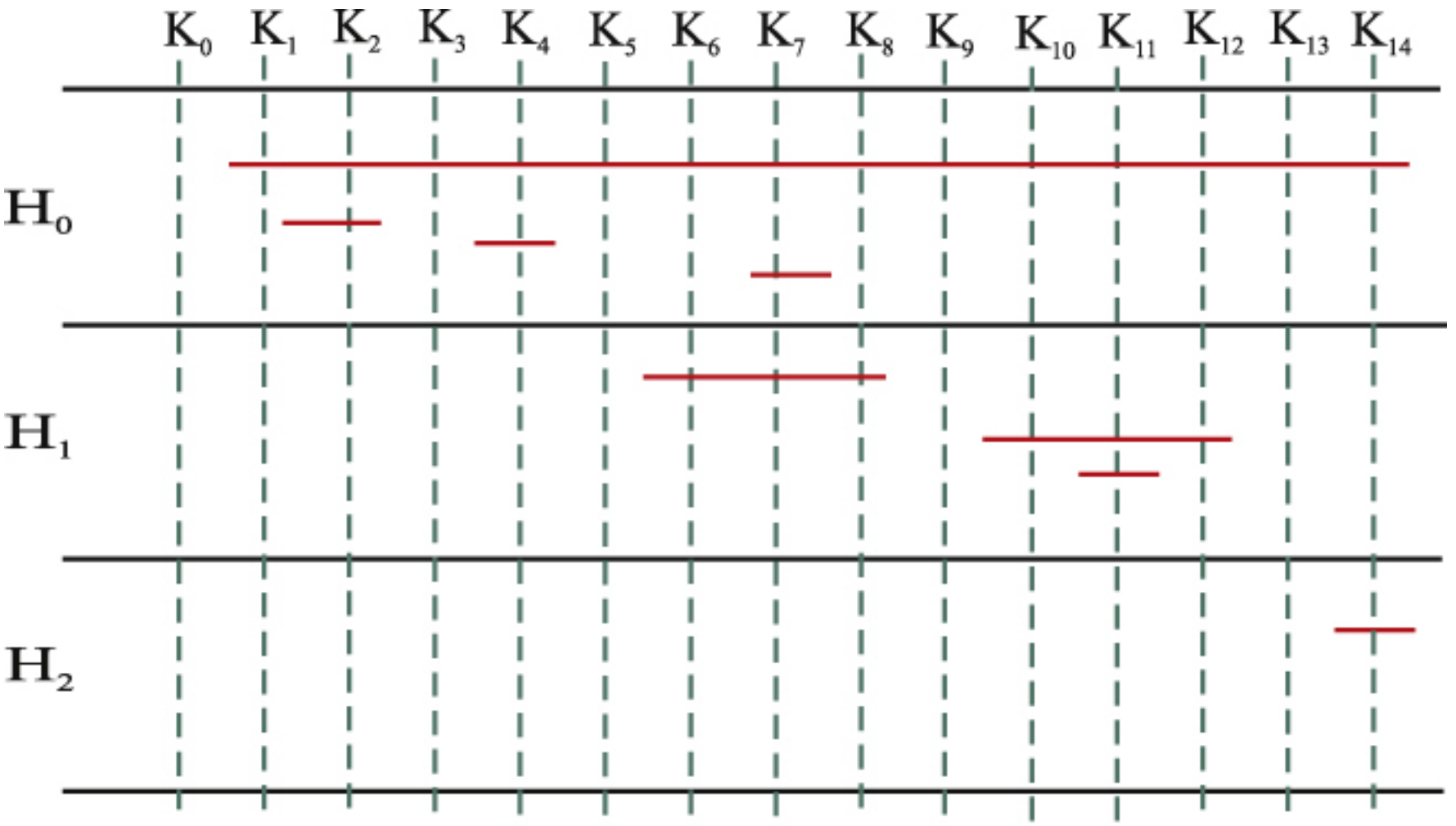
Inference

The Niyogi, Smale, Weinberger (2008) estimator:

1. Estimate density. (h)
2. Throw away low density points. (t)
3. Form a Cech complex. (ϵ)
4. Apply an algorithm from computational geometry.

Usually, the results are summarized as a function of ϵ in a barcode plot (or a persistence diagram).

Example: from Horak, Maletic and Rajkovic (2008)



The Usual Heuristic

Usually assume that small barcodes are topological noise.

This is really a statistical problem with many tuning parameters.

Currently, there are no methods for choosing the tuning parameters.

What we want: the simplest topology consistent with the data.

Working on this with Sivaraman Balakrishnan, Aarti Singh, Alessandro Rinaldo and Don Sheehy.

Summary

We still don't know how to choose simple models.

Summary

THE END